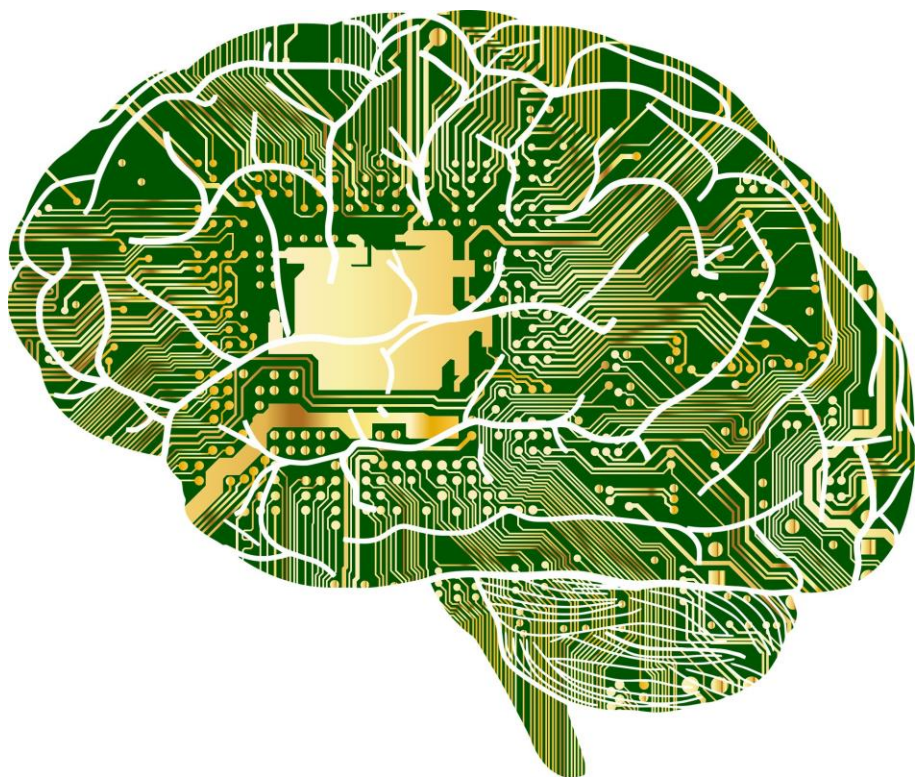




МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

Хмельницький національний університет

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

*Методичні вказівки до самостійної роботи студентів спеціальності
121 «Інженерія програмного забезпечення» та 113 «Прикладна
математика»*

*Затверджено на засіданні кафедри
інженерії програмного забезпечення
Протокол № 4 від 27.12.2018*

м. Хмельницький, 2019

Методи інтелектуального аналізу даних в інформаційних технологіях : методичні вказівки до самостійної роботи студентів спеціальностей 121 «Інженерія програмного забезпечення» та 113 «Прикладна математика» / Н.В. Грипинська, П.М. Радюк. – Хмельницький : ХНУ, 2019. – 222 с.

Укладач: Грипинська Н.В., канд. фіз-мат. наук, доц.
Радюк П.М., аспірант

Рецензент: Бедратюк Л.П., докт. фіз-мат.н., професор

Редактор-коректор: Яремчук В. С.

© ХНУ, 2019

Навчальна дисципліна «Методи інтелектуального аналізу даних в інформаційних системах» є обов'язковою (цикл професійної та практичної підготовки) для студентів напрямків підготовки 12 «Інформаційні технології» та 11 «Математика і статистика» спеціальностей 121 «Інженерія програмного забезпечення» та 113 «Прикладна математика». Лекційний обсяг дисципліни становить 34 години. Практична частина дисципліни (в даних методичних вказівках не розглядається) складається з лабораторних та самостійних робіт, виконання яких спрямоване на опанування вміння в проведенні інтелектуального аналізу даних.

Метою вивчення дисципліни «Методи інтелектуального аналізу даних в інформаційних технологіях» є формування компетентностей, а саме:

ФК03. Здатність обирати та застосовувати математичні методи для розв'язання прикладних задач, моделювання, аналізу, проектування, керування, прогнозування, прийняття рішень.

ФК05. Здатність проектувати бази даних, інформаційні системи та ресурси.

ФК07. Здатність експлуатувати та обслуговувати програмне забезпечення автоматизованих та інформаційних систем різного призначення.

ФК09. Здатність до проведення математичного і комп'ютерного моделювання, аналізу та обробки даних, обчислювального експерименту, розв'язання формалізованих задач за допомогою спеціалізованих програмних засобів.

К14. Здатність брати участь у проектуванні програмного забезпечення, включаючи проведення моделювання (формальний опис) його структури, поведінки та процесів функціонування.

К18. Здатність аналізувати, вибирати і застосовувати методи і засоби для забезпечення інформаційної безпеки (в тому числі кібербезпеки).

К19. Володіння знаннями про інформаційні моделі даних, здатність створювати програмне забезпечення для зберігання, видобування та опрацювання даних.

К20. Здатність застосовувати фундаментальні і міждисциплінарні знання для успішного розв'язання завдань інженерії програмного забезпечення.

Навчальна дисципліна «Методи інтелектуального аналізу даних в інформаційних технологіях» передбачає засвоєння методів

сучасної обробки даних – огляд методів, програмних продуктів і різних інструментальних засобів, які використовуються в аналізі даних; підготовка студентів до самостійної роботи з вирішення задач засобами інтелектуального аналізу даних і розробки інтелектуальних систем.

Передбачений робочим планом час на аудиторні заняття з дисципліни не дає можливості в необхідному обсязі викласти навчальний матеріал дисципліни. Тому передбачається самостійне опанування студентом частини курсу. Виконання самостійної роботи це – поглиблення, узагальнення й закріплення теоретичних знань і практичних умінь студентів із дисципліни, що вивчається, шляхом вироблення вміння самостійної роботи з навчальною і фаховою науково-технічною літературою.

Методичні рекомендації цього видання такі:

- ознайомити студентів із повним обсягом навчального матеріалу з дисципліни, у тому числі і з тією частиною, яка повністю виноситься на самостійне вивчення;

- навести необхідну навчальну літературу щодо кожної тематики дисципліни;

- надати методичні вказівки та контрольні питання для самоперевірки знань;

- ознайомити студентів із заходами контролю засвоєння навчального матеріалу з дисципліни.

Отриманні знання будуть використовуватися та доповнюватися під час подальшого вивчення дисциплін «Теорія прийняття рішень», «Методи та системи штучного інтелекту».

РОЗДІЛ 1.

ОСНОВНІ ПОНЯТТЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

1.1 Поняття інтелектуального аналізу даних

Інтелектуальний аналіз даних (ІАД) – термін, що застосовується для опису отримання знань у базах даних, дослідження даних, обробки зразків даних, очищення і збору даних. Це процес виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій.

Термін *Data Mining* дістав назву від двох понять: дані – з англ. “data”, і переробка сирого матеріалу (гірської руди) – з англ. “mining” [1].

Data Mining – предметна область що виникла й розвивається на основі таких наук, як прикладна статистика, розпізнавання образів, системи штучного інтелекту, теорія баз даних тощо.

Виникнення й розвиток *Data Mining* зумовлені різними чинниками, серед яких виділемо основні: вдосконалення програмно-апаратного забезпечення; вдосконалення технологій зберігання й запису даних; накопичення великої кількості ретроспективних даних; вдосконалення алгоритмів обробки інформації.

Існують різні означення *Data Mining*. Але вони збігаються у головному, оскільки мають чотири основні ознаки, які наявні у кращому означенні технології *Data Mining*, запропонованому у 1996 р. Г. П'ятецьким-Шапіро: «*Data Mining* – дослідження та виявлення “машинною” (алгоритмами, засобами штучного інтелекту) в сирих даних приваних, раніше невідомих, нетривіальних, практично корисних та доступних для інтерпретації людиною знань, необхідних для прийняття рішень у різних сферах людської діяльності» [2].

Data Mining – це процес виявлення в необроблених даних раніше невідомих нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах діяльності [3].

За визначенням SAS Institute, *Data Mining* – це процес виділення, дослідження й моделювання великих обсягів даних для виявлення невідомих до цього структур (patterns) для досягнення переваг у бізнесі [4].

В основу технології *Data Mining* покладено концепцію шаблонів (patterns), що є закономірностями, які властиві вибіркам даних і можуть бути подані у формі, зрозумілій людині [5].

У процесі роботи з даними багаторазово виконуються операції перетворення над “сирими” даними (відбір ознак, стратифікація,

кластеризація, візуалізація й регресія), що призначені для знаходження:

- структур, які інтуїтивно зрозумілі для людей і краще розкривають суть бізнес-процесів;

- моделей, які можуть передбачити результат або значення певних ситуацій, використовуючи історичні або суб'єктивні дані.

Метою ІАД є виявлення неявних закономірностей у наборах даних. Як науковий напрям він став активно розвиватися в 90-х роках ХХ століття, що було викликано широким поширенням технологій автоматизованої обробки інформації та накопиченням у комп'ютерних системах великих обсягів даних. І хоча наявні технології дозволяли, наприклад, швидко знайти в базі даних потрібну інформацію, цього в багатьох випадках було недостатньо. Виникла потреба пошуку взаємозв'язків між окремими подіями серед великих обсягів даних, для чого знадобилися методи математичної статистики, теорії баз даних, теорії штучного інтелекту та низки інших областей.

Суть і ціль технології Data Mining полягає у пошуку неочевидних, об'єктивних і корисних на практиці закономірностей у великих обсягах даних. В основу сучасної технології Data Mining покладена концепція шаблонів (патернів), які відображають фрагменти багатоаспектних взаємовідносин в даних. Ці фрагменти представляють собою закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у зрозумілій людині формі. Пошук шаблонів здійснюється методами, не обмеженими рамками апіорних припущень стосовно структури вибірки та виду розподілу значень змінних, що аналізуються.

Поняття Data Mining з'явилося у 1989 році, але високу популярність у сучасному трактуванні набуло приблизно у першій половині 1990-х років. Data Mining широко використовується у багатьох галузях, для яких характерне використання великих об'ємів даних: в науці, торгівлі, телекомунікаційній сфері, банківській справі, промисловому виробництві та інших галузях, де виникає задача автоматичного аналізу даних і прийняття рішень на його основі. Завдяки мережі Інтернет Data Mining використовується кожний день користувачами пошукових систем на просторах Інтернету. Data Mining лежить на перетині декількох наук, основні з яких – це системи баз даних, статистика та штучний інтелект.

До *методів та алгоритмів ІАД* відносять такі: штучні нейронні мережі, дерева рішень, символні правила, методи найближчого сусіда і k -найближчого сусіда, метод опорних векторів, байєсівські мережі, лінійну регресію, кореляційно-регресійний аналіз;

ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, у тому числі алгоритми k -середніх і k -медіани; методи пошуку асоціативних правил, наприклад, алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних та інші методи. До задач ІАД відносять: класифікацію, кластеризацію, асоціацію, послідовну асоціацію або просто послідовність, прогнозування, визначення відхилень або викидів, оцінювання параметрів і станів, аналіз зв'язків, візуалізацію, підбиття підсумків.

З огляду на різноманітність форм подання даних, використовуваних алгоритмів і сфер застосування, ІАД може проводитися з допомогою програмних продуктів таких класів [5]:

- спеціалізованих «коробкових» програмних продуктів для інтелектуального аналізу;
- математичних пакетів;
- електронних таблиць (і різного роду надбудов над ними);
- засобів інтегрованих у системи керування базами даних (СКБД);
- інших програмних продуктів.

У рамках даного курсу розглядаються засоби ІАД, інтегровані зі СКБД. Прикладом таких засобів є СКБД *Microsoft SQL Server* зі службами *Analysis Services*, що входять до її складу. Ці служби забезпечують користувачів засобами аналітичної обробки даних у режимі online (OLAP) та ІАД, які вперше з'явилися в *Microsoft SQL Server 2002*. Не тільки *Microsoft*, але й інші провідні розробники СКБД мають у своєму арсеналі засоби ІАД.

1.2 Поняття даних

Дослідження у сфері інтелектуального аналізу даних середини ХХ століття були зосереджені в основному на дослідженні табличних даних [2, 3]. Проте, з часом виникла потреба в обробці та аналізі як звичайних, так і специфічних текстових даних: html-документів, xml-файлів, json-файлів тощо. Особливість обробки та аналізу текстових даних у тому, що останні містять інформацію і в своїй структурі, і у семантичних зв'язках, що, в свою чергу, вимагає досліджень у сфері баз знань. Окрім того, сучасний стан розвитку інформаційних технологій зумовлює дослідження цікавих та складних об'єктів, що раніше не представляли інтересу, наприклад, соціальних мереж. Технологія *Data Mining* передбачає аналіз даних представлених у різних формах, придатних для зберігання, передачі та обробки.

У широкому розумінні, дані представляють собою факти, текст, графіки, картинки, звуки, аналогові або цифрові відеосегменти [6]. Дані можуть бути отримані в результаті *вимірів, експериментів, арифметичних і логічних операцій*. Іншими словами, *дані* – це необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

Найчастіше зустрічаються дані, що складаються із записів (англ. Record Data). Прикладами таких наборів даних є *табличні, матричні, документальні та транзакційні (або операційні)*.

Табличні дані складаються із записів, кожен з яких складається з фіксованого набору *атрибутів*. *Транзакційні дані* представляють собою особливий тип даних, де кожен запис, що є *транзакцією*, включає набір значень. Прикладами *графічних даних* є: WWW-дані, молекулярні структури, графи, карти тощо. Перевагою графічних даних є простота сприйняття у порівнянні з табличними та матричними даними.

Особливим чином організовані і збережені в електронному вигляді дані називають *базою даних* (англ. Database). Під «особливим чином» мається на увазі такий спосіб, що дає можливість полегшити їх пошук і доступ до них для одного або декількох застосунків. Така організація передбачає наявність мінімісованої надмірності даних.

Бази даних є одним з різновидів інформаційних технологій, а також формою зберігання даних. Метою створення баз даних є побудова такої системи, що була б незалежною від програмного забезпечення та застосовуваних технічних засобів. Побудова такої системи має забезпечувати несуперечливу і цілісну інформацію. під час проектування бази даних передбачається багатоцільове її використання. У найпростішому випадку база даних представляються у вигляді системи двовимірних таблиць.

Набір даних і їх атрибутів

Найпростіший і досить популярний формат зберігання даних – *таблиці*. Вони широко використовуються під час дослідження та аналізу даних.

Таблиця (англ. Table) – це набір елементів даних (*значень*), які організовані з використанням моделі вертикальних стовпчиків (з різними іменами) і горизонтальних рядків. По горизонталі таблиці розташовуються *атрибути* об'єкта (його ознаки). По вертикалі таблиці – власне *об'єкти*. Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці.

Атрибут (англ. Attribute) – це властивість, що характеризує об’єкт (наприклад: колір очей людини, температура води тощо). Атрибут також називають полем таблиці, вимірюванням, характеристикою.

Змінна (англ. Variable) – властивість або характеристика, загальна для всіх досліджуваних об’єктів, прояв якої може змінюватися від об’єкта до об’єкта. *Значення* (англ. Value) змінної є проявом ознаки.

Аналіз даних, як правило, не передбачає можливості розглянути всю цікаву сукупність об’єктів. Обробка та вивчення великих обсягів даних є дорогим процесом, який вимагає значних затрат часу, а також неминуче призводить до помилок, пов’язаних з людським фактором. Для більшості задач буває цілком достатньо розглянути деяку частину всієї сукупності – *вибірку*. Розмір вибірки під час цього залежить від різноманітності об’єктів, представлених в *генеральній сукупності* (англ. Population) – всій сукупності досліджуваних об’єктів.

Таким чином *вибірка* (англ. Sample) – це частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та характеристики генеральної сукупності. У вибірці повинні бути представлені різні комбінації і елементи генеральної сукупності.

Числові характеристики генеральної сукупності називають *параметрами*, а числові характеристики вибірки – *статистиками*.

Досить часто дослідження ґрунтуються на *гіпотезах*, які перевіряються за допомогою даних. *Гіпотеза* – це припущення щодо параметрів сукупності об’єктів, яке повинно бути перевірено на її частині. Гіпотеза є частково обґрунтованою закономірністю знань, що служить або для зв’язку між різними емпіричними фактами, або для пояснення факту групи фактів.

Вимірювання

У процесі підготовки даних вимірюється характеристики об’єкту. *Вимірювання* – процес присвоєння чисел характеристикам досліджуваних об’єктів згідно з визначеним правилом – шкалою.

Зазвичай виділяють п’ять типів шкал: номінальна, порядкова, інтервальна, відносна і дихотомічна.

Номінальна шкала (англ. Nominal Scale) містить тільки категорії; дані в ній не можуть бути впорядковані, над ними неможливо проводити арифметичні дії. Номінальна шкала складається з назв, категорій, імен для класифікації і сортування об’єктів або

спостережень за певною ознакою. Для цієї шкали можна застосовувати лише прості операції порівняння: дорівнює (=) та не дорівнює ($\neq$). Прикладами номінальної шкали є: назви факультетів, прізвища тощо.

Порядкова шкала (англ. Ordinal Scale) характеризується тим, що в ній числа привласнюють об'єктам для позначення відносної позиції об'єктів, але не величини відмінностей між ними, тобто вимірювання в порядковій шкалі містять інформацію тільки про порядок величин, але не відповідає на питання типу “наскільки одна величина більша за іншу”. Для цієї шкали можна застосовувати лише такі операції, як: дорівнює (=), не дорівнює ($\neq$), більше ($>$) та менше ($<$). Приклади порядкової шкали: освіта (початкова, середня, вища), вчене звання (доцент, старший дослідник, професор) тощо.

В *інтервальної шкалі* (англ. Interval Scale) різниці між значеннями можуть бути обчислені, проте їхні відношення не мають сенсу. Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної і порядкової шкал, а також дозволяє визначити кількісну зміну ознаки. Інтервальна шкала, на відміну від дискретних номінальної і порядкової, є неперервною. Вона дозволяє здійснювати точні вимірювання ознаки і виконувати арифметичні операції додавання, віднімання, множення, ділення. Для цієї шкали можна застосовувати лише такі операції, як дорівнює (=), не дорівнює ($\neq$), більше ($>$), менше ($<$), операції додавання (+) і віднімання (-). Прикладом такої шкали є температура повітря: можна говорити, що -10°C на 20 градусів менше, ніж $+20^{\circ}\text{C}$, проте порівняння у скільки разів менше не має сенсу.

Відносна шкала (англ. Ratio Scale) характеризується існуванням певної точки відліку, що наповнює сенсом відношення між значеннями. Для цієї шкали можна застосовувати такі операції, як дорівнює (=), не дорівнює ($\neq$), більше ($>$), менше ($<$), операції додавання (+) і віднімання (-), множення (*) та ділення (/). Прикладом такої шкали є заробітна плата.

Дихотомічна шкала (англ. Dichotomous Scale) містить тільки дві категорії. Приклад такої шкали: стать (чоловіча і жіноча). Це частковий випадок категоріальної шкали.

Самостійна робота №1. Надбудова інтелектуального аналізу даних для Microsoft Office

Одним з можливих варіантів проведення ІАД засобами SQL Server випуску 2012 – використання надбудов для пакета Microsoft Office 2010. У цьому випадку, джерелом даних для аналізу може

служити електронна таблиця Excel [6]. Дані передаються на SQL Server 2012, там обробляються, а результати повертаються Excel для відображення.

Для використання подібної «зв'язки», вам повинен бути доступний Microsoft SQL Server 2012 року в одній з версій, що підтримують інструменти Data Mining (Enterprise, Developer або с деякими обмеженнями – Standard), Microsoft Office 2010 в версії Professional або більш старшої [7].

Особливих складнощів процес встановлення додаткових програм не викликає. Єдине, що хочеться відзначити, за замовчуванням пропонується встановлювати не всі компоненти. Але для виконання подальших самостійних робіт краще встановити усі компоненти (рис. 1.1).

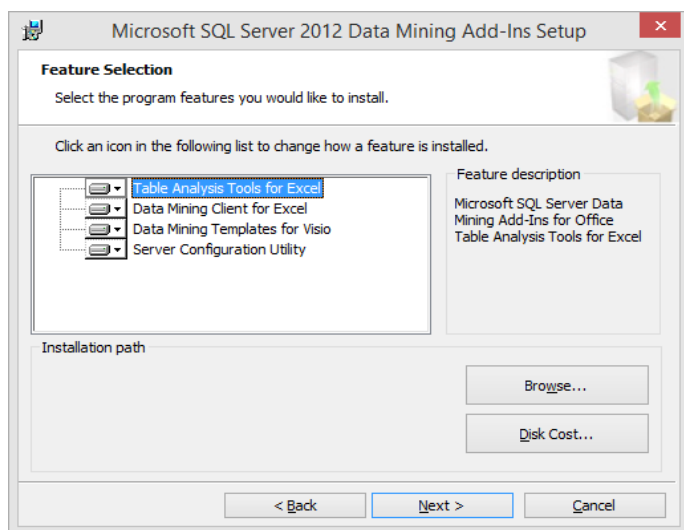


Рис. 1.1 – Вікно вибору компонентів Microsoft SQL Server 2012 Data Mining

Наступний крок – конфігурація Microsoft SQL Server для роботи з надбудовами. Для цього використовується майстер «Приступаючи до роботи» (**Getting Started**), що запускається з головного меню (рис. 1.2). Для того, щоб виконати конфігурацію Microsoft SQL Server 2012 треба мати права адміністратора.

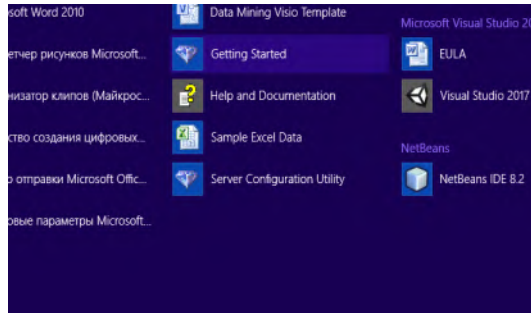


Рис. 1.2 – Вікно запуску майстра «Початок роботи»

На першому кроці майстер пропонує вибрати, завантажити чи пробну версію Microsoft SQL Server, конфігурувати існуючий екземпляр сервера, де у користувача адміністраторські права, або використовувати сервер, на якому користувач не є адміністратором (рис. 1.3).

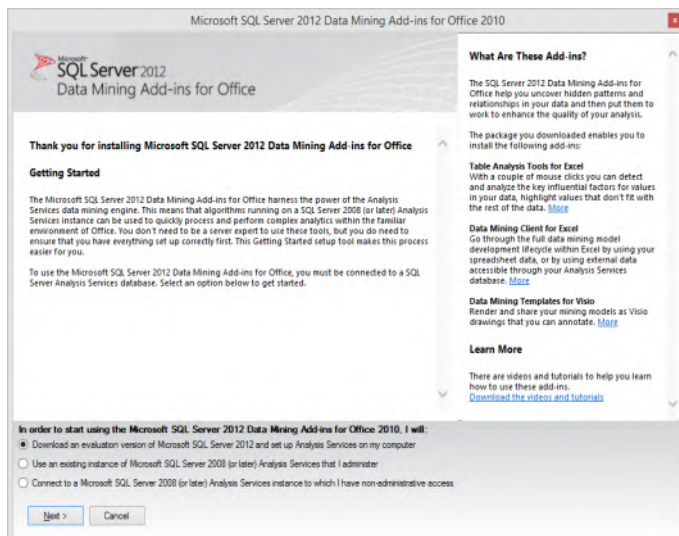


Рис. 1.3 – Вікно вибору сервера баз даних SQL Server 2012

Розглянемо варіант 2, під час вибору якого майстер покаже вікно з посиланням на інструмент «Засіб налаштування сервера». Його також можна запустити з меню **Пуск → Надбудови ІАД → Засіб налаштування сервера** (рис. 1.4).



Рис. 1.4 – Вікно налаштування сервера SQL Server 2012

Наступне вікно пропонує вибрати конфігурується сервер (рис. 1.5). За замовчуванням стоїть **localhost**, що відповідає неіменованому примірнику Microsoft SQL Server, встановленому на той же комп'ютер, на якому запущено «засіб настройки». Якщо це не так, треба вказати ім'я сервера або для іменованого примірника <ім'я сервера> \ <ім'я екземпляра>.

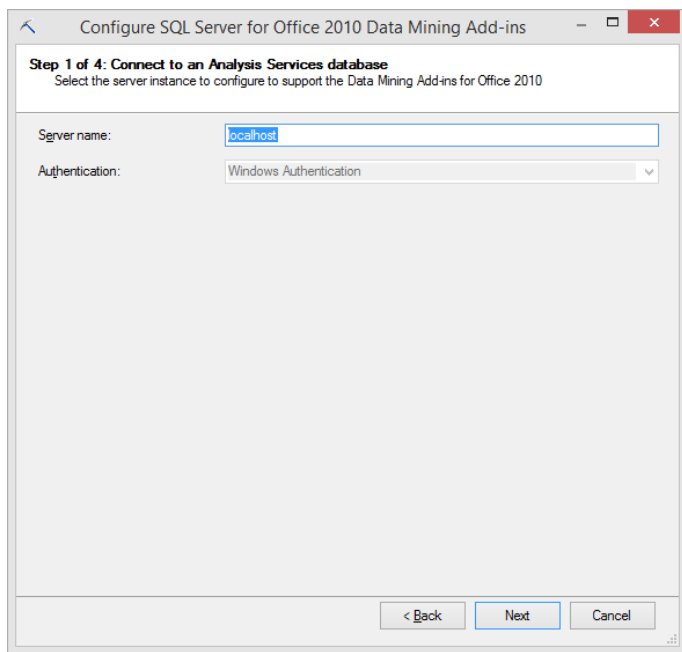


Рис. 1.5 – Вікно вибору назви сервера SQL Server 2012

У вікні, що представлено на рисунку 1.6, дається дозвіл на створення тимчасових моделей інтелектуального аналізу (**Allow creating temporary mining models**). Тимчасова модель відрізняється від постійної тим, що створюється тільки на час сеансу користувача. Коли користувач, який проводить аналіз за допомогою надбудов, завершить сесію (закриє Excel), модель буде видалена, але результати аналізу зберігаються в електронній таблиці. Постійна модель автоматично не видаляється, зберігається на сервері, і до роботи з нею можна повернутися.

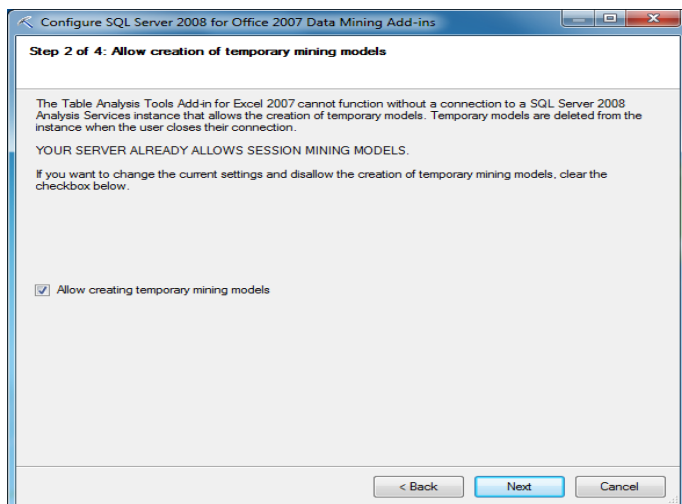


Рис. 1.6 – Вікно надання дозволу для створення тимчасових моделей інтелектуального аналізу даних

Після цього пропонується створити нову базу даних аналітичних служб або вибрати для роботи існуючу (рис. 1.7).

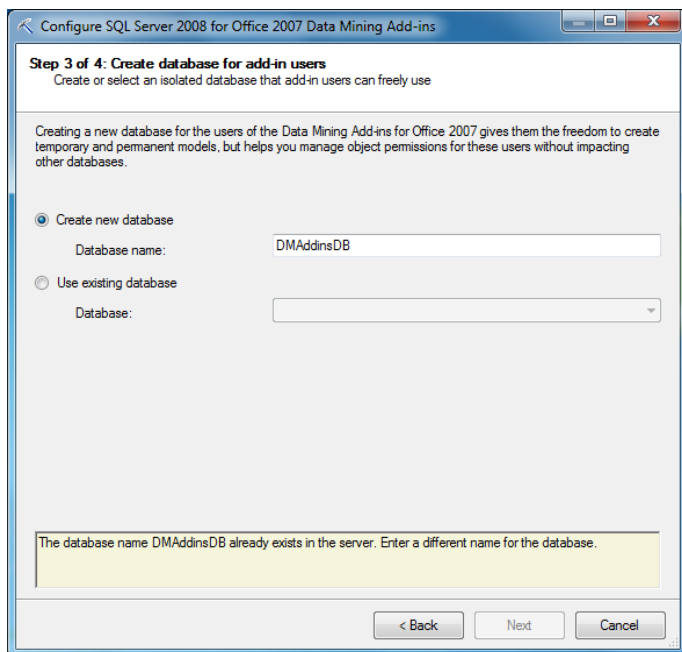


Рис. 1.8 – Вікно створення бази даних SQL Server 2012 з аналітичними службами

У вікні, представленому на рисунку 1.9, можна додати користувачів до списку адміністраторів створеної бази даних. Це потрібно для створення на сервері постійних моделей. Якщо використовувати тільки тимчасові моделі, права адміністратора користувачеві необов'язкові.

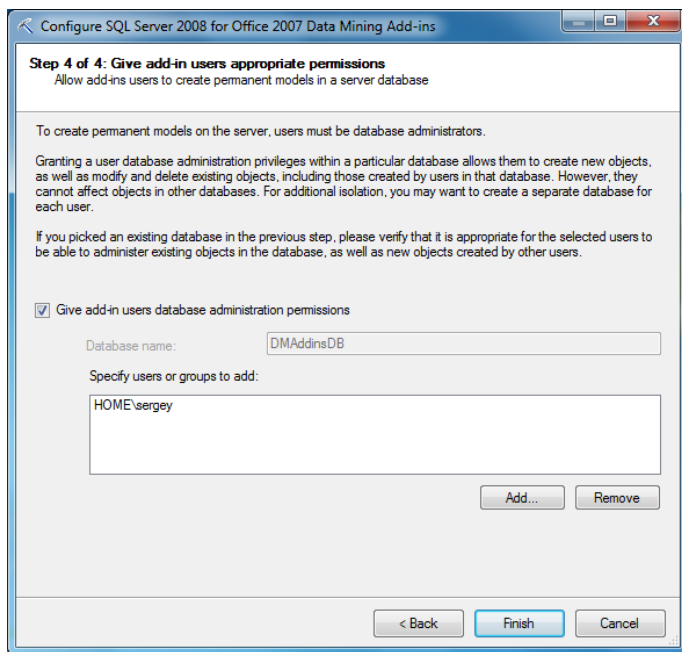


Рис. 1.9 – Вікно додавання користувачів до списку адміністраторів бази даних

Після закінчення настройки, можна відкрити Excel (а під час використання майстра «Пристаюючи до роботи», він буде запущений автоматично з документом «Зразки даних ...») і протестувати підключення до сервера. Для цього треба перейти на вкладку **Data Mining** і в розділі **Connection** (підкреслять на рисунку 1.9) натиснути кнопку **DMAddinsDB**. З'явиться вікно, яке відображає налаштовані з'єднання. Кнопка **Test Connection** дозволяє протестувати своє з'єднання. Якщо налаштованого з'єднання немає, і кнопка **DMAddinsDB** виглядає як на рисунку 1.10, то потрібно створити нове з'єднання, вибравши у вікні **Analysis Services Connection** (рис. 1.11) кнопку **New**.

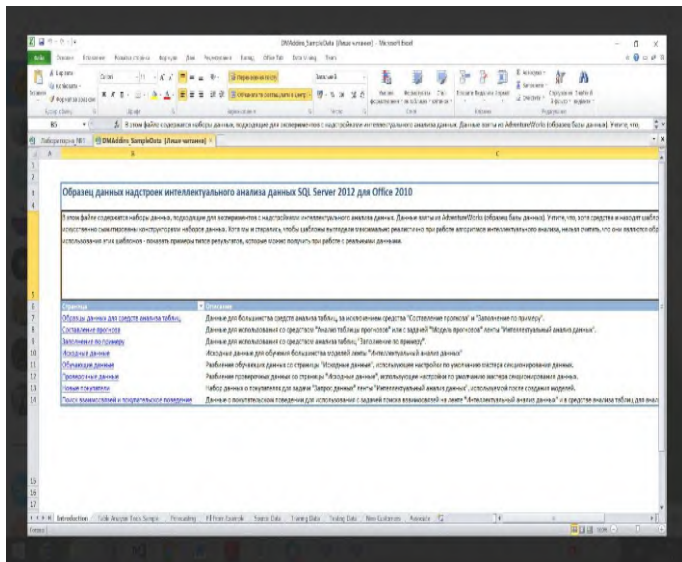


Рис. 1.10 – Вікно з вкладкою аналітичної служби Data Mining

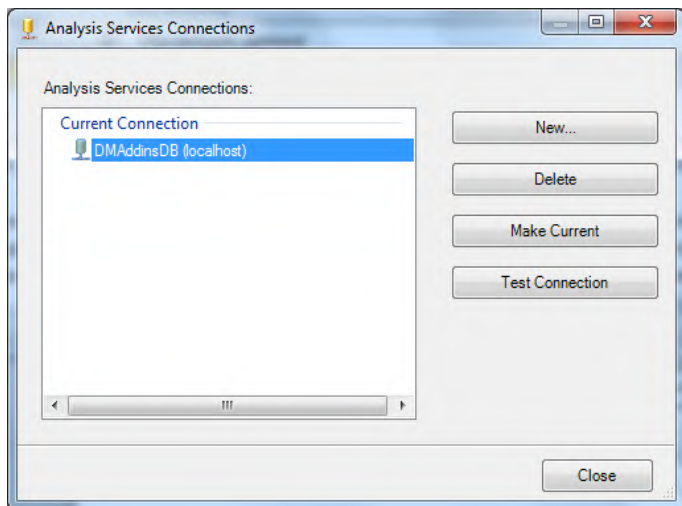


Рис. 1.11 – Вікно налаштування з'єднання із сервером SQL Server 2012

Під час створення нового підключення треба вказати сервер, до якого плануєте підключатися, і в розділі **Catalog name**

рекомендується явно вказати базу даних, з якої буде працювати надбудови. Коли з'єднання створено і перевірено, можна починати роботу (рис. 1.12).

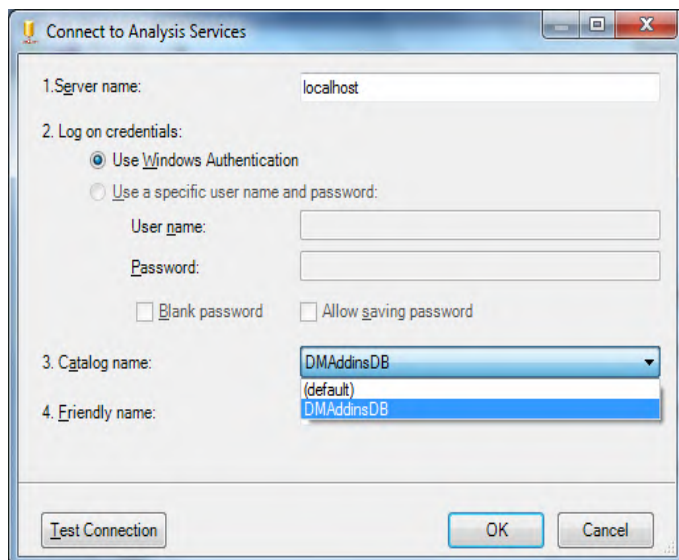


Рис. 1.12 – Вікно створення нового підключення з сервером SQL Server 2012

Якщо ви плануєте працювати з власними даними, необхідно враховувати, що інструменти інтелектуального аналізу таблиць працюють з даними, відформатовані у вигляді таблиці. Тому ваші дані в Excel потрібно виділити і вибрати «Форматувати як таблицю» (рис. 1.13). Після цього треба вибрати стиль таблиці і вказати заголовок. Вкладка **Analyze** з інструментами **Table Analysis Tools** з'явиться під час натискання в області таблиці (рис. 1.14).

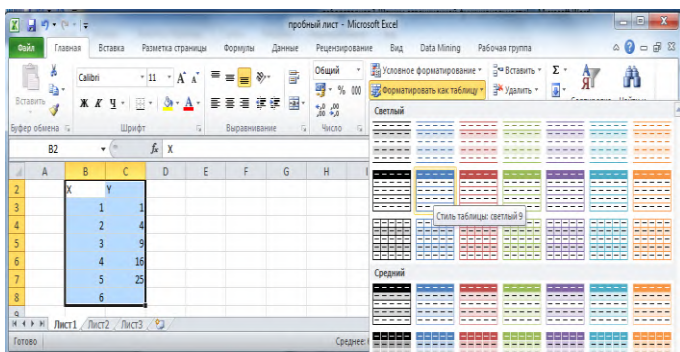


Рис. 1.13 – Вікно форматування підготовлених даних у MS Office 2010

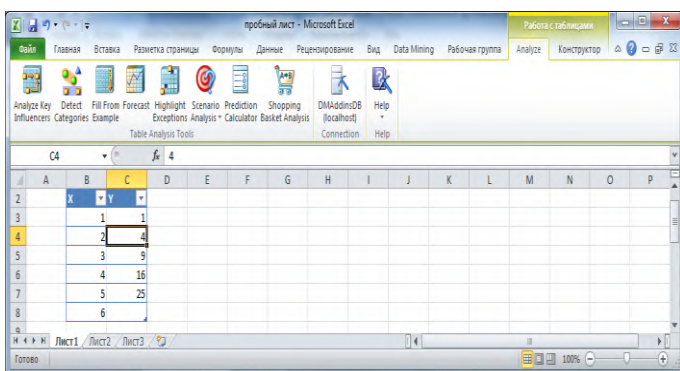


Рис. 1.14 – Вікно з інструментами інтелектуального аналізу даних

Завдання 1. Встановіть надбудови інтелектуального аналізу даних для Microsoft Office 2007. Виконайте необхідну конфігурацію MS SQL Server 2012 (2008 R2) для роботи з надбудовами. Створіть і протестуйте підключення.

Завдання 2. Підготовлений набір даних (для прикладу, можна взяти наведений на рис. 1.14) отформатуйте як таблицю. Переконайтеся, що ви можете отримати доступ до вкладки з інструментами інтелектуального аналізу таблиць.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Якою є мета інтелектуального аналізу даних?
2. Наведіть визначення інтелектуального аналізу даних.

3. За допомогою програмних продуктів яких класів може проводитися інтелектуальний аналіз даних?
4. Що досліджується в ході інтелектуального аналізу даних?
5. Перерахуйте основні задачі інтелектуального аналізу даних.
6. Дайте визначення поняттю даних. Які типи даних ви знаєте?
7. Що собою являє об'єкт та атрибут в табличних даних?
8. Які методи застосовуються в Data Mining? Чим відрізняються статистичні та кібернетичні методи Data Mining?
9. Що собою являє попередній аналіз даних?
10. Які властивості методів Data Mining ви знаєте?

ЛІТЕРАТУРА

1. Witten I.H. Data Mining: Practical Machine Learning Tools and Techniques / I.H. Witten, E.Frank. – Morgan Kaufmann, 2005. – 558 p.
2. Чубукова И.А. Data Mining / И.А. Чубукова. – М.: Интернет-университет информационных технологий, Бином. Лаборатория знаний, 2008 – 382 с.
3. Lohr S. Data-Ism: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else / S. Lohr. – HarperBusiness, 2015. – 489 p.
4. Data Mining History and Current Advances [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: https://www.sas.com/en_us/insights/analytics/data-mining.html.
5. Бармак О.В. Інтелектуальний аналіз даних. Методичні вказівки до самостійного вивчення курсу «Інтелектуальний аналіз даних» для студентів напрямку підготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.
6. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.
7. Службы SQL Server Analysis Services – интеллектуальный анализ данных. – [Електронний ресурс]. – Режим доступу: <https://docs.Microsoft.com/ru-ru/sql/analysis-services/analysis-services-tutorials-ssas?view=sql-server-2017>

РОЗДІЛ 2.

ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В СКБД MICROSOFT SQL SERVER 2012

В процесі інтелектуального аналізу даних проводиться дослідження множини об'єктів (або варіантів). Здебільшого його можна подати у вигляді таблиці, кожний рядок якої відповідає одному з варіантів, а в стовпцях містяться значення параметрів, які його характеризують. *Залежна змінна* – параметр, значення якого розглядається як залежне від інших параметрів (*незалежних змінних*) [1]. Цю залежність і необхідно визначити, використовуючи методи ІАД.

2.1 Задачі інтелектуального аналізу даних

Розглянемо основні задачі інтелектуального аналізу даних.

Задача класифікації полягає в тому, що для кожного варіанту визначається категорія або клас, якому він належить. Прикладом задачі класифікації є кредитоспроможність потенційного позичальника: класами виступають «кредитоспроможний» і «некредитоспроможний». Необхідно зазначити, що для виконання завдання потрібно, щоби множина класів була відома заздалегідь і була б кінцевою й розрахунковою. Також виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних – класи; за цими ознаками новий об'єкт можна віднести до того або іншого класу. Для розв'язання задач класифікації можуть використовуватися методи: найближчий сусід (Nearest Neighbor); k -найближчий сусід (k -Nearest Neighbor); баєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (Neural Networks) [2-3].

Задача регресії де в чому схожа з задачею класифікації, але під час її рішення проводиться пошук шаблонів для визначення неперервного числового значення [4].

Задача класифікації й регресії розв'язується у два етапи. На першому етапі виділяється навчальна вибірка. До неї входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. На підставі навчальної вибірки будується модель визначення значення залежної змінної. Її часто називають функцією класифікації або регресії. Для отримання максиміально точної функції до навчальної вибірки висуваються такі основні вимоги:

– кількість об'єктів, що входять до вибірки, має бути досить великою;

– до вибірки мають входити об’єкти, що представляють усі можливі класи в задачі класифікації або всю область значень у задачі регресії;

– для кожного класу в задачі класифікації або кожного інтервалу області значень у задачі регресії вибірка має містити достатню кількість об’єктів.

На другому етапі побудовану модель застосовують до об’єктів аналізу. Задача класифікації й регресії має геометричну інтерпретацію.

Окремо виділяється **задача прогнозування** нових значень на підставі наявних значень числової послідовності (або декількох послідовностей, між значеннями в яких спостерігається кореляція) [5]. Водночас можуть враховуватися наявні тенденції (тренди), сезонність, інші чинники. Класичним прикладом є прогнозування цін акцій на біржі.

Тут потрібно зробити невеликий відступ. За способом розв’язання задачі інтелектуального аналізу можна розділити на два класи: *навчання з вчителем* (від англ. Supervised learning) і *навчання без вчителя* (від англ. Unsupervised learning). У першому випадку, потрібен навчальний набір даних, на якому створюється й навчається модель інтелектуального аналізу. Готова модель тестується і згодом використовується для передбачення значень у нових наборах даних. Іноді в цьому ж випадку говорять про **керовані алгоритми інтелектуального аналізу**. Задачі класифікації й регресії належать саме до цього типу.

В іншому випадку, метою є виявлення закономірностей в наявному наборі даних. За такої умови навчальна вибірка не потрібна. У якості прикладу можна навести задачу аналізу споживчого кошика, коли під час дослідження виявляються товари, які найчастіше купуються разом [6]. До цього ж класу належить **задача кластеризації**.

Також можна говорити про класифікацію задач ІАД за призначенням, відповідно до якої вони діляться на *описові* (від англ. Descriptive) і *прогнозні* (від англ. Predictive). Мета розв’язку описових завдань – краще зрозуміти досліджувані дані, виявити наявні в них закономірності, навіть якщо в інших наборах даних вони траплялися не будуть. Для прогнозних задач характерно те, що під час їхнього вирішення на підставі набору даних з відомими результатами будується модель для передбачення нових значень.

Завдання кластеризації – полягає в розподілі безлічі об’єктів на групи (кластери) схожих за параметрами [7]. під час цього, на відміну від класифікації, число кластерів і їх характеристики можуть бути заздалегідь невідомі і визначатися в ході побудови кластерів

виходячи зі ступеня близькості об'єднуються об'єктів за сукупністю параметрів.

Інша назва кластеризації – сегментація. Наприклад, інтернет-магазин може бути зацікавлений в проведенні подібного аналізу бази своїх клієнтів, для того, щоб потім сформувані спеціальні пропозиції для виділених груп, враховуючи їх особливості.

Завдання визначення взаємозв'язків, також звана завданням пошуку асоціативних правил, полягає у визначенні часто зустрічаються наборів об'єктів серед безлічі подібних наборів. Класичним прикладом є аналіз споживчого кошика, який дозволяє визначити набори товарів, які найчастіше зустрічаються в одному замовленні (або в одному чеку). Ця інформація може потім використовуватися під час розміщенні товарів у торговельному залі або під час формуванні спеціальних пропозицій для групи пов'язаних товарів.

Задача **визначення взаємозв'язків**, яку ще називають *задачею пошуку асоціативних правил*, полягає у визначенні наборів об'єктів, що часто трапляються серед множини подібних наборів [8]. Класичним прикладом є аналіз споживчого кошика, який дозволяє визначити набори товарів, які найчастіше трапляються в одному замовленні (або в одному чеку). Ця інформація може потім використовуватися у разі розміщення товарів у торговельному залі або під час формування спеціальних пропозицій для групи пов'язаних товарів. Дана задача також належить до класу «навчання без учителя».

Аналіз послідовностей або **сіквенціальний аналіз** одними авторами розглядається, як варіант попередньої задачі, іншими – виділяється окремо [8]. Метою аналізу послідовностей є виявлення закономірностей у послідовностях подій. Подібна інформація дозволяє, наприклад, попередити збій у роботі інформаційної системи, отримавши сигнал про настання події, яка часто передую збою подібного типу. Інший приклад застосування – аналіз послідовності переходів сторінками користувачів web-сайтів.

Аналіз відхилень дозволяє відшукати серед множини подій ті, які істотно відрізняються від норми [9]. Відхилення може сигналізувати про якусь незвичну подію (несподіваний результат експерименту, шахрайська операція банківськими картами тощо) або, наприклад, про помилку введення даних оператором [10].

Зазначені вище задачі поділяються на описові і предиктивні.

Описові, або дескриптивні (з англ. Descriptive), задачі пов'язані з поліпшенням розуміння аналізованих даних. Ключовий

момент у таких моделях – простота і прозорість результатів для сприйняття людиною. До такого типу задач належать кластеризація й пошук асоціативних правил.

Рішення *предиктивних* (з англ. Predictive), або *прогнозуючих*, задач поділяється на два етапи. На першому етапі на підставі набору даних із відомими результатами будується модель. На другому етапі вона використовується для прогнозу результатів на підставі нових наборів даних. Вимагається, щоби побудовані моделі працювали максирисьно точно. До цього типу задач відносять задачі класифікації й регресії. Сюди можна віднести й задачу пошуку асоціативних правил, якщо результати її рішення можуть бути використані для прогнозу появи деяких подій.

У табл. 2.1. наведено приклади задач ІАД з різних областей [11].

Таблиця 2.1 – Методи інтелектуального аналізу даних за сферами застосування

Задача	Інформаційні технології	Торгівля	Фінансова сфера
Класифікація			Оцінка кредитоздатності
Регресія			Оцінка допустимого кредитного ліміту
Прогнозування		Прогнозування продажів	Прогнозування цін акцій
Кластеризація		Сегментація клієнтів	Сегментація клієнтів
Визначення взаємозв'язків		Аналіз споживчого кошики	
Аналіз послідовностей	Аналіз переходів по сторінкам web-сайту		
Аналіз відхилень	Виявлення вторгнень в інформаційні системи		Виявлення шахрайства з банківськими картками

Більш детально деякі з перерахованих завдань будуть розглянуті в наступних розділах. Крім того, можна рекомендувати звернутися до додаткових джерел [12-13].

2.2 Інтелектуальний аналіз даних засобами СКБД Microsoft SQL Server 2012

Розглянемо реалізацію засобів ІАД в СКБД на прикладі Microsoft SQL Server 2012 для Microsoft Excel 2010. Як було зазначено в підрозділі 1., вирішуються ці завдання службами Analysis Services. На рисунку 2.1 схематично подані інтегровані служби та компоненти СКБД Microsoft SQL Server 2012 і виділена підсистема ІАД [14].

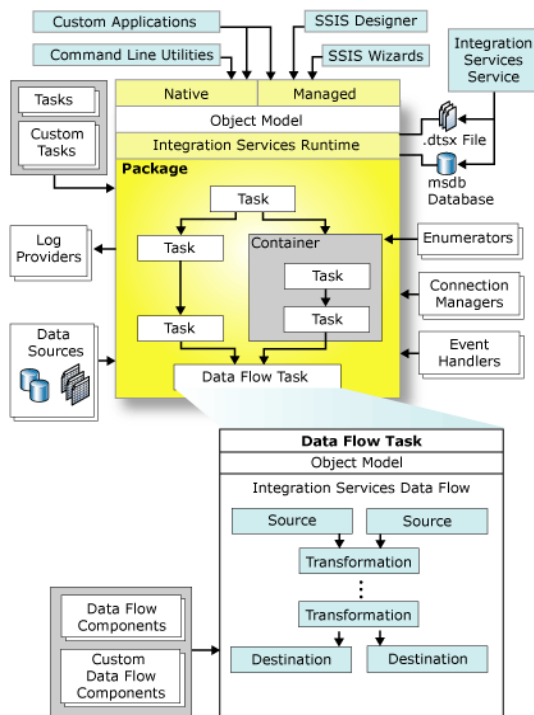


Рис. 2.1 – Інтегровані служби СКБД Microsoft SQL Server 2012

Служби Analysis Services надають наступні засоби для створення рішень з ІАД:

- 1) Набір стандартних алгоритмів ІАД.

2) Конструктор ІАД, призначений для створення і перегляду моделей ІАД, керування ними та побудови прогнозів.

3) Мову розширень ІАД (Data Mining eXtensions to SQL).

Для роботи з наданими засобами інтелектуального аналізу використовується інструментальне середовище Business Intelligence Development Studio, скорочено BI Dev Studio.

Також Microsoft SQL Server 20012 підтримують створення, управління та використання моделей ІАД з Microsoft Excel за допомогою надбудов ІАД SQL Server 2012 для Office 2010.

Докладно з перерахованими алгоритмами ми познайомимося в наступних розділах. У табл. 2.2, наведені деякі приклади використання ІАД і відповідні їм алгоритми [15].

Таблиця 2.2 – Приклади використання методів інтелектуального аналізу даних до прикладних задач [10]

Задача і приклад	Відповідні алгоритми
Прогнозування дискретного атрибуту. <i>Наприклад, чи купить одержувач цільової розсилки певний продукт.</i>	Алгоритм дерева прийняття рішень. Спрощений алгоритм Баеса. Алгоритм кластеризації. Алгоритм нейронної мережі.
Прогнозування безперервного атрибуту. <i>Наприклад, прогноз продажів на наступний рік.</i>	Алгоритм дерева прийняття рішень. Алгоритм часових рядів.
Прогнозування послідовності. <i>Наприклад, аналіз маршруту переміщення по веб-вузлу компанії</i>	Алгоритм кластеризації послідовностей
Знаходження груп спільних елементів у транзакціях. <i>Наприклад, використання аналізу купівельної поведінки для пропозиції додаткових продуктів замовнику.</i>	Алгоритм взаємозв'язків. Алгоритм дерева прийняття рішень.
Знаходження груп схожих елементів. <i>Наприклад, розбиття демографічних даних на групи для кращого розуміння зв'язків між атрибутами.</i>	Алгоритм кластеризації. Алгоритм кластеризації послідовностей.

На закінчення розглянемо схему взаємодії аналітичних служб SQL Server із зовнішніми програмами (рис. 2.2). На рисунку видно, що зовнішні програми, використовуючи різноманітні засоби (ADODATA MINING.Net тощо) і прокол XMLA (XML for Analysis) «поверх» протоколу TCP або HTTP (характерно для web-рішень) можуть взаємодіяти з аналітичними службами. Водночас, в залежності від типу запиту, задіюється або підсистема OLAP, або підсистема ІАД. Оброблятися запити інтелектуального аналізу можуть за допомогою як стандартних алгоритмів Microsoft, так і алгоритмів розробки третіх фірм. Результат за допомогою XMLA передається назад застосуванню.

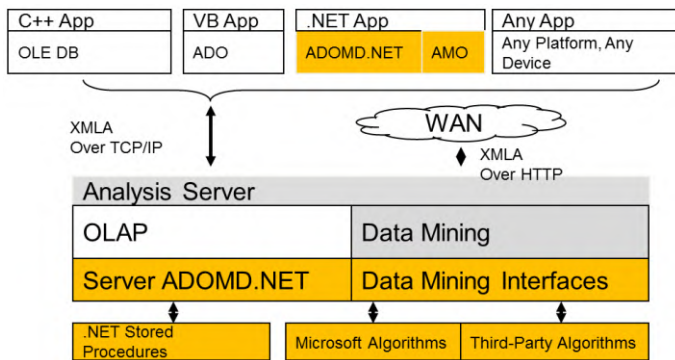


Рис. 2.2 – Схема взаємодії аналітичних служб Microsoft SQL Server 2012 із зовнішнім програмним забезпеченням

Під час вирішення передбачувальних задач інтелектуального аналізу якості прогнозу, що видається моделлю, можна оцінити на тестувальному наборі даних, для якого відомі значення прогнозованого параметра. В Microsoft SQL Server 2012 служби Analysis Services надають засоби, що спрощують поділ даних на навчальний і тестувальний набори. Таке секціонування можна виконати автоматично під час побудови моделі ІАД. Точність прогнозів можна перевірити за допомогою таких засобів, як діаграма точності прогнозів і матриця класифікації.

Самостійна робота №2. Використання інструментів «Analyze Key Influencers» та «Detect Categories»

Розпочнемо безпосереднє вивчення інструментів ІАД. До складу пакета надбудов для Microsoft Office 2007 входить електронна

таблиця із зразками даних. Її можна відкрити з меню **Пуск -> Надбудови ІАД Microsoft SQL Server 2012**. Але переведено вміст файлу тільки частково – перша сторінка зі змістом і деякі заголовки. Тому в роботі буде використовуватися локалізований набір даних для аналізу, доступний для скачування за адресою: <http://russianDataMiningadds.codeplex.com/>.

Завантажимо файл та відкриємо його і відформатуємо дані на аркуші «клієнти» як таблицю. Перейдіть на вкладку **Analyze** (рис. 2.3).

№	Семейный полонем	Пол	Возраст	Дети	Образование	Тип работы	Квал-фикация	Был ли авто	Расстояние до работы	Ремонт	Возраст	Прислал велосипед
1	12498	Мужской	40000	1	Бакалавр	Квалифицированный	Да	0-0.1 км	Европа	42	Нет	
2	24307	Мужской	30000	3	Неоконченное высшее	Офисный работник	Да	1-0.1 км	Европа	43	Нет	
3	14177	Мужской	80000	5	Неоконченное высшее	Профессионал	Нет	2-2.5 км	Европа	60	Нет	
4	24382	Одинокий(ая)	70000	0	Бакалавр	Профессионал	Да	1-5-10 км	Россия	41	Да	
5	25587	Одинокий(ая)	30000	0	Бакалавр	Офисный работник	Нет	0-0.1 км	Европа	36	Да	
6	11507	Мужской	10000	2	Неоконченное высшее	Физический труд	Да	0-0.1 км	Европа	50	Нет	
7	27974	Одинокий(ая)	160000	2	Среднее	Управление	Да	4-0.1 км	Россия	33	Да	
8	19364	Мужской	40000	1	Бакалавр	Квалифицированный	Да	0-0.1 км	Европа	43	Да	

Рис. 2.3 – Вікно із вхідними даними

Таблиця аналізу даних містить дані фірми, що продає велосипеди. У ній зібрано інформацію про клієнтів (ідентифікатор, сімейний стан, стать та ін.) і зазначено, придбав клієнт велосипед чи ні.

Аналіз ключових факторів впливу

Інструмент **Analyze Key Influencers** дозволяє дізнатися залежність між параметрами. Для цього важливо правильно визначити залежність параметрів між собою. Власне кажучи, в цьому і полягає майстерність аналітика, що ґрунтується на знаннях предметної області та використаних методів Data Mining.

У зв'язку з тим, що будуть оцінюватися степені взаємного впливу параметрів між собою, необхідно одразу позбутися від розгляду повністю незалежних і залежних атрибутів. Наприклад, ми хочемо оцінити вплив різних факторів на рівень заробітної плати людини. Якщо, є зміна, що містить в собі ідентифікатор (порядковий номер запису в таблиці, номер паспорта), то цей атрибут потрібно забрати з розгляду, оскільки порядковий номер не впливає на досліджуваний параметр – Рівень заробітної плати. Нехай розглядатимемо показник заробітної плати за місяць і за рік. Дані

значення показників завжди взаємопов'язані між собою, знайти залежність одного від іншого засобами Data Mining не має жодного значення, а наявність сильної залежності приховується впливом інших факторів на ці фактори, які необхідно виявити.

Далі необхідно визначити від якого параметру залежить рішення клієнта про покупку велосипеда. Натиснемо на кнопку **Analyze Key Influencers** та вкажемо в якості цільового стовпця, стовпець «Придбати велосипед» (рис. 2.4). Перейдемо за посиланням “Choose columns to be used for analysis”, щоб вказати параметри впливу, які необхідно оцінити (рис. 2.5). Далі скасуємо мітку навпроти “ID” та “Придбати велосипед” (хоча останню мітку не обов'язково скасовувати).

Після запуску процедури аналізу (натиснути кнопку **Run**, що зображена на рисунку 2.4) буде згенеровано звіт про всі фактори впливу та запропоновано сформувати додатковий звіт для порівняння.

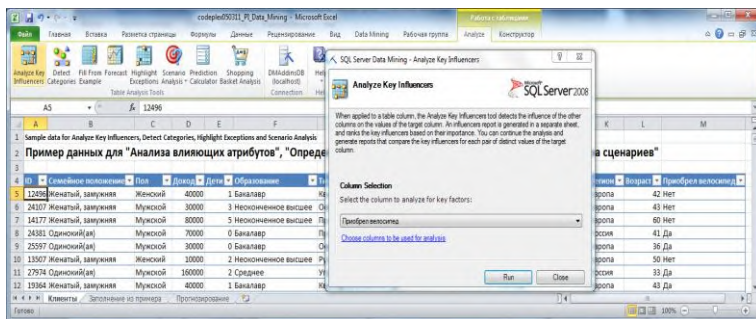


Рис. 2.4 – Вікно вибору залежного параметра для аналізу

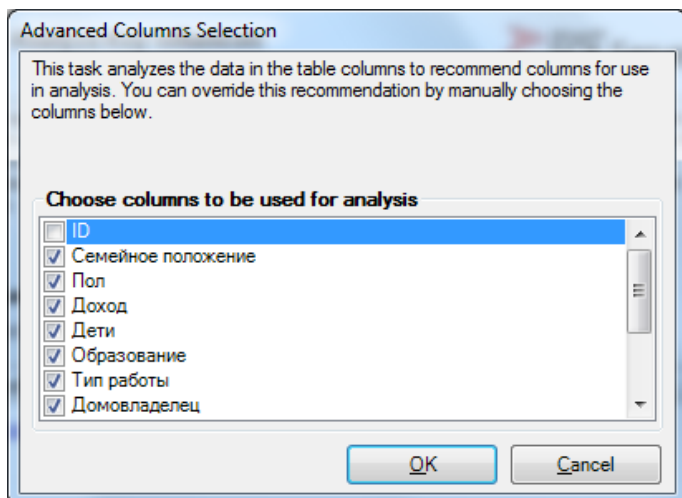


Рис. 2.5 – Вікно вибору параметрів для дослідження залежностей

В головному звіті вказано стовпець (Column), його значення (Value), значення цільового стовпця з яким воно зв'язується (Favors) та рівень впливу (Relative Impact), рівень впливу можна оцінити в інтервалі [0,100] балів. З звіту, що представлено на рисунку 2.6 видно, що на рішення відхилити покупку велосипеда в найбільшій мірі впливає наявність 2-х автомобілів.

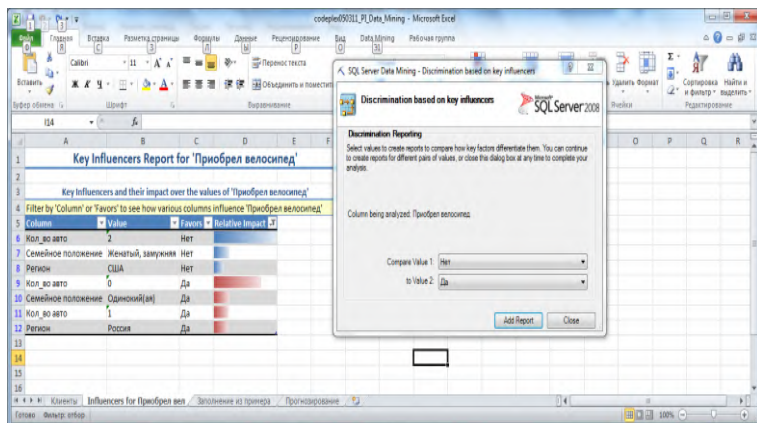


Рис. 2.6 – Основна таблиця результату

Також необхідно розуміти оцінку 100 балів як випадок у якому в 100% подій, власники 2-х авто, відхиляли необхідність в купівлі велосипеда (перегляньте набір даних, там є подія «2 машини – велосипед придбано», але таких випадків менше 4%).

Інший за рівнем впливу на відмову від покупки є наступний фактор – “Сімейний стан” = “одружений, одружена”. Найбільший вплив на позитивне рішення про покупку велосипеда є відсутність у клієнта автомобіля. Якщо додати порівняльний звіт до двох вибраних значень (див. рис. 2.7), то можна побачити різницю вибору між одним значенням цільового стовпця та іншим значенням цього ж стовпця.

В даному прикладі відбулося перегрупування вихідної таблиці залежних факторів, оскільки таких факторів всього два. В інших випадках додаткова таблиця дозволяє провести детальне порівняння двох обраних факторів. Як зазначається в джерелі [1], якщо цільовий або інший стовпці, що обробляються інструментом **Analyze Key Influencers** містять багато різних числових значень, то тоді проводиться дискретизація. Весь інтервал значень ділиться на декілька інтервалів, кожен з яких розглядається як одне з можливих значень (наприклад, замість точного значення 2,5 отримаємо інтервал [2;3]). Згенерована таблиця даних буде доступною лише в тому випадку, якщо відкрити даний файл в Excel на іншому комп’ютері (без підключення до аналітичних служб SQL Server) (рис. 2.7).

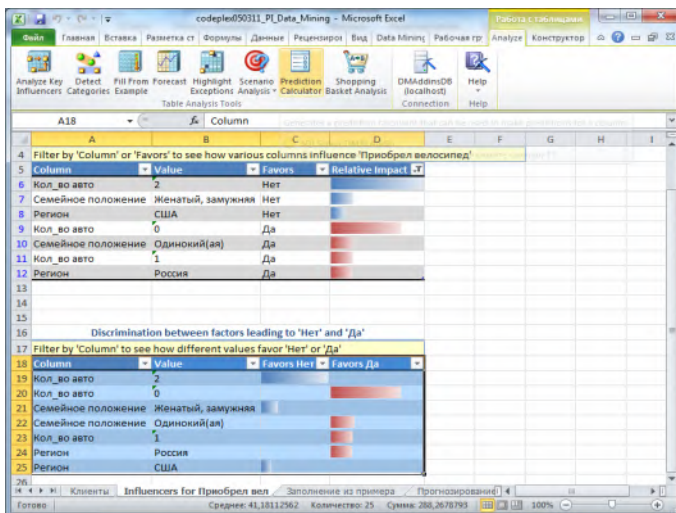


Рис. 2.7 – Дві таблиці для порівняння результатів

Щоб повернутися до початкових даних необхідно видалити листи з вмістом згенерованих таблиць залежностей.

Знаходження категорій

Інструмент **Detect Categories** дозволяє розв'язати задачу кластеризації, тобто задачу розділення всієї вхідної сукупності даних на групи об'єктів з спільними характеристиками. Подібна задача також називається задачею сегментації, оскільки, вхідна сукупність ділиться на окремі групи сегменти.

В вхідному наборі даних для Excel описано множину клієнтів. Необхідно розділити їх на невелику кількість груп з певними характеристиками об'єктів. У зв'язку з тим, що в процесі роботи інструмент додає данні в вихідну таблицю, рекомендується перед початком роботи створити її копію (рис. 2.8).

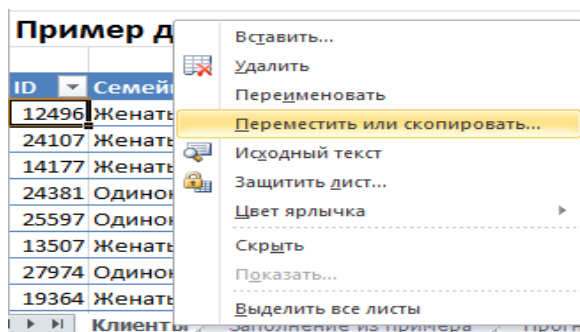


Рис. 2.8 – Створення копії вхідної таблиці даних

Далі необхідно натиснути кнопку **Detect Categories** та налаштувати параметри (рис. 2.9). Необхідно звернути увагу на атрибут ID, значення якого неважливо враховувати в процесі аналізу. В даному випадку, решту атрибутів можна позначати виходячи з предметної області.

Окрім, вибору параметрів, можна явно задати число категорій(або зупинити налаштування за замовчуванням – автоматичне визначення). Також за умовчуванням встановлено флажок “Append a Category column to the original Excel table”, який означає, що записів в вхідній таблиці додано стовпець з назвами категорій.

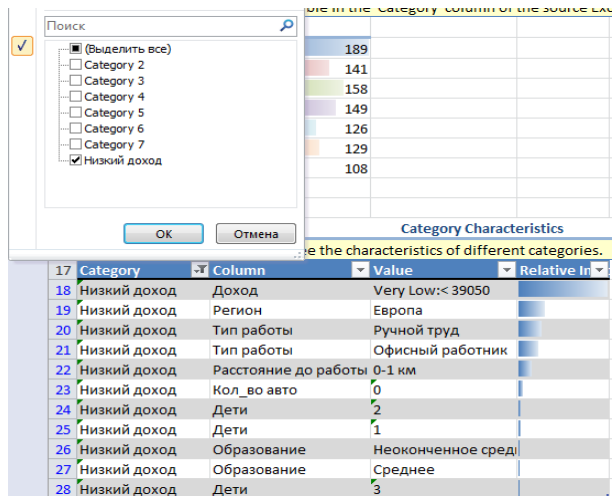


Рис. 2.11 – Опис категорій

Третій розділ – це діаграма профілів категорій. Вона відображає кількість рядків даних в кожній категорії з кожним значенням вибраних параметрів. За замовчуванням відображається тільки один параметр. Для розглянутого прикладу це вік. На рисунку 2.12 зображено для кожної категорії розподіл по віку та доходам.

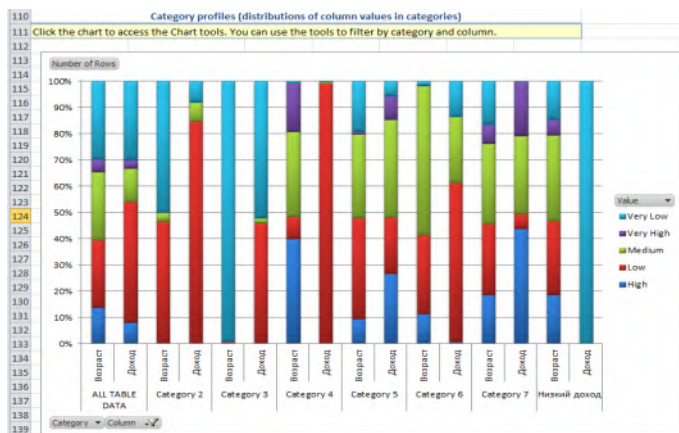


Рис. 2.12 – Діаграма категорій

Рис. 2.13 показує, що для кожного запису вхідної таблиці співставлена категорія до якої цей запис відноситься, а за допомогою фільтра можна переглянути записи що відносяться до обраної категорії.

The screenshot shows an Excel spreadsheet with the following data (rows 4-12):

Пол	Доход	Дети	Образование	Тип работы	Демографическая группа	Кол-во авто	Расстояние до работы	Регистр	Возраст	Приобретение велосипеда	Category
Женский	40000	1	Бакалавр	Квалифицированный	Да	0	0-1 км	Европа	42	Нет	Category 2
Мужской	30000	3	Неоконченное высшее	Офисный работник	Да	1	0-1 км	Европа	43	Нет	Низкий доход
Мужской	80000	5	Неоконченное высшее	Профессионал	Нет	2	2-5 км	Европа	60	Нет	Category 5
Мужской	70000	0	Бакалавр	Профессионал	Да	1	5-10 км	Россия	41	Да	Category 5
Мужской	30000	0	Бакалавр	Офисный работник	Нет	0	0-1 км	Европа	36	Да	Низкий доход
Женский	10000	2	Неоконченное высшее	Ручной труд	Да	0	1-2 км	Европа	50	Нет	Низкий доход
Мужской	160000	2	Среднее	Управление	Да	4	0-1 км	Россия	33	Да	Category 7
Мужской	40000	1	Бакалавр	Квалифицированный	Да	0	0-1 км	Европа	43	Да	Category 2

Рис. 2.13 – Співставлення категорій по записам у вхідній таблиці

Рисунок 2.13 показує, що всіх записів вихідної таблиці тепер співставлена категорія, до якої вони належать. А за допомогою фільтрів можна переглянути записи, які стосуються обраної категорії.

Завдання 1. Переименуйте категорію Category 3.

Завдання 2. Проведіть аналіз параметрів, що характеризують залишилися категорії, і дайте їм осмислені назви.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Якими компонентами MS SQL SERVER вирішуються задачі інтелектуального аналізу?
2. Які засоби надають служби Analysis Services для створення рішень з інтелектуального аналізу даних?
3. Яка інструментальна середовище в MS SQL SERVER використовується для подання засобів інтелектуального аналізу даних?
4. Дайте визначення структури інтелектуального аналізу даних.
5. Дайте визначення моделі інтелектуального аналізу даних.
6. Дайте визначення алгоритму інтелектуального аналізу даних.

7. Перерахуйте алгоритми інтелектуального аналізу даних, реалізовані в Microsoft SQL Server 2008 R2 і приведіть приклади завдань для цих алгоритмів.

8. Опишіть схему взаємодії аналітичних служб Microsoft SQL Server 2012 із зовнішнім програмним забезпеченням.

9. Для чого призначений конструктор інтелектуального аналізу даних служби Analysis Services в СКБД Microsoft SQL Server 2012?

10. Яке призначення інструментів «Analyze Key Influencers» та «Detect Categories» в СКБД Microsoft SQL Server 2012?

ЛІТЕРАТУРА

1. Барсегян А.А. Анализ данных и процессов : учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д.Тесс. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.

2. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: «БХВ-Петербург», 2004. – 336 с.

3. Baker L.D. Distributional Clustering of Words for Text Classification In ACM SIGIR 98 / L.D. Baker, A.K. McCallum. – 1998.

4. Воронцов К.В. Машинное обучение: курс лекций [Електронний ресурс] / К.В. Воронцов. – Режим доступу: [http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_\(курс_лекций,_К.В.Воронцов](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов)

5. Башмаков А.И. Интеллектуальные информационные технологии: учеб. пособ. / А.И. Башмаков, И.А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

6. Data Mining and Image Processing Toolkits. – [Електронний ресурс]. – Режим доступу: <http://Data Mining.itsc.uah.edu/adam/>

7. Дюк В. Data Mining: учебный курс / В. Дюк, А. Самойленко. – СПб: Изд. Питер, 2001. – 368 с.

8. Palmer M. Data is the New Oil / M. Palmer // CMO News, Michael Palmer, Reinventing Marketing, Technology's Impact, The Customer. – 2006. – Електронний ресурс. – Режим доступу: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

9. Акіменко В.В. Проектування СПІР на основі нечіткої логіки : навчально-методичний посібник / В.В. Акіменко, Ю.В. Загородній. – К.: Вид-во КНУ, 2007. – 94 с.

10. Бармак О.В. Інтелектуальний аналіз даних. Методичні вказівки до самостійного вивчення курсу «Інтелектуальний аналіз

даних» для студентів напрямку підготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.

11. Дюк В.А. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях / В.А. Дюк, А.В. Флегонтов, И.К. Фомина // Известия Российского государственного педагогического университета им. А.И. Герцена. Область наук: Информатика. – 2011. – С. 77–84.

12. Нікольський Ю. Дерева прийняття рішень та їхнє застосування для прогнозування діагнозу у медицині / Ю. Нікольський, Ю. Щербина, Р. Якимечко // Вісник Львівського університету. Серія прикладна математика та інформатика. – № 6, 2003. – С. 191–211.

13. Кречетов Н. Продукты для интеллектуального анализа данных / Н. Кречетов // Рынок программных средств, N14-15_97. – 1997. – С. 32–39.

14. Службы Analysis Services – видеоролики по интеллектуальному анализу данных (SQL Server 2012). URL: [http://Microsoftdn.Microsof.com/ru-ru/library/dd776389\(v=SQL.100\).aspx](http://Microsoftdn.Microsof.com/ru-ru/library/dd776389(v=SQL.100).aspx)

15. Киселев М. Средства добычи знаний в бизнесе и финансах / М. Киселев, Е. Соломатин // Открытые системы. – 1997. – № 4. – С. 41–44.

РОЗДІЛ 3.

ОСНОВНІ ЕТАПИ ПРОВЕДЕННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Розглянемо етапи проведення інтелектуального аналізу даних. На рисунку 3.1 схематично подані перераховані етапи і вказані засоби Microsoft SQL Server, за допомогою яких вони виконуються [1]. Зазначені етапи не обов'язково будуть пройдені один за іншим. Наприклад, на одному з проміжних етапів може з'ясуватися, що в поточній постановці для вирішення завдання не вистачає даних, і потрібно ще раз повернутися до першого етапу [2].

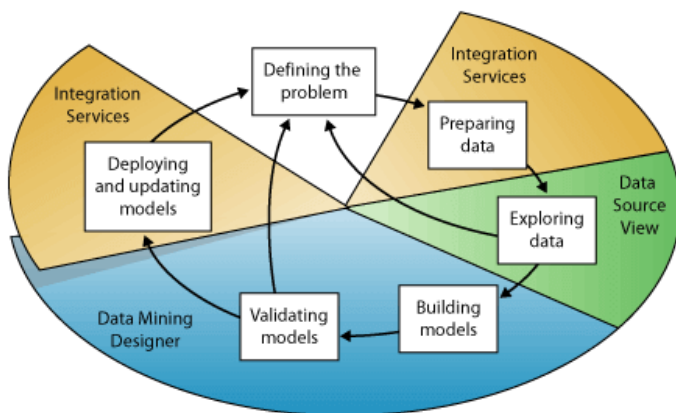


Рис. 3.1 – Основні етапи проведення інтелектуального аналізу даних

На етапі постановки задачі потрібно визначити, що є метою аналізу [3]. Зокрема, потрібно відповісти на ряд запитань, головне з яких – що саме необхідно визначити в результаті аналізу.

Виокремимо два типи задач, розв'язуваних із різною ефективністю різними методами ІАД (хоча, втім, реальні задачі дослідження даних можуть охоплювати обидва типи).

Задачі *першого типу* полягають у побудові на підставі наявних даних різних моделей, якими можна скористатися з метою прогнозування та ухвалення рішення в майбутньому, за схожої ситуації.

Задачі *другого типу* характерні тим, що наголос у них робиться на з'ясуванні сутності залежностей у множині даних, а також взаємовпливу, тобто на побудові емпіричних моделей різних систем,

які легко може сприймати людина. під час цього не так уже й важливо, щоб система добре передбачала і працювала в майбутньому, а важливо зрозуміти взаємні впливи досліджуваних закономірностей (що і чим визначається в наявному масиві даних). І навіть якщо встановлені закономірності належатимуть до специфічних особливостей саме конкретних досліджуваних даних і більше ніде не траплятимуться, але нам усе одно потрібно їх з'ясувати.

Розглянемо *головні етапи (кроки)*, характерні для будь-якого дослідження даних за допомогою методів ІАД і становлять основний цикл пошуку нового знання та його оцінювання (рис. 3.2) [4].



Рис. 3.2 – Схема інтелектуального аналізу даних і оцінювання виявленого нового знання

Перший етап полягає у зведенні даних до форми, придатної для застосування конкретних реалізацій систем ІАД. Нехай, скажімо, інформацію подано у вигляді текстів і потрібно побудувати автоматичний рубрикатор, класифікатор якихось анотацій, описів тощо. Вхідна інформація являє собою тексти в електронному вигляді, але практично жодна з наявних систем ІАД не здатна працювати безпосередньо з текстами. Щоб працювати з певним текстом, ми маємо з вихідної текстової інформації заздалегідь дістати деякі похідні параметри (наприклад, частоту появи ключових слів, середню довжину

речень, параметри, що характеризують сполучуваність тих чи інших слів у реченні тощо), тобто побудувати чіткий набір кількісних або якісних параметрів даного тексту. Ця задача найменш автоматизована в тому сенсі, що систему шуканих параметрів формує людина, хоча значення параметрів можуть обчислюватися автоматично в рамках відповідної технології первинної обробки даних. Вибравши параметри, дані можна подати у вигляді прямокутної таблиці, де кожний рядок характеризує окрему ознаку (стан, властивість) досліджуваного об'єкта, а кожний стовпець – ознаки (стани, властивості) всіх досліджуваних об'єктів. Рядки такої таблиці в теорії ІАД, як і в теорії баз даних, називають записами, а стовпці – полями.

Практично всі наявні системи ІАД працюють тільки зі щойно описаними прямокутними таблицями.

На *другому етапі* отримуємо прямокутну таблицю – «сировинний» матеріал для застосування методів ІАД, і дані, що входять до неї, необхідно передусім обробити. По-перше, таблиця може містити параметри (ознаки об'єктів), що мають однакові значення в якомусь зі стовпців. Коли б досліджувані об'єкти риси тільки такі ознаки, усі вони були б абсолютно ідентичними. Звідси випливає, що відповідні ознаки жодним чином не характеризували б досліджуваних об'єктів, а отже, їх потрібно вилучити з аналізу. Можлива й така ситуація, що деяка категоріальна ознака в усіх її записах має різні значення, через що відповідне поле не придатне для аналізу даних і його також доведеться вилучити. Нарешті може статися так, що полів буде дуже багато, і якщо ми всі їх намагатимемося досліджувати, то надто відчутно збільшиться час розрахунків, оскільки практично для всіх методів ІАД характерна сильна залежність часу розрахунків від кількості параметрів, тоді як залежність часу розрахунків від кількості записів лінійна або близька до неї.

Тому у процесі попередньої обробки даних необхідно, по-перше, розглянути множину всіх ознак, що стосуються шуканої залежності, вилучити з неї ті, які явно не придатні для подальшого дослідження, та виокремити ті, що найімовірніше ввійдуть у шукану залежність. Для цього, як правило, застосовують статистичні методи, що ґрунтуються на застосуванні кореляційного аналізу, лінійних регресій, тобто методи, що дають змогу швидко, хоча й наближено оцінити вплив одного параметра на інші.

Третій етап – безпосереднє застосування методів ІАД за різними сценаріями, що містять складні комбінації тих методів, які

допомагають аналізувати дані з різних поглядів. Власне, цей етап дослідження і називають Data Mining (добування даних).

Четвертий етап – верифікація та перевірка результатів, найчастіше здійснювані в такий спосіб. Усі наявні дані, що мають бути проаналізовані, розбивають на дві (як правило, не однакові за розміром) групи. У більшій групі даних за допомогою тих чи інших методів ІАД дістають моделі й залежності, а в меншій виконують їх перевірку. Далі за різницею в точності між результатами, здобутими для обох груп, доходять висновку щодо адекватності й статистичної значущості побудованої моделі. Існує багато інших, складніших способів верифікації (перехресна перевірка, бутстреп-аналіз тощо), які дають змогу оцінити значущість побудованих моделей без розбиття даних на дві групи.

Нарешті, на *n'ятому етапі* знання, що їх здобула людина, автоматично інтерпретуються з метою їх використання для прийняття рішень та внесення сформульованих правил і залежностей до баз знань тощо. Цей етап часто передбачає застосування методів, що є проміжними між технологією ІАД і технологією експертних систем. Від того, наскільки ефективним він буде, значною мірою залежить успіх розв'язання поставленої задачі.

Метою даного етапу є оцінка якості роботи створеної моделі перед початком її використання у виробничому середовищі. Якщо створювалося кілька моделей, то на цьому етапі робиться вибір на користь тієї, що дасть найкращий результат. На рисунку 3.2 вказано схему вибору найкращої моделі.

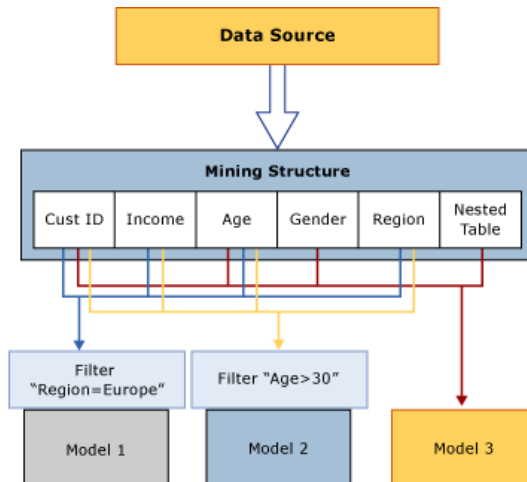


Рис. 3.2 – Схема моделей, що створені в рамках єдиної структури

Цим етапом і закінчується цикл ІАД. Остаточне оцінювання вагомості здобутого нового знання виходить за рамки аналізу, автоматизованого чи традиційного, і стає можливим тільки після впровадження на практиці рішення, прийнятого на основі такого знання. Дослідженням практичних результатів, досягнутих за допомогою здобутого засобами ІАД нового знання, завершується його оцінювання (див. рис. 5.1).

Залежно від задачі кількість етапів, а також обсяг виконуваних на кожному з них дій можуть змінюватися, але загалом усі вони необхідні і так чи інакше мають належати процесу інтелектуального аналізу даних.

Інший підхід, так звана перехресна перевірка, полягає в тому, що створюються підмножини даних і порівнюються результати роботи моделі на кожній підмножині. Такий підхід може використовуватися як під час вирішення передбачувальних, так і описових задач. Засоби автоматизації перехресної перевірки доступні під час використання Microsoft SQL Server 2012 версії Enterprise або Developer.

Найбільш ефективні моделі *розгортаються* у виробничому середовищі. Водночас, можливі сценарії інтеграції засобів ІАД і користуальницьких застосувань. І кінцевий користувач, у відповідь на сформований запит, буде отримувати результати аналізу у вигляді звіту. Під час формування звітів про результати проведеного аналізу можуть використовуватися можливості служби SQL Server Reporting Services.

З часом, характеристики предметної області можуть мінятися, що вимагатиме і зміни шаблонів ІАД. Може знадобитися перенавчання існуючих моделей або створення нових. У ряді випадків, SQL Server дозволяє автоматизувати процес оновлення моделей за рахунок використання служб Integration Services.

Самостійна робота № 3. Використання інструментів «Fill From Example» та «Forecast»

Інструменти: Fill From Example та Forecast використовуються для розв'язання задач прогнозування невідомих значень параметрів, тому для задач прогнозування необхідно мати вхідні дані за певний проміжок часу для побудови математичної моделі та майбутнього прогнозування.

Нехай для дослідження вхідних даних буде використовуватися той самий локальний приклад для Excel з попередньої самостійної роботи №2. Вхідні дані знаходяться в листі «Заповнення з прикладу» (рис. 3.3).

ID	Семейное положение	Пол	Возраст	Дети	Образование	Тип работы	Владеет дом.	Машина	Расстояние до работы	Регион	Высоко-доходный клиент
12496	Женатый, замужни	Женский	40000	1	Бакалавр	Квалифицирован	Да	0-0.1 км	Европа	42	Да
24107	Женатый, замужни	Мужской	30000	3	Неоконченное выс	Офисной работ	Да	1.0-1 км	Европа	43	Да
14177	Женатый, замужни	Женский	80000	5	Неоконченное выс	Профессионал	Нет	2.2-5 км	Европа	60	Да
24381	Одиночный(ая)	Мужской	70000	0	Бакалавр	Профессионал	Да	1.5-10 км	Россия	41	Нет
25397	Одиночный(ая)	Мужской	30000	0	Бакалавр	Офисной работ	Нет	0.0-1 км	Европа	36	Да
13507	Женатый, замужни	Женский	10000	2	Неоконченное выс	Ручной труд	Да	0.1-2 км	Европа	50	Нет
27974	Одиночный(ая)	Мужской	160000	2	Среднее	Управление	Да	4.0-1 км	Россия	31	Нет
19364	Женатый, замужни	Мужской	40000	1	Бакалавр	Квалифицирован	Да	0.0-1 км	Европа	43	Да
22155	Женатый, замужни	Мужской	20000	2	Неоконченное сре	Офисной работ	Да	2.5-10 км	Россия	58	Нет
19380	Женатый, замужни	Мужской	20000	2	Неоконченное выс	Ручной труд	Да	1.0-1 км	Европа	48	Да
22173	Женатый, замужни	Женский	30000	3	Среднее	Квалифицирован	Нет	2.1-2 км	Россия	54	
12697	Одиночный(ая)	Женский	90000	0	Бакалавр	Профессионал	Нет	4.10-9 км	Россия	36	
11434	Женатый, замужни	Мужской	170000	5	Неоконченное выс	Профессионал	Да	4.0-1 км	Европа	55	

Рис. 3.3 – Вікно набору даних для служби Fill From Example

Набір даних описує ряд клієнтів магазину. Для деяких з них помічено те, чи являється клієнт «високодохідним», чи ні. Ці рядки будуть використовуватися для навчальної вибірки. Задачею аналізу в даній самостійній роботі буде являтися оцінка решти клієнтів за заданим прикладом. Тому для розв'язання поставленої задачі використовується алгоритм Microsoft Logistic Regression. Для створення моделі з вхідної вибірки даних необхідно, щоб в усій вибірці вхідних даних були представлені усі варіанти з усіма можливими значеннями цільового стовпця. Як правило, чим більше характерних прикладів в вхідній вибірці, тим якісніше буде проведено навчання моделі.

Даний інструмент не придатний для задачі прогнозування значень параметру, що необхідно дослідити та може приймати неперервні числові значення. Ще однією особливістю є те, що аналіз проводиться по стовпчикам (передбачається значення стовпця). Якщо, ряд який необхідно заповнити зберігається в вигляді стрічки перед початком аналізу необхідно виконати транспонування (скопювати буфер обміну даних, вибрати в контекстному меню «Спеціальна вставка» позначити прапорець «Транспонувати»).

Запустимо інструмент Fill From Example. В першому вікні буде запропоновано вибрати стовпець, що містить вже заповнені дані,

в даному випадку він автоматично визначений як – «Високодохідний клієнт». Як і в попередніх випадках за посиланням «Choose columns to be used for analysis», можна вибрати стовпці які враховуються під час аналізу. Євристичний механізм визначив, що поле ID враховувати не потрібно. На практиці рекомендовані налаштування варто змінювати лише у випадку, якщо точно відомо про взаємну незалежність параметрів.

Після запуску, інструмент формує звіт по знайденим шаблонам даних (рис. 3.4) та додає стовпець з прогнозованими значеннями до вхідної таблиці. В звіті описано наявність залежності між значеннями цільового стовпця та значеннями інших стовпців.

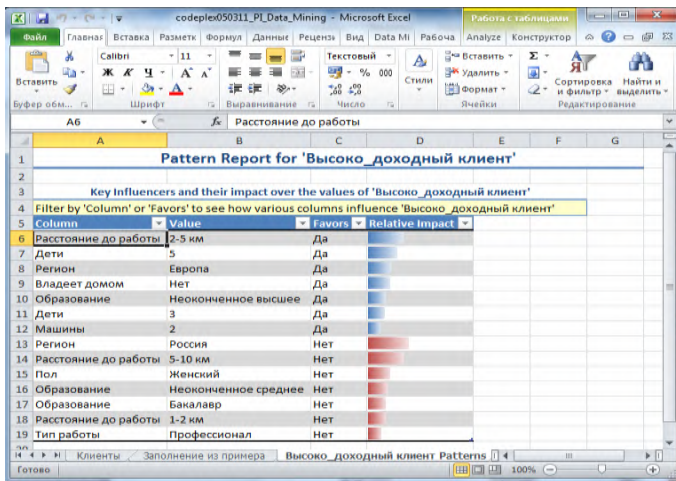


Рис. 3.4 – Таблица початкових факторів та показників залежностей між факторами

На рисунку 3.4 видно, що значення параметру «Відстань до роботи» рівне 2-5 км має самий найбільший рівень приналежності під час вибору варіанту «Так». Це можна інтерпретувати як «відстань 25 кілометрів до роботи у багатьох випадках визначає вибір на користь покупки велосипеда». Тому для кожного рядка розраховується загальна оцінка для кожного варіанту (для прикладу «Так» та «Ні»), а також робиться вибір на користь значень з найбільшим сумарним ваговим коефіцієнтом. Дане значення заноситься в стовпець з суфіксом «Високодохідний клієнт_Extended»). Для вхідних даних на котрих навчалася побудована модель, значення цього стовпця співпадає з зразком.

Припустимо, що було отримано додаткові данні про деяких клієнтах магазину. Можна змінити зразок після чого необхідно знову запустити інструмент. Нові значення будуть отримані з урахуванням уточнень на наборі даних для навчання. Подібні ітерації дозволяють послідовно уточнювати оцінку значень моделі.

ID	Семейное положение	Пол	Возраст	Дети	Образование	Тип работы	Владеет ли	Маш.	Расстояние до работы	Регион	Возраст	Высота	Высоко лояльный клиент	Секс
12496	Женатый, замужем	Женский	40000	1	Бакалавр	Квалифицирован	Да	0-0.1 км	Европа	42	Да	Да		
24127	Женатый, замужем	Мужской	30000	3	Неоконченное выс	Офисный работ	Да	1-0.1 км	Европа	43	Да	Да		
14177	Женатый, замужем	Женский	80000	3	Неоконченное выс	Профессионал	Нет	2-3 км	Европа	60	Да	Да		
24381	Одинокий(ая)	Мужской	70000	0	Бакалавр	Профессионал	Да	1-5-10 км	Россия	41	Нет	Нет		
25597	Одинокий(ая)	Мужской	30000	0	Бакалавр	Офисный работ	Нет	0-0.1 км	Европа	36	Да	Да		
13507	Женатый, замужем	Женский	10000	2	Неоконченное выс	Ручной труд	Да	0-1.2 км	Европа	50	Нет	Нет		
27974	Одинокий(ая)	Мужской	160000	2	Среднее	Управление	Да	4-0.1 км	Россия	33	Нет	Нет		
15364	Женатый, замужем	Мужской	40000	1	Бакалавр	Квалифицирован	Да	0-0.1 км	Европа	43	Да	Да		
22155	Женатый, замужем	Мужской	20000	2	Неоконченное сре	Офисный работ	Да	2-5-10 км	Россия	58	Нет	Нет		
15280	Женатый, замужем	Мужской	20000	2	Неоконченное выс	Ручной труд	Да	1-0.1 км	Европа	48	Да	Да		
22175	Женатый, замужем	Женский	30000	3	Среднее	Квалифицирован	Нет	2-1.2 км	Россия	54	Нет	Нет		
12897	Одинокий(ая)	Женский	90000	0	Бакалавр	Профессионал	Нет	4-10+ км	Россия	36	Нет	Нет		
11454	Женатый, замужем	Мужской	170000	5	Неоконченное выс	Профессионал	Да	4-0.1 км	Европа	55	Да	Да		
25123	Женатый, замужем	Мужской	40000	2	Неоконченное выс	Офисный работ	Да	1-1.2 км	Европа	35	Да	Да		
25542	Одинокий(ая)	Мужской	60000	1	Неоконченное выс	Квалифицирован	Нет	1-0.1 км	Россия	45	Да	Да		
20870	Одинокий(ая)	Женский	10000	2	Среднее	Ручной труд	Да	1-0.1 км	Европа	38	Нет	Да		

Рис. 3.5 – Значення оцінок моделі

Для того, щоб повністю видалити результати роботи інструменту достатньо видалити робочий лист з звітом та доданий стовпець в таблиці з вхідними даними.

Виконаємо прогнозування. Інструмент Forecast дозволяє виконувати задачі пов'язані з прогнозуванням значень числових рядів. Значення ряду повинні бути заповнені в стовпець таблиці (якщо досліджувані дані представлені у рядка то його потрібно транспонувати). В раніше використовуваному файлі Excel на листі «Прогнозування» є набір даних по сумах продажів велосипедів марки M200 по місячно в трьох різних регіонах. Таким чином для дослідження є три числових послідовності, що можливо взаємопов'язані між собою (рис. 3.6).

The screenshot shows a Microsoft Excel window titled 'codepilot090311_Pt_Data_Mining - Microsoft Excel'. The ribbon includes 'Главная', 'Вставка', 'Разметка страниц', 'Формулы', 'Данные', 'Рецензирование', 'Вид', 'Data Mining', 'Рабочая группа', and 'Таблица с таблицами'. The 'Table Analysis Tools' ribbon is also visible, with options like 'Analyze Key Influencers', 'Detect', 'Fill From Forecast', 'Highlight Exceptions', 'Scenario Analysis', 'Prediction', 'Shopping Basket Analysis', 'DMAdminDB localhost', 'Map', and 'Help'. The active sheet is 'Россия, руб.' and contains a table with the following data:

Год/месц	Европа, руб.	США, руб.	Россия, руб.
200107	20324,94	20324,94	64424,81
200108	20349,94	23724,93	60899,82
200109	16949,95	16974,95	10174,97
200110	16949,95	20299,94	54174,94
200111	27124,92	23749,93	57599,83
200112	27049,92	47399,86	57474,83
200201	27124,92	30474,91	64349,81
200202	23699,93	30424,91	6799,98
200203	27049,92	30499,91	74524,78
200204	27099,92	33874,9	77824,77
200205	23699,93	60924,82	67699,8
200206	30524,91	43999,87	74549,78
200207	24678,464	39156,0798	47330,1512
200208	32897,1782	45325,6958	55571,1868

Рис. 3.6 – Набір значень для прогнозування продажів помісячно велосипедів моделі M200 в різних регіонах

В процесі роботи інструмент будує модель з використанням алгоритму часових рядів (Microsoft Time Series). Для роботи інструменту необхідно, щоб в усіх стовпцях були лише числові значення, пропуски також не допускаються. Прогнозувати можна неперервні числові або грошові (currency) значення. Інструмент не розрахований для прогнозування дат.

Як зазначалося в [5], інструмент проводить пошук в числовій послідовності на основі якої відбувається прогнозування шаблонів наступних типів:

1) Тренд – закономірність зміни значень. Тренд може бути зростаючим та спадаючим.

2) Періодичність (сезонність) – події повторюються через певну кількість інтервалів.

3) Взаємна кореляція – залежність значень одного ряду від інших (наприклад, вартість акцій нафтових компаній від цін на нафту). Алгоритми які проводять перевірку на наявність кореляційного зв'язку між факторами входять до пакету Microsoft SQL Server 2012 версії Enterprise або Developer, крім версії Standard.

Налаштування параметрів полягає у виборі стовпців для аналізу, кількості значень для прогнозування, а також вибору періодичності (рис. 3.7).

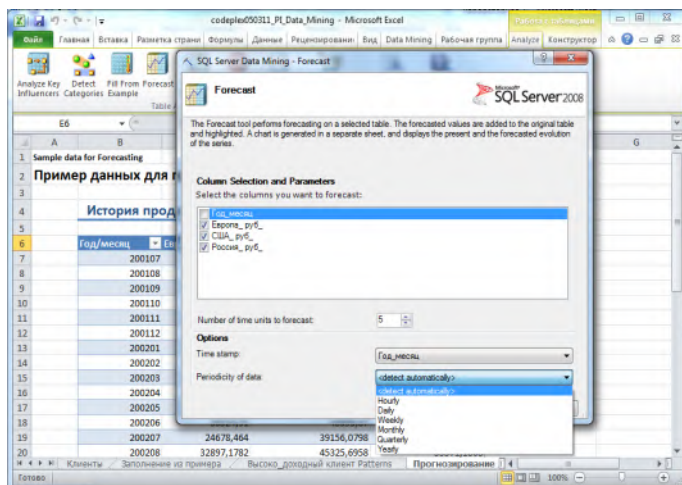


Рис. 3.7 – Вікно для налаштування параметрів для інструменту Forecast

В даному випадку, в якості мітки часу логічно обрати поле «Рік/місяць» для сумісності з системними вимогами SQL server, далі поставити мітку, яка виключає поле «Рік/місяць» з списку значень для прогнозування. Слід зазначити, що значення стовпця використовується в якості часової мітки і вони повинні бути унікальними. Що стосується періодичності, то існує два варіанти для вибору періодичності:

1. Hourly(почасова) – періодичність 12.
2. Daily (щоденна) – періодичність 5 або 7 (робочий тиждень та дні відповідно);
3. Weekly (тижнева) – періодичність 4 або 13 (кількість тижнів в місяці та кварталі).
4. Monthly (місячна) – періодичність 12 (число місяців в році).
5. Yearly – пакет інструментів автоматично буде визначати періодичність.

Якщо періодичність невідома, то рекомендується позначити «detect automatically», щоб пакет проаналізував вхідні дані на наявність періодичності та визначив її самостійно, якщо вона присутня.

За результатами аналізу генерується звіт у вигляді таблиці та графіка (рис. 3.8), де зображено тренд побудований за вхідними даними. Пунктирною лінією зображено прогнозовані значення тренду. Слід звернути увагу на те, що часові позначки для даних, що прогнозувалися на графіку не представлені.

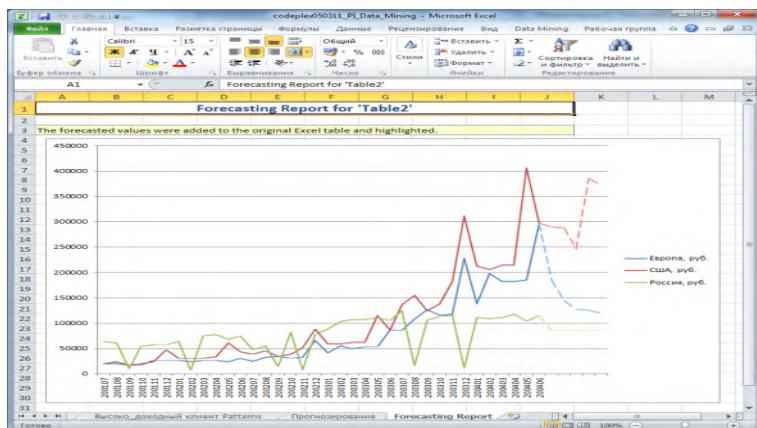


Рис. 3.8 – Графік прогнозованих значень.

До вихідної таблиці додано прогнозні результати (рис. 3.9).

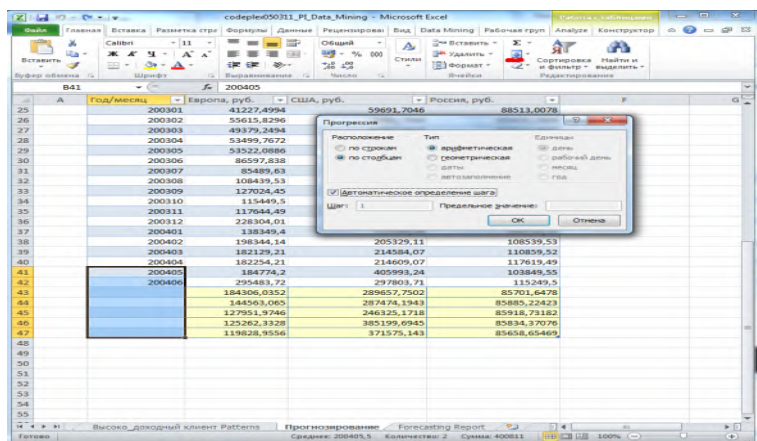


Рис. 3.9 – Заповнені прогнозовані значення для початкової таблиці

На рисунку 3.9 щойно додані значення зображено кольором фону. Щоб продовжити часовий ряд можна виділити декілька останніх значень стовпця «Рік/місяць» та область що незаповнена даними, вибрати на панелі керування Головна далі Заповнити із списку, що з'явився вибрати пункт «Прогресія» та вказати автоматичне визначення кроку. Дані котрих не вистачає будуть додані автоматично. Для того, щоб видалити результати потрібно видалити лист.

Щоб прибрати результати роботи інструменту, треба видалити аркуш звіту і рядки вихідної таблиці з передбаченими значеннями.

Завдання. За допомогою інструменту побудуйте прогноз продажів на рік (12 значень). Проаналізуйте графік. На ваш погляд, який тип періодичності виявив інструмент у вихідних даних і використовує для передбачення?

Самостійна робота № 4. Використання інструментів
аналізу «Highlight Exceptions» та «Scenario
Analysis»

Відокремлення виключень

З назви видно, що інструмент дозволяє виявити данні, які виділяються якимись характеристиками серед решти даних. Це може бути корисно, коли дані є помилковими (наприклад, результат помилки оператора під час введення даних з клавіатури), також аналіз виключень може розглядатися як метод ІАД за допомогою інших методів, це дозволяє забрати грубі помилки, тобто значення які не схожі на решту вхідних даних вибірки.

Під час роботи інструмент інтелектуального налізу даних «Highlight Exceptions» створює тимчасову модель інтелектуального аналізу з використанням алгоритму Microsoft Clustering. Для кожного рядка з значеннями, що аналізується оцінюється степінь приналежності виявленим кластерам, значення які знаходяться далеко від кластерів позначаються, як виключення.

Під час запуску інструмента можна позначити стовпці, які не враховуються в процесі аналізу. В списку джерел [3, 5] рекомендується виключати з аналізу стовпці з унікальними значеннями (імена, ідентифікатори), а також стовпці які містять багато пустих значень або текст. На рисунку 3.10 видно, що під час аналізу набору даних «Клієнти» інструментом виключається з розгляду поле «ID».

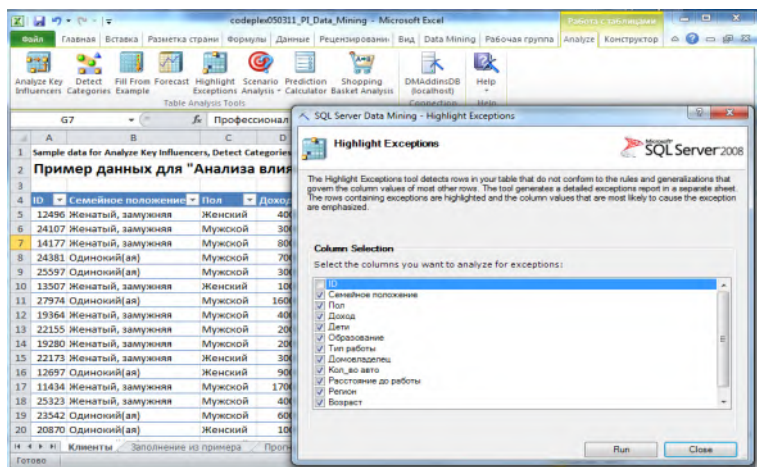


Рис. 3.10 – Вікно запуску інструменту Highlight Exceptions

За результатами роботи формується таблиця (рис. 3.10), а у вихідному наборі значень виключення виділяються іншим кольором (рис. 3.11).

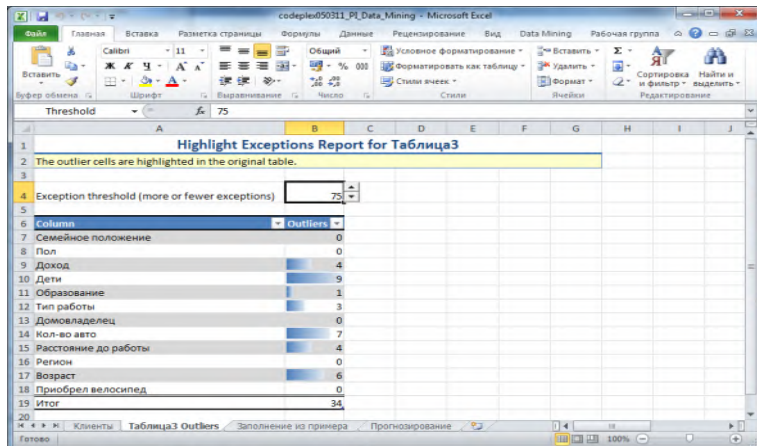


Рис. 3.11 – Результат проведеного аналізу

На рисунку 3.12 інструмент дозволяє вказати інтервал відхилення від норми значень (Exception threshold), показник вимірюється у відсотках (оцінка ймовірності того, що розглянуте значення відноситься до виключення).

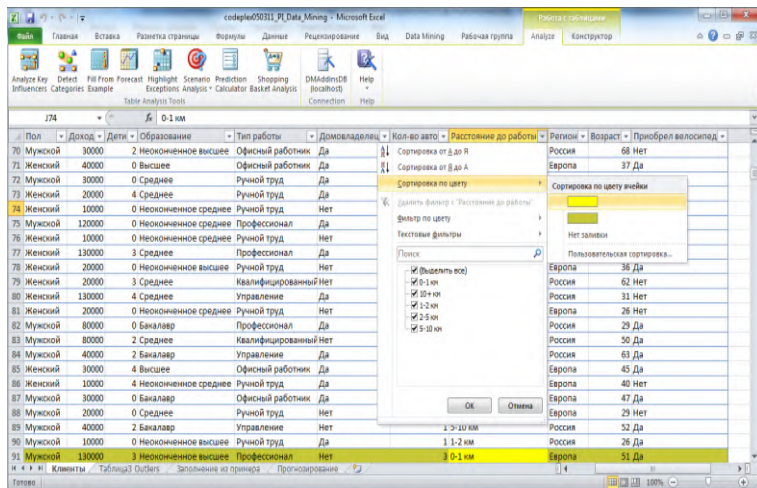


Рис. 3.12 – Данні, які увійшли до виключень, виділяють жовтим кольором, що дозволяє провести сортування

Зменшення інтервалу призведе до того, що більше значень увійде до виключення, збільшення інтервалу – навпаки. під час значенні рівним 7,5 % в даному наборі вхідних даних знайдено 34 виключення. Таблиця показує в яких стовпцях та яка кількість значень увійшла до виключення.

Певернемося до листа Excel з вихідними даними на весь рядок, який зафарбований коричневим кольором, а точні значення, що цікавлять нас жовтим (рис. 3.13).

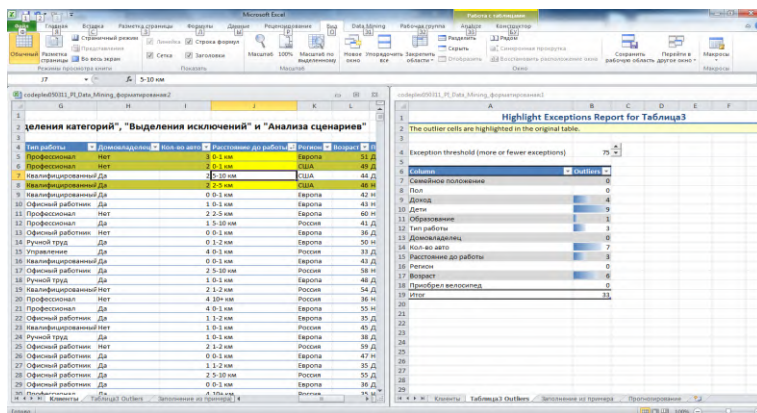


Рис. 3.13 – Вікно з редагуванням помилових значень

Для того, щоб згрупувати потрібні рядки для подальшого аналізу, необхідно скористатися вбудованими функціями Excel, лише вони дозволяють провести сортування даних за кольором. Також можна скористатися інструментами на вкладці «Вид», щоб створити новий лист, щоб розмістити робочу таблицю поруч з даними (рис. 3.13). Нехай у вибраному масиві даних ми знайшли помилку: «Відстань від роботи деякого клієнта з США, який має два авто не 0-1 км, а 5-10 км (саме тому клієнту потрібно в сім'ї необхідно два авто)». Якщо змінити значення, то відбудеться автоматичний перерахунок результатів по даних, що були змінені. У випадку, що представлений на (рис. 3.13) нові значення вже не входять в інтервал помилок (виключень).

Зверніть увагу на те, що в таблиці відбувся обрахунок результатуючих ознак по редагованим значенням та змінився фон комірок для значень, в даному випадку на рисунку відображаються 33 виключення. Автоматичний перерахунок працює лише тоді, коли час роботи з аналітичними службами SQL Server залишається відкритим. Якщо таблиця Excel була закрита, а потім повторно відкрита то в даному випадку автоматичного перерахунку не буде (необхідно знову провести повторний аналіз).

Також, слід зазначити, що інструмент реагує тільки на зміни даних в обраному діапазоні комірок робочого листа, які використовуються під час навчання. Якщо після початку роботи інструменту в кінець таблиці додати нові рядки з значеннями, вони не будуть братися до уваги під час подальших розрахунків.

Як відзначалося раніше, якщо необхідно розглядати лише найбільш важливіші похибки, можна збільшити інтервал відхилень і інструмент самостійно змінить оцінки у відповідності з заданими значеннями (рис. 3.14).

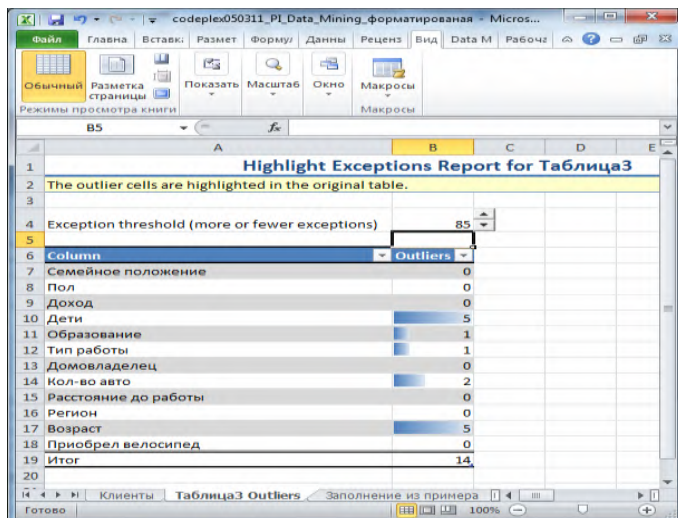


Рис. 3.14 – Збільшення інтервалу відхилень, зменшує число виключень

Під час повторного запуску, інструмент видалить результати попереднього аналізу, що проводився раніше. Також слід врахувати те, що зміни які провів інструмент є досить складними (тонування рядків таблиці та ін.), якщо необхідно видалити результати роботи то необхідно запустити повторний аналіз інструментом та підтвердити видалення даних у діалоговому вікні яке викличе інструмент, діалогове вікно буде схоже на вікно на (рис. 4.15), якщо необхідно скасувати виклик аналізу то необхідно натиснути кнопку **Close**.

Аналіз сценаріїв

Інструмент Scenario Analysis дозволяє моделювати вплив, що відбувається під час зміни одного із параметрів (значень одного із стовпців) на інший, який має залежність з першим стовпцем. В основі роботи інструменту лежить використання алгоритму Microsoft Logistic Regression. Для формування тимчасової моделі необхідно використати статистичну сукупність даних на основі якої відбудеться навчання з кількістю елементів не менше 50 записів.

В інструмент Scenario Analysis вбудовано два основні інструменти: «Аналіз сценарію для пошуку рішень» (Goal Seek) та «Аналіз можливих варіантів» (What-If).

Аналіз сценарію пошуку рішень (Good Seek)

Користуватися інструментом Good Seek дуже легко та швидко, оскільки, саме він дозволяє оцінити та досягнути бажаного результату в цільовому стовпці вхідних даних та дозволяє змінювати значення обраного параметра. За допомогою обраного інструмента можна проводити аналіз як для одного значення так і для всієї вибраної сукупності. Крім того, неможна забувати, що запит до даних необхідно формувати з урахуванням знань в предметній області. Наприклад, необхідно зробити запит, якщо людина хоче збільшити річний дохід на 20% то, чи необхідно йому придбати для цього велосипед? І після сформованого запиту отримати який результат. Не зрозумілим є те, що навіть в такій постановці запиту формувати не логічне питання про велосипед. Припустимо, що ми хочемо отримати знання про те, як буде впливати освіта на рівень матеріального забезпечення людини. Для початку проведемо аналіз для одного значення. Наприклад нас цікавить клієнт з ідентифікатором 12496 (перше значення в наборі вхідних значень). Відкриємо набір даних «Клієнти» і на вкладці **Analysis** виберемо **Scenario Analysis → Goal Seek** (рис. 3.15).

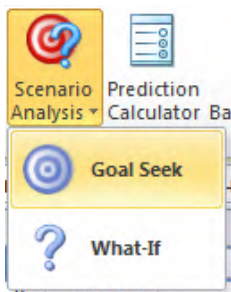


Рис. 3.15 – Вікно запуску інструмента Goal Seek

У вікні параметрів (рис. 3.16) позначимо цільовий стовпець – «Дохід», а також бажане значення 125% від поточного доходу. В даному випадку інструмент обраховує результат, значення якого не менше за бажане в прикладі $40000 \times 1,25 = 50000$ та більше. Якщо значення яке необхідно задати для параметра буде меншим за 100% та результат буде меншим за згаданий дохід.

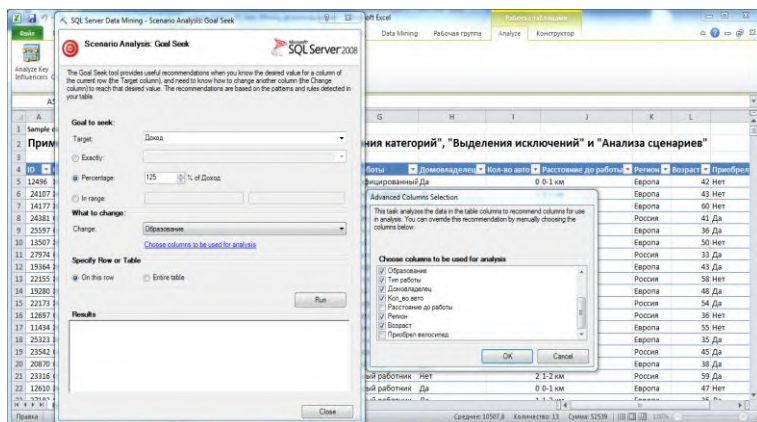


Рис. 3.16 – Вікно налаштування параметрів для Goal Seek

Також можна вказати точне значення та діапазон(вибравши «In range»). Для значень які не є числовими варіант «Percentage» та «In range» будуть не активні. Для досягнення шуканого значення необхідно замінити стовпець «Освіта».

Перейшовши за наступним посиланням «Choose columns to be used for analysis» слід зазначити, що під час аналізу необхідно розглянути стовпці «ID», «Діти», «Відстань до роботи», «Придбати велосипед». Після закриття вікна «Advanced Columns Selection» необхідно ще раз перевірити налаштування блоку «Goal Seek» тому, що інколи під час переходу між вікнами: «Exactly», «Percentage» та «In range» перемикачем скидається значення по замовчуванню «Exactly» (рис. 3.17).

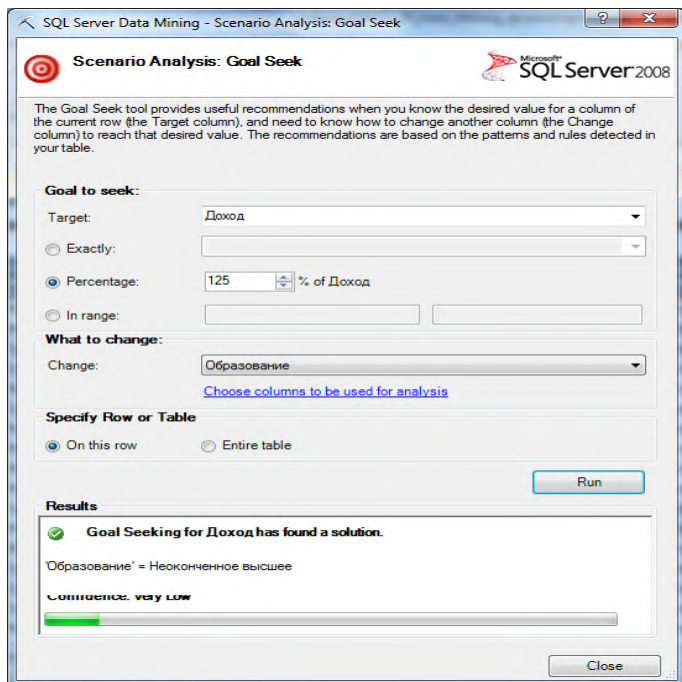


Рис. 3.17 – Вікно з результатом аналізу; для одного рядка знайдено розв’язок

Для обраного рядка знайдено шаблон, що рекомендується для параметру «Освіта» значення «Незакінчена вища». Водночас рівень надійності даного результату – Confidence (інколи верхня частина надпису стирається так само як на рисунку), оцінюється як дуже низький «Very low». Якщо переміститися на наступний рядок значень, та знову натиснути Run то отримаємо результат для нових даних (рис 3.18). В даному випадку необхідний розв’язок (відповідь) не був знайдений, але був наближений найбільш підходящий варіант.

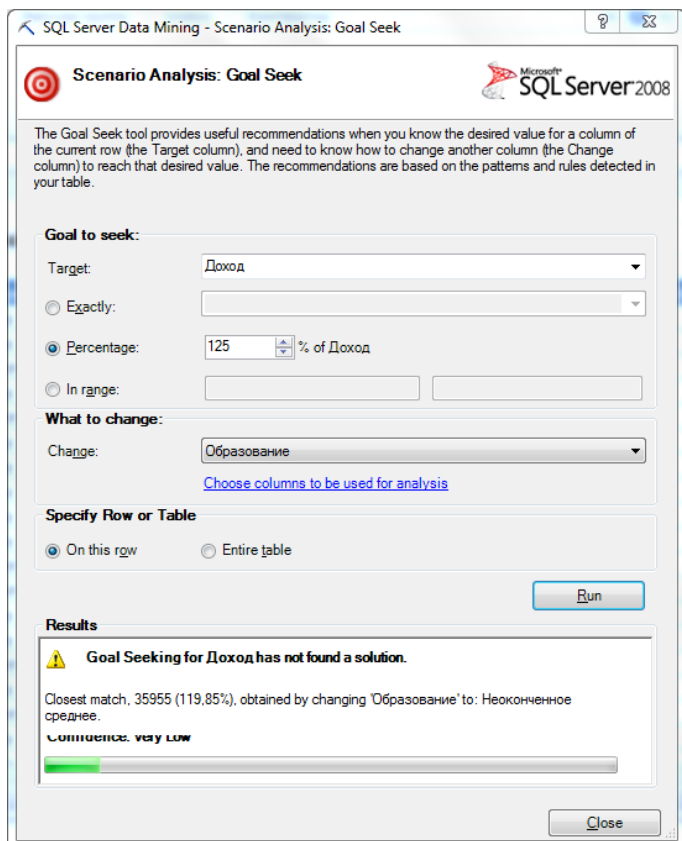


Рис. 3.18 – Вікно з результатом аналізу для рядка даних

Якщо у секції «Specify Row or Table» встановити перемикач в «Entire table», тоді сценарій вибірки буде обраховано для всіх рядків значень (рис. 3.19).

	G	H	I	J	K	L	M	N	O
4	Тип работы	Домашнее	Кол-во авто	Расстояние до работы	Регион	Возраст	Приобрел велосипед	Goal: Доход (change by 125.00 %)	Recommended Образование
5	Квалифицированный	Да	0-0.1 км	Европа	42	Нет	✓	✓	Неоконченное высшее
6	Офисный работник	Да	1-0.1 км	Европа	43	Нет	✗	✗	Неоконченное среднее
7	Профессионал	Нет	2-5-10 км	Россия	40	Нет	✓	✓	Неоконченное среднее
8	Профессионал	Да	1-5-10 км	Россия	41	Да	✓	✓	Неоконченное высшее
9	Офисный работник	Нет	0-0.1 км	Европа	36	Да	✓	✓	Неоконченное высшее
10	Ручной труд	Да	0-1-2 км	Европа	50	Нет	✓	✓	Бакалавр
11	Управление	Да	4-0-1 км	Россия	33	Да	✓	✓	Неоконченное высшее
12	Квалифицированный	Да	0-0-1 км	Европа	43	Да	✓	✓	Неоконченное высшее
13	Офисный работник	Да	2-5-10 км	Россия	58	Нет	✓	✓	Неоконченное высшее
14	Ручной труд	Да	1-0-1 км	Европа	48	Да	✗	✗	Неоконченное среднее
15	Квалифицированный	Нет	2-1-2 км	Россия	34	Да	✓	✓	Бакалавр
16	Профессионал	Нет	4-10+ км	Россия	36	Нет	✓	✓	Неоконченное высшее
17	Профессионал	Да	4-0-1 км	Европа	35	Нет	✗	✗	Неоконченное среднее
18	Офисный работник	Да	1-1-2 км	Европа	35	Да	✓	✓	Неоконченное среднее
19	Квалифицированный	Нет	1-0-1 км	Россия	45	Да	✗	✗	Неоконченное среднее
20	Ручной труд	Да	1-0-1 км	Европа	38	Да	✗	✗	Неоконченное высшее

Рис. 3.20 – Вікно з результатом аналізу для всієї вибірки даних

Результат роботи вказано у двох стовпцях які були додані до вхідної таблиці даних. Для тих рядків які позначені хрестиком в червоному колі, відповідного шаблону сценарію не було знайдено.

Запуск процедури аналізу для ряду інших комбінацій стовпців (наприклад - цільової стовпець "покупка велосипеда" = "так", незалежна змінна - "відстань до роботи") призводить до помилки "Query (1, 50) Синтаксичний аналізатор Невірний синтаксис" value ". ", мабуть пов'язаної з некоректною обробкою деяких типів даних.

Аналіз можливих варіантів «What-if

Інструмент What-if дозволяє розв'язати обернену по відношенню до Goal Seek задачу та оцінити значення цільової змінної у разі зміни значення параметра. Наприклад, можна оцінити як зміниться рівень доходу людини якщо підвищити його рівень освіти. Перейдемо до запису за ідентифікатором 12697 та запустимо інструмент: **Scenario Analysis -> What-if**. Вкажимо параметри для сценарію: освіта змінюється на «Вищу» та цільовий стовпець «Дохід». Отриманий результат показує, що у разі зміни рівня освіти дохід може збільшитися (вхідне значення 90000, середнє значення для нового шаблону 104448). Відповідно результатів на рисунку 3.21, ступінь приналежності для прогнозування не надто високий.

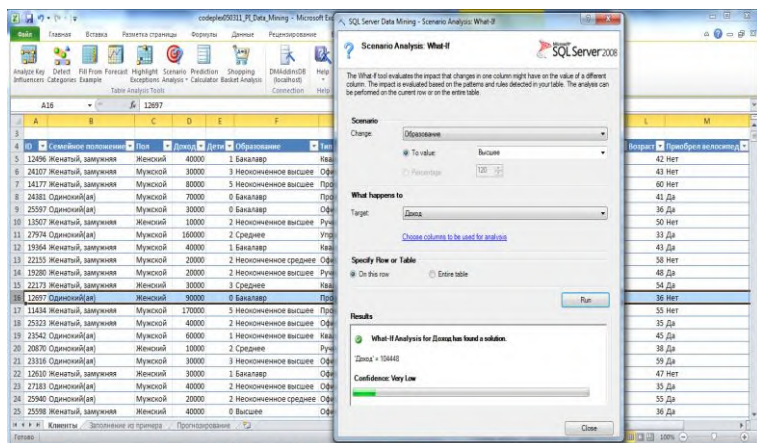


Рис. 3.21 – Работа інструменту What-If для окремого рядка даних

Аналогічно як і під час попереднього аналізу інструментом, подібний аналіз для сценарію можна зробити і для всієї таблиці одразу. В даному випадку до вхідної таблиці додаються два стовпці значень, де перший містить значення цільового параметра, інший – оцінку достовірності (рис. 3.22). Достовірність оцінюється числом від 0 до 100, де 100 – це максирисьна достовірність (абсолютна ймовірність виконання прогнозу), 0 – мінірисьна.

Пол	Доход	Дата	Образование	Тип работы	Домовладение	Маш на авто	Расстояние до работы	Ремонт	Возраст	Приобрел велосипед	Нов Доход	Confidence
Женский	40000	1	Бакалавр	Квалифицированный	Да	0-0.1 км	Европа	42	Нет		58875	
Мужской	30000	3	Неоконченное высшее	Офисный работник	Да	1-0.1 км	Европа	43	Нет		32586	
Мужской	80000	5	Неоконченное высшее	Профессионал	Нет	2-2.5 км	Европа	60	Нет		68975	
Мужской	70000	0	Бакалавр	Профессионал	Да	1-3-10 км	Россия	41	Да		67504	
Мужской	30000	0	Бакалавр	Офисный работник	Нет	0-0.1 км	Европа	36	Да		38673	
Женский	10000	2	Неоконченное высшее	Ручной труд	Да	0-1-2 км	Европа	50	Нет		17295	
Мужской	160000	2	Среднее	Управление	Да	4-0-1 км	Россия	33	Да		125621	
Мужской	40000	2	Бакалавр	Квалифицированный	Да	0-0.1 км	Европа	43	Да		62024	
Мужской	20000	2	Неоконченное среднее	Офисный работник	Да	2-3-10 км	Россия	38	Нет		62301	
Мужской	30000	3	Неоконченное высшее	Ручной труд	Да	1-0-1 км	Европа	48	Да		22449	
Женский	30000	3	Среднее	Квалифицированный	Нет	2-1-2 км	Россия	34	Да		51595	
Женский	90000	0	Бакалавр	Профессионал	Нет	4-10+ км	Россия	36	Нет		104448	
Мужской	170000	3	Неоконченное высшее	Профессионал	Да	4-0.1 км	Европа	35	Нет		106317	
Мужской	40000	2	Неоконченное высшее	Офисный работник	Да	1-1-2 км	Европа	35	Да		38591	
Мужской	60000	1	Неоконченное высшее	Квалифицированный	Нет	1-0-1 км	Россия	45	Да		60032	

Рис. 3.22 – Прогноз для всієї сукупності даних за допомогою What-If

Для того, щоб видалити результати роботи з таблицею інструментів What-If та Scenario Analysis, достатньо видалити додані стовпці. Під час роботи з окремими рядками, ніяких додаткових дій непотрібно виконувати.

Завдання 1. Проведіть аналіз винятків для набору даних "Клієнти" і значення порога в 90%. Запропонуйте інтерпретацію отриманих результатів.

Завдання 2. Проведіть аналіз винятків для набору даних "Прогнозування" (продажі по місяцях в різних регіонах). Запропонуйте інтерпретацію отриманих результатів.

Завдання 3. Проведіть аналіз для окремого рядка і таблиці, аналогічний описаному вище. Прокоментуйте результати.

Завдання 4. Проведіть аналіз даних, аналогічний описаному вище.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Наведіть етапи проведення інтелектуального аналізу даних (варіант Майкрософт)
2. Наведіть питання, відповіді на які необхідно одержати на етапі постановки завдання.
3. Що включає в себе етап підготовки даних?
4. Що з'ясовується на етапі вивчення даних?
5. Для чого призначений етап побудови моделей?
6. Яка мета етапу перевірки моделі?

ЛІТЕРАТУРА

1. Службы SQL Server Analysis Services – интеллектуальный анализ данных. – [Електронний ресурс]. – Режим доступу: <https://docs.Microsoft.com/ru-ru/sql/analysis-services/analysis-services-tutorials-ssas?view=sql-server-2017>

2. Бармак О.В. Интеллектуальный анализ данных. Методичні вказівки до самостійного вивчення курсу «Інтелектуальний аналіз даних» для студентів напрямку підготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.

3. Службы Analysis Services – видеоролики по интеллектуальному анализу данных (SQL Server 2012). URL: [http://Microsoftdn.Microsoft.com/ru-ru/library/dd776389\(v=SQL.100\).aspx](http://Microsoftdn.Microsoft.com/ru-ru/library/dd776389(v=SQL.100).aspx)

4. Baker L.D. Distributional Clustering of Words for Text Classification In ACM SIGIR 98 / L.D. Baker, A.K. McCallum. – 1998.
5. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.

РОЗДІЛ 4.

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: СПРОЩЕНИЙ АЛГОРИТМ БАЕСА ТА ДЕРЕВА РІШЕНЬ

4.1 Спрощений алгоритм Баєса

Спрощений алгоритм Баєса – це алгоритм класифікації, заснований на обчисленні умовної ймовірності значень прогнозованих атрибутів [1]. Водночас передбачається, що входні атрибути є незалежними, і визначений хоча б один початковий атрибут.

Алгоритм заснований на використанні формули Баєса. Нехай $A_1, A_2, \dots$ – повна група несумісних подій, а B – деяка подія, ймовірність якої додатна. Тоді умовна ймовірність події A_i , якщо в результаті експерименту спостерігалася подія B , обчислюється за формулою:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}. \quad (4.1)$$

Розглянемо наступний приклад. У 2002 році у Конгресі США було 51% представників республіканської партії і 49% демократів. Є інформація про те, як голосували представники тієї й іншої партії по ряду законопроектів (табл. 4.1). Необхідно визначити партійну приналежність якогось конгресмена.

Таблиця 4.1 – Результати голосування за законопроекти з розбиттям за належності до партій (Дем.-демократ, Респ.-республіканець)

	Зак-т 1		Зак-т 2		Зак-т 3		Зак-т 4		Партія	
	Дем.	Рес.	Дем.	Рес.	Дем.	Рес.	Дем.	Рес.	Дем.	Рес.
За	41	214	87	211	184	172	178	210	211	223
Проти	166	4	114	6	11	36	23	1		
За	20%	98%	43%	97%	94%	83%	89%	99.5	49%	51%
Проти	80%	2%	57%	3%	6%	17%	11%	0.5		

Без додаткової інформації ми можемо сказати, що з ймовірністю 0,49 він буде демократом, а з ймовірністю 0,51 – республіканцем. Це апіорна («задана до досвіду») ймовірність, у

формулі (4.1) позначена як $P(A_i)$, де подія A_i може полягати в тому, що даний депутат є демократом (Дем.) або республіканцем (Респ.).

Але ми можемо збільшити точність прогнозу, якщо знаємо, як голосував даний депутат (табл. 4.2), і загальні результати голосування членів Конгресу з тих самих питань (табл. 4.1). Формула Баєса дозволяє обчислити умовні (апостеріорні, «отримані після досвіду») ймовірності.

Таблиця 4.2 – Результати голосування шуканого представника

Зак-т 1	Зак-т 2	Зак-т 3	Зак-т 4	Партія
за	проти	за	за	?

Якщо відомо, що кожен член конгресу повинен належати лише до однієї із двох перерахованих партій, то повна група подій – це $\{Дем., Респ.\}$. Позначимо результати голосування з табл. 4.2 як «Рез.» і скористаємося правилом множення ймовірностей. Тоді умовні ймовірності можна розрахувати як:

$$P(Рез.|Дем.) \times P(Дем.) = 0,2 \times 0,57 \times 0,94 \times 0,89 \times 0,49 \approx 0,047;$$

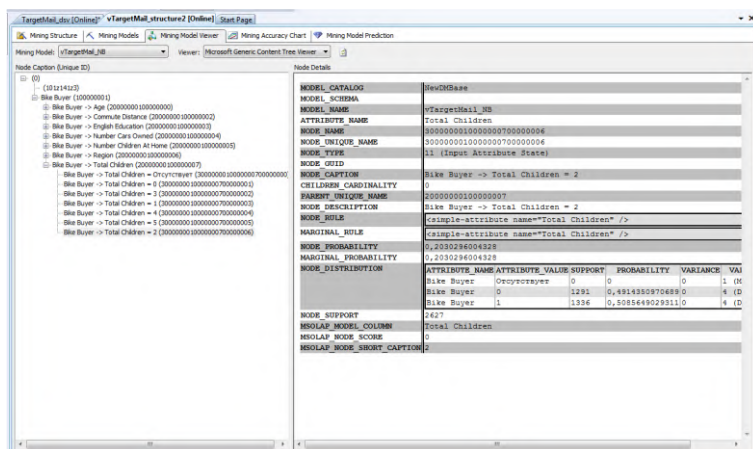
$$P(Рез.|Респ.) \times P(Респ.) = 0,98 \times 0,03 \times 0,83 \times 0,99 \times 0,51 \approx 0,012;$$

$$P(Дем.|Рез.) = \frac{0,047}{0,047 + 0,012} = 0,79;$$

$$P(Респ.|Рез.) = \frac{0,012}{0,047 + 0,012} = 0,21.$$

У підсумку можна зробити висновок, що з ймовірністю близько 0,79 розглянутий конгресмен є представником демократичної партії.

Якщо говорити про СКБД Microsoft SQL Server 2012, то у ній опис моделі інтелектуального аналізу зберігається у вигляді ієрархії вузлів. Для спрощеного алгоритму Баєса ієрархія має 4 рівні. Середовище розробки BI Dev Studio дозволяє переглянути вміст моделі (рис. 4.1). На верхньому рівні ієрархії знаходиться вузол самої моделі. Найнижчий рівень містить інформацію про те, з якою частотою зустрічалися в навчальній вибірці різні значення початкового параметра в поєднанні з вказаним значенням обраного вхідного параметра (рис. 4.1).



Для коректного використання спрощеного алгоритму Баєса необхідно враховувати що [2]:

Для коректного використання спрощеного алгоритму Баєса необхідно враховувати що [2]:

- вхідні атрибути повинні бути взаємно незалежними;
- атрибути можуть бути тільки дискретними або дискретизовані (в процесі дискретизації безліч значень безперервного числового атрибута розбивається на інтервали і далі йде робота з номером інтервалу);

– алгоритм вимагає меншої кількості обчислень, ніж інші алгоритми інтелектуального аналізу, що подаються Microsoft SQL Server 2008, тому він часто використовується для початкового дослідження даних. З тієї ж причини, даний алгоритм кращий для аналізу великих наборів даних з великою кількістю вхідних атрибутів.

З тієї ж причини, даний алгоритм кращий для аналізу великих наборів даних з великою кількістю вхідних атрибутів.

4.2 Алгоритм дерева рішень

Дерева прийняття рішень – це спосіб представлення правил в ієрархічній структурі, де кожному об'єкту відповідає єдиний вузол, що дає результуюче рішення. Під правилом розуміється логічна конструкція, представлена у вигляді «якщо ... то ...» (рис. 4.2).

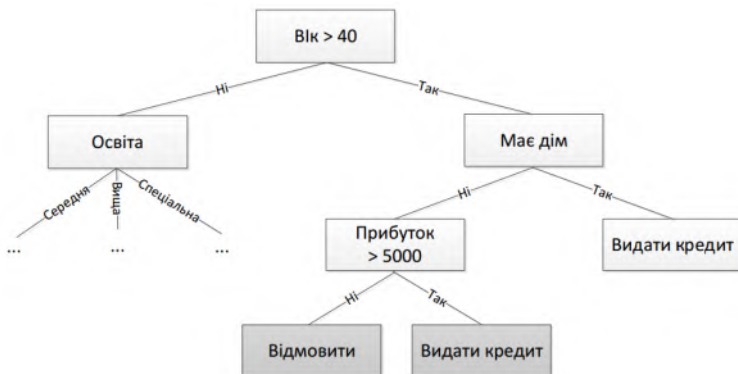


Рис. 4.2 – Схема методу дерева прийняття рішень

Дерева прийняття рішень (дерева класифікацій, регресійні дерева) – один з методів автоматичного аналізу даних.

Ситуації, в яких варто застосовувати дерева прийняття рішень, зазвичай виглядають наступним чином: є множина випадків, кожен з яких описується деяким кінцевим набором дискретних атрибутів, і в кожному випадку дано значення деякої (невідомої) булевої функції, що залежить від цих атрибутів.

Область застосування дерев рішень у даний час широка, але всі задачі, які вирішуються цим апаратом можуть бути об'єднані в наступні три класи:

- опис даних, за якого дерева рішень дозволяють зберігати інформацію про дані в компактній формі, у вигляді дерева рішень, що містить точний опис об'єктів;

- класифікація, з якою дерева рішень відмінно справляються; проте цільова змінна для даної задачі має мати дискретні значення;

- регресія; якщо цільова змінна має безперервні значення, дерева рішень дозволяють встановити залежність цільової змінної від незалежних (вхідних) змінних; наприклад, до цього класу відносяться завдання чисельного прогнозування (передбачення значень цільової змінної).

Розглянемо застосування методу дерев прийняття рішень на прикладі задачі прогнозування.

Припустимо, маючи дані про виступи футбольних клубів за останній час, ми бажаємо знати, чи виграє клуб «Динамо» свій наступний матч. Ми знаємо, що, насправді, це залежить від багатьох параметрів; проте перерахувати їх усі – задача безнадійна, тому обмежимося такими:

- чи вище суперник по турнірній таблиці;
- чи вдома грається матч;
- чи пропускає матч хтось з лідерів команди;
- чи йде дощ.

У нас є деяка статистика (табл. 4.3) і тепер ми хочемо зрозуміти, чи виграє «Динамо» в наступній грі. Припустимо, що сьогоднішній суперник стоїть нижче «Динамо» по турнірній таблиці, гра відбувається в гостях, лідери на місці і дощу немає. Спробуємо побудувати дерево прийняття рішень. В його вузлах, що не є листям, знаходяться атрибути, за якими різняться випадки. В листі знаходиться значення цільової функції. По ребрам будемо спускатися, щоб класифікувати випадки, які вже маємо.

Таблиця 4.3. – Статистична вибірка даних футбольних клубів

Суперник	Граємо	Лідери	Дощ	Перемога
Вище	Вдома	На місці	Так	Ні
Вище	Вдома	На місці	Ні	Так
Вище	Вдома	Пропускають	Ні	Так
Нижче	Вдома	Пропускають	Ні	Так
Нижче	В гостях	Пропускають	Ні	Ні
Нижче	Вдома	Пропускають	Так	Так
Вище	В гостях	На місці	Так	Ні
Нижче	В гостях	На місці	Ні	???

Використаємо атрибути в порядку, в якому вони вказані в таблиці. Тоді одержимо структуру, що зображена на рисунку 4.3.а). Використання дерева рішень для відповіді на питання, що нас цікавить, зводиться до того, щоб пройти по дереву згори вниз і визначити, в який з листків потрапляє ситуація, що нас цікавить.

Для того, щоб реалізувати інтелектуальний аналіз, треба змінювати дерево прийняття рішень у відповідності до нової інформації. Наприклад, якщо «Динамо» програв, то нічого змінювати не треба. Інакше довелося би доповнити дерево ще одним вузлом.

Побудоване дерево не ідеальне. Спробуємо взяти за основу інший атрибут, наприклад, помістимо в корінь дерева питання, чи йде дощ. Тоді глибина дерева буде дорівнювати 2 (рис. 4.3.б).

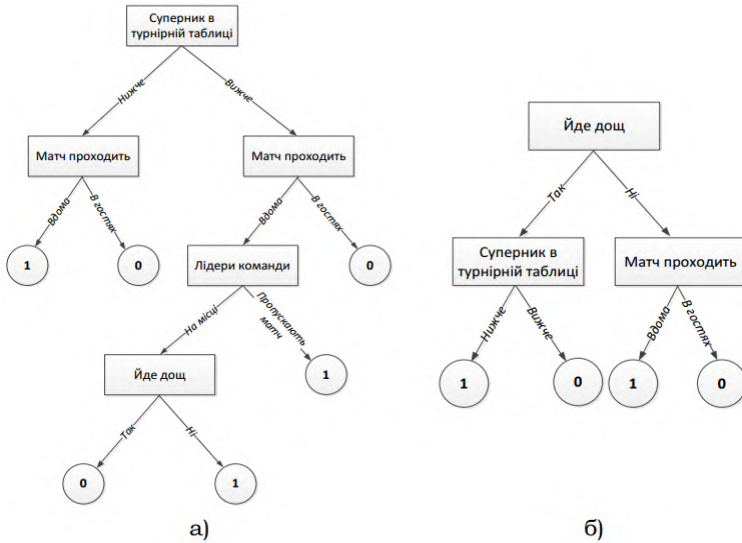


Рис. 4.3 – Схеми алгоритмів дерева прийняття рішень

В іншому випадку, алгоритм прийняття рішень працює під час побудови дерева для прогнозування неперервного стовпця [3]. У цьому випадку, кожен вузол містить регресійну формулу, а розбиття здійснюється в точці нелінійності у цій формулі. Даний підхід показаний на рисунку 4.4. Нехай найбільш точно наявні дані можна моделювати двома з'єднаними лініями – рисунку 4.4.а). Дерево рішень в цьому випадку будується так, як показано на рисунку 4.4.б).

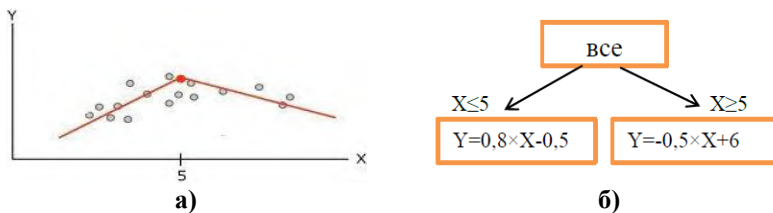


Рис. 4.4 – Схема побудови дерева рішень для неперервного прогнозованого атрибута

Але повернемося до дерев рішень. Під час розв'язку задачі класифікації, пройшовши від кореневого вузла до кінцевого («листа»), ми отримуємо результат. Із поданого на рисунку 4.5 фрагменту

“стандартний” аналіз наявних даних. У той же час, цей набір інструментів не надає особливих можливостей з підготовки даних до аналізу, оцінці результатів та ін. З Excel це можна зробити, використовуючи клієнт інтелектуального аналізу даних (Data MiningClient), який також входить в набір надбудов інтелектуального аналізу. Під час “Надбудови інтелектуального аналізу даних для Microsoft Office”, зазначалося, що бажано зробити повну установку надбудов, в яку входить і Data Mining Client.

Відкриємо вже використовувався нами набір даних, що входить в поставку надбудов (**Файл → Надбудови інтелектуального аналізу даних → Зразки даних Excel**). Щоб можна було спокійно вносити зміни, краще зберегти його під новим іменем. Перейдіть на лист “Вихідні дані” (SourceData) і клацніть на закладці Data Mining. Стрічка з пропонованими інструментами представлена на рисунку 4.6.

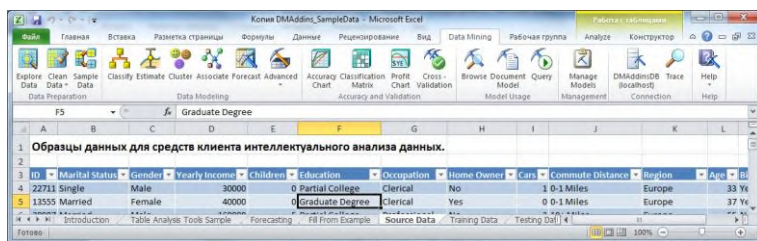


Рис. 4.6 – Інструменти Data Mining Client

Перша група інструментів (**Data Preparation – Підготовка даних**), дозволяє провести перше знайомство з набором даних і підготувати його для подальшого аналізу.

Наприклад, в попередніх роботах ми неодноразово стикалися з тим, що ряд алгоритмів (MicrosoftNaiveBayes і ін.) вимагають попередньої дискретизації безперервних значень числових параметрів. Але в ряді випадків користувачеві бажано подивитися можливі діапазони, уточнити їх число та ін. Окремий інтерес може представляти і розподіл рядків за значенням обраного параметра.

Дослідження даних

Інструмент Explore Data дозволяє проаналізувати значення стовпця (або діапазону комірок) і відобразити їх на діаграмі. Розглянемо його роботу на прикладі значення річного доходу клієнта (Income). Додатковий інтерес становить те, що це значення може

розглядатися і як безперервне, і як дискретне. Отже, запускаємо інструмент, що зображено на рисунку 4.7.

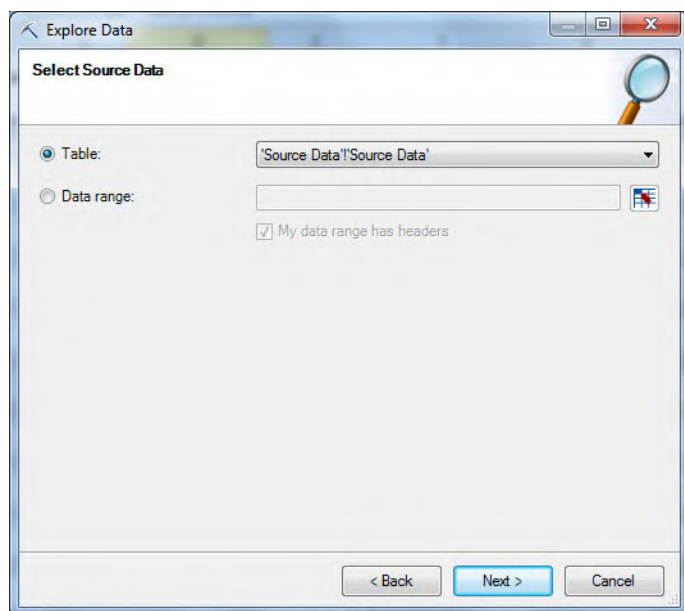


Рис. 4.7 – Вікно запуску аналітичної служби Explore Data

В процесі роботи буде потрібно вказати, для якої таблиці (або діапазону комірок) і стовпця буде проводитися аналіз (рис. 4.8).

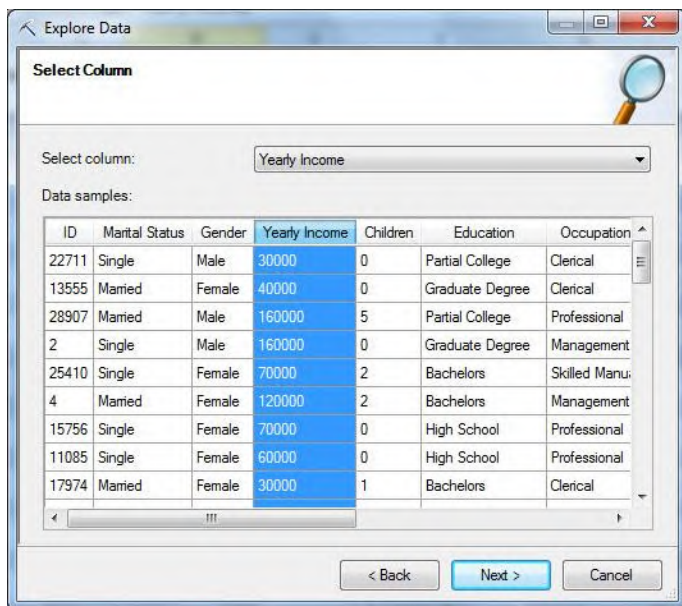


Рис. 4.8 – Вікно вибору діапазону колонок аналітичної служби Explore Data

Після цього вказані значення будуть проаналізовані і результат представлений у вигляді гістограми (рис. 4.9).

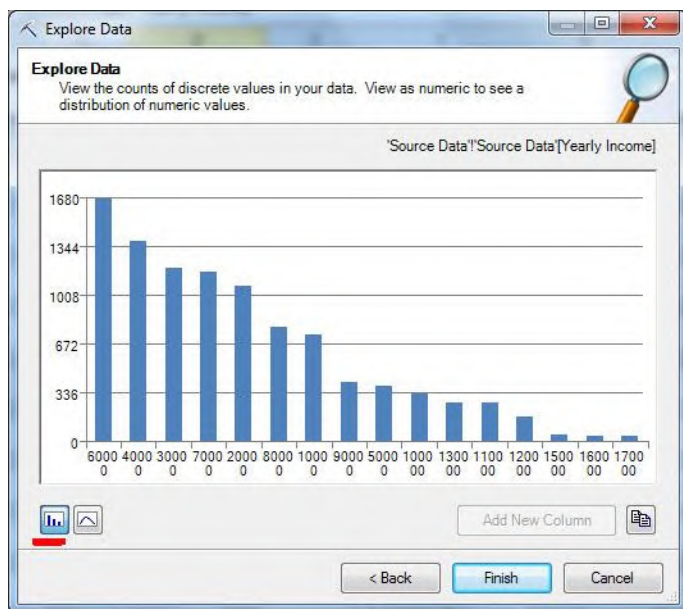


Рис. 4.9 – Вікно роботи аналітичної служби Explore Data

Як вже зазначалося вище, значення річного доходу можна розглядати і як безперервне, і як дискретне (за рахунок того, що в нашому наборі даних присутні тільки значення, кратні 10 тисячам). Для безперервного значення буде запропонований варіант розбиття на діапазони.

Число діапазонів можна поміняти і діаграма з розподілом значень будуть побудована заново. Натиснувши кнопку **Add New Column** можна додати в вихідну таблицю новий стовпець з інтервалами річного доходу. Наприклад, якщо для рядка значення Yearly Income = 30000, то значення нового параметра Yearly Income 2 під час використання представленого на рисунку розбиття буде “30000 – 50000” (саме так, з апострофом на початку, щоб розглядалося як строкове). Під час інтелектуального аналізу, отриманий стовпець може використовуватися замість справжнього (включення обох стовпців одночасно небажано). Кнопками із зображеннями графіка і гістограми (рис. 4.9), можна вказати тип аналізованого значення – безперервне або дискретне.

Якщо значення річного доходу розглядаємо як дискретне, то для нього буде побудована діаграма, що показує розподіл числа рядків за значенням річного доходу. Водночас сортування проводиться по

спадаючій числа рядків з даних значенням, через що перший стовпець гістограми відповідає значенню “60000”, другий – “40000” та ін. Сформовану гістограму можна скопіювати в буфер (кнопка правіше кнопки **Add New Column**, і використовувати для подальшої роботи.

Очищення даних

Інструмент Clean Data дозволяє підготувати дані для аналізу, відкинувши нетипові або помилкові дані (викиди), а також провівши заміну окремих значень (рис. 4.10). Як зазначається в документації, під викидом мається на увазі значення даних, що є проблематичним по одній з наступних причин:

- значення знаходиться за межами очікуваного діапазону;
- дані були введені неправильно;
- значення відсутній;
- дані представляють собою пробіл або порожній рядок;
- значення може значно відхилитися від розподілу, якому підпорядковуються дані в моделі.

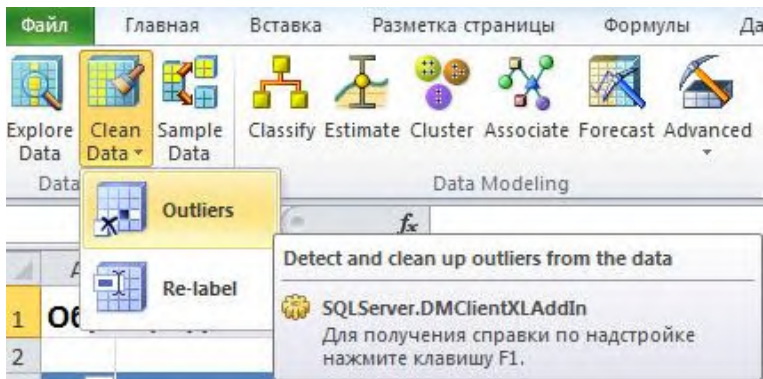


Рис. 4.10 – Вікно вибору аналітичної служби Clean Data

Використання даного інструменту проілюструємо на прикладі все тієї ж таблиці з даними про клієнтів (лист **Source Data**). Звернемося до колонку з віком. Нехай нам потрібно очистити набір даних від інформації про нехарактерних за віком покупців. Запускаємо інструмент **Clean Data** → **Outliers**, у вікні аналогічному представленому на рисунку 4.8 вибираємо таблицю для аналізу, потім у вікні **Select Column** – стовпець **Age**.

В даному наборі даних є рядки зі значеннями стовпчика **Age** від 25 до 96 років. Якщо цей параметр вважаємо безперервним, то він

буде представлений графіком, де по осі X вказується вік, по осі Y-число клієнтів з таким віком. У наборі даних частка клієнтів похилого віку дуже рідка. На рисунку 4.11 показано, що встановивши порогове значення в 75 років, ми відкидаємо заштрихований “хвіст”, що включає нехарактерні значення (покупці велосипедів у віці від 76 до 96 років, яких переважна меншість).

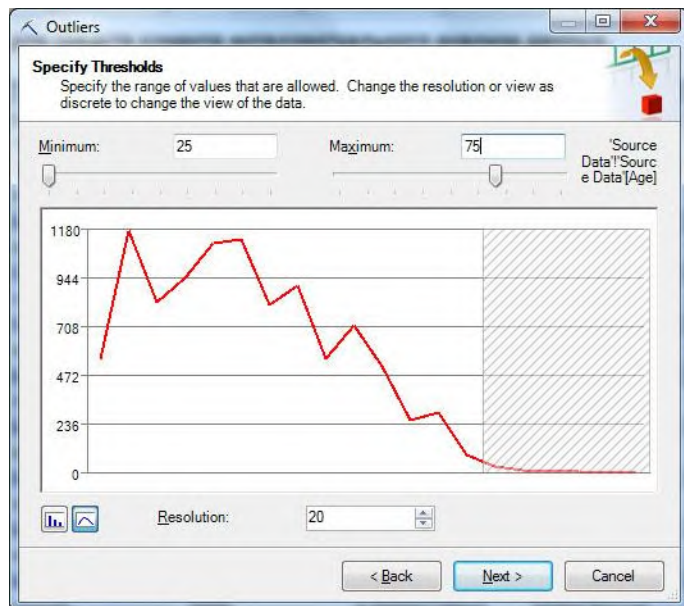


Рис. 4.11 – Вікно вибору параметрів аналітичної служби Clean Data

Багато в чому аналогічно виглядає робота з параметром, які приймають дискретні значення. Для нього будується гістограма, а для визначення порогу потрібно вказати мінімальне число прикладів, “підтримують” значення. Наприклад, на рисунку 4.12, встановлено граничне значення в 15.

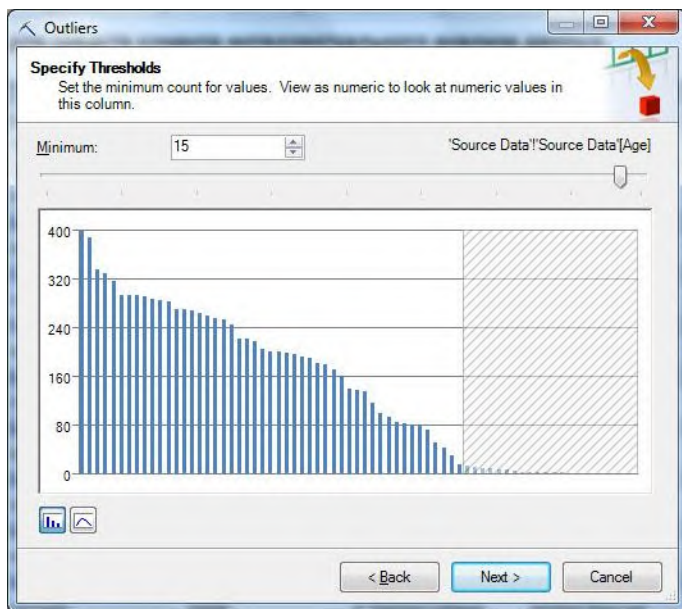


Рис. 4.12 – Вікно вибору параметрів аналітичної служби Clean Data

На жаль, у випадку великого числа стовпців гістограми, значення параметра на ній не відображаються. Тому не зрозуміти, що саме потрапляє в “хвіст” розподілу.

Отже, ми виділили нехарактерні дані. Тепер потрібно визначити, що з ними робити. Пропоновані майстром рішення дещо відрізняються для випадків безперервного і дискретного параметра. Відповідний рядок можна видалити (Delete rows containing outliers) або замінити значення параметра на порожнє (Change value to null). Крім того, для безперервних даних (рис. 4.13) можна замінити нехарактерне значення середнім або граничним (зверху чи знизу, в залежності від того, який діапазон відкидається).

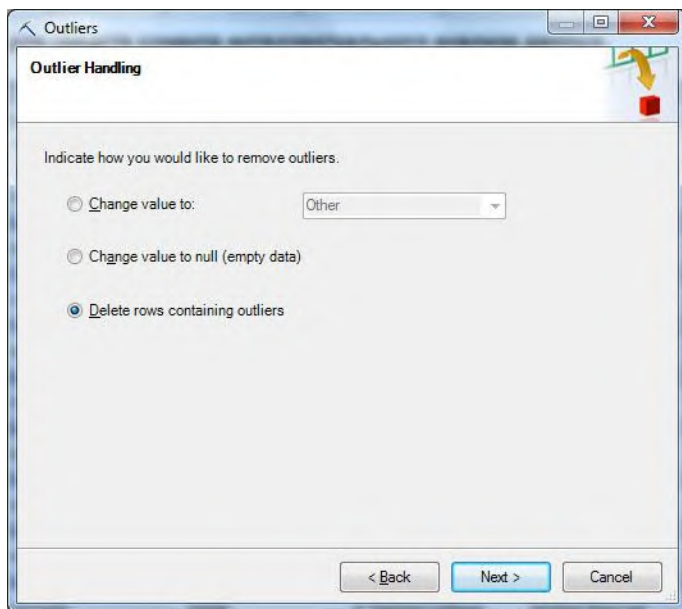


Рис. 4.13 – Використання інструменту Clean Data для виключення викидів

Для дискретного параметра можна вказати значення (з числа вже наявних в наборі), на яке будуть замінюватися “викиди”.

Останнє вікно майстра пропонує вибрати, куди заносити зміни – в вихідні дані (Change data inplace), в їх копію на новому аркуші Excel (Copy sheet data with changes to a new work sheet) або в новий стовпець в вихідній таблиці (Add as a new column to the current work sheet). Останній опція для випадку видалення рядків недоступна.

CleanData.Re-label

У деяких випадках у вихідних даних можуть бути значення, які ускладнюють автоматизований аналіз. Наприклад, є параметр “місто” і серед його значень – Санкт-Петербург, С-Петербург, СПб. Для того, щоб в процесі інтелектуального аналізу цих значень враховувалися коректно, треба їх замінити на одне. Для цього можна використовувати інструмент Re-label. Його ж можна застосувати, якщо потрібно знизити рівень деталізації значень параметра. Треба відзначити, що інструмент працює тільки з дискретними значеннями (ну або розглядаються як дискретні).

Для прикладу, в таблиці з інформацією про клієнтів нам треба зменшити число значень параметра CommuteDistance (відстань щоденних поїздки). Вихідні значення “0-1 Miles”, “1-2 Miles”, “2-5 Miles”, “5-10 Miles”, “10+ Miles” (рис. 4.14).

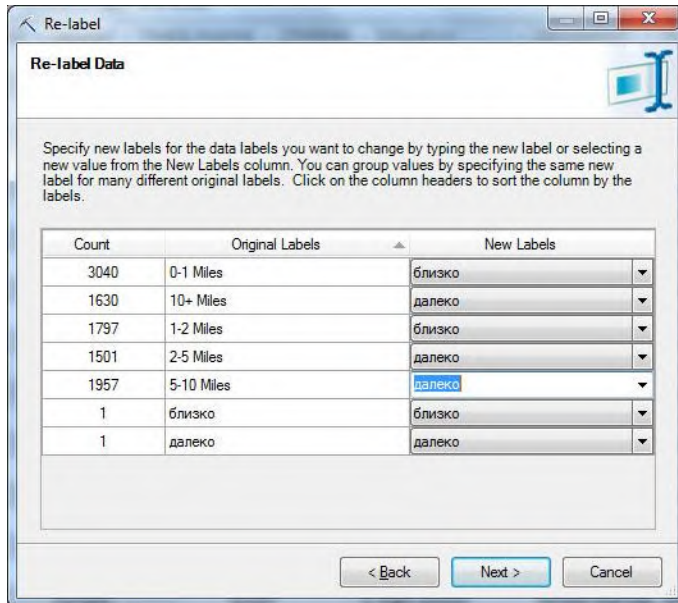


Рис. 4.14 – Вікно із заміною позначень аналітичної служби Re-label Data

Нехай все, що менше 2 миль, буде “близько”, решта – “далеко”. Додамо в таблицю два порожні рядки і вкажемо для однієї CommuteDistance “близько”, для іншого – “далеко”. Робиться це тому, що значення, на які замінюємо, теж повинні бути присутніми в стовпці.

Запустимо інструмент: **CleanData -> Re-label**. Перші два екрани, як і раніше, дозволяють вказати таблицю і стовпець. Далі вказуємо порядок заміни і вибираємо створення нового стовпчика, щоб не втратити вихідні дані. Заміна буде проведена, після чого не забудемо видалити додані порожні рядки з “близько” – “далеко”.

Завдання. Проведіть описану в лабораторній роботі оцінку обраного набору даних.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Опишіть суть спрощеного алгоритму Байєса. Що необхідно для коректного використання алгоритму Байєса?
2. Сформулюйте алгоритм побудови елементарних правил.
3. Яка закономірність лежить в основі наївного байєсівського класифікатора?
4. Про що говорить формула Баєса?
5. Сформулюйте алгоритм наївного байєсівського класифікатора.
6. Сформулюйте алгоритм побудови дерева прийняття рішень.
7. Опишіть суть алгоритму дерев рішень. Наведіть приклади.
8. Наведіть приклад схеми алгоритмів дерева прийняття рішень.
9. Що таке ентропія і як вона застосовується при побудові дерева прийняття рішень?
10. Вкажіть інструменти СКБД Microsoft SQL Server 2012 для побудови баєсівського класифікатора та алгоритму дерева прийняття рішень.

ЛІТЕРАТУРА

1. Хайкин С. Нейронные сети : полный курс, 2-е издание : пер. с англ / С.Хайкин. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
2. Бармак О.В. Інтелектуальний аналіз даних. Методичні вказівки до самостійного вивчення курсу «Інтелектуальний аналіз даних» для студентів напрямку підготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.
3. Нікольський Ю. Древа прийняття рішень та їхнє застосування для прогнозування діагнозу у медицині / Ю. Нікольський, Ю. Щербина, Р. Якимечко // Вісник Львівського університету. Серія прикладна математика та інформатика. – № 6, 2003. – С. 191–211.

РОЗДІЛ 5.

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: ЛІНІЙНА РЕГРЕСІЯ ТА АНАЛІЗ ЧАСОВИХ РЯДІВ

5.1 Метод лінійної регресії

Зв'язок між ознаками може бути функціональним (повним) і кореляційним (статистичним).

Функціональним називають такий зв'язок між ознаками, за якого кожному значенню однієї змінної (аргументу) відповідає тільки одне значення іншої змінної (функції). Такі зв'язки зустрічаються в математиці, фізиці, хімії і в інших науках.

Наприклад, площа кола ($S = \pi R^2$) і довжина кола ($l = 2\pi R$) повністю визначаються радіусом R , площа трикутника – його сторонами.

В соціально-економічних явищах функціональні зв'язки між ознаками зустрічаються рідко. Найчастіше зустрічаються такі зв'язки між змінними величинами, у разі яких числовому значенню однієї з них відповідає декілька значень інших. Такий зв'язок між ознаками називається *кореляційним* (статистичним). Наприклад, відомо, що врожайність залежить від кількості внесених добрив, але на неї впливають і інші фактори (якість ґрунту, шкідники, опади тощо). Крім того, однакова доза добрив під час однакових умов часто по різному впливає на врожайність.

Кореляційний зв'язок є неповним, оскільки він виявляється у випадку великої кількості спостережень та під час порівняння середніх значень результативної і факторної ознак. Він виражається відповідними математичними рівняннями.

В економічних дослідженнях часто розглядається лінійна форма зв'язку. Наприклад, зв'язок між витратами на відпустку та складом родини; витратами на рекламу та обсягом продукції, що випускається.

Ця форма зв'язку виражається рівнянням прямої лінії:

$$y = a + bx, \quad (5.1)$$

де $\tilde{y}$ – вектор розрахункових значень результативної ознаки (залежна змінна): $\tilde{y} = \{ \tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n \}$;

x – вектор спостережень за незалежною змінною (значення факторної ознаки): $x = \{x_1, x_2, \dots, x_n\}$;

a, b – розрахункові параметри регресійної моделі.

Моделі, які встановлюють лінійну залежність між двома змінними і лінійні за своїми параметрами називаються простими лінійними регресійними моделями або лінійними моделями.

Рівняння типу $\tilde{y} = a + bx$, називається рівнянням регресії або кореляційним рівнянням. Головною його метою є встановлення кількісного взаємозв'язку між ознаками.

Алгоритм лінійної регресії дозволяє подати залежність між вхідними і вихідними змінними як лінійну, а потім використовувати отриманий результат під час прогнозування. Подібний приклад зображений на рисунку 5.1.

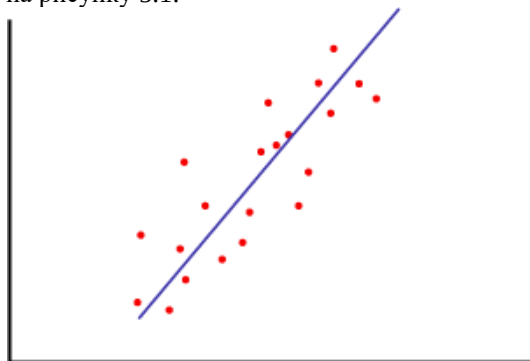


Рис. 5.1 – Графік лінійної регресії

Лінія на діаграмі, що накладається на кореляційне поле (сукупність точок, розташованих у системі координат, що характеризує взаємозв'язок між двома ознаками: факторною x_i та результативною y_i) є найкращим лінійним поданням даних.

У випадку однієї незалежної змінної (одного фактора), задача може бути сформульована наступним чином. Для i -ї точки буде справедливою рівність: $y_i = a + bx_i + e_i$, де e_i – різниця між фактичним значенням y_i і обчисленим $\tilde{y}_i$ у відповідності з рівнянням регресії. Параметри рівняння a і b визначаються методом найменших квадратів, який дає можливість знайти таку теоретичну лінію регресії, яка у порівнянні з іншими проходить найбільш близько до всіх точок кореляційного поля, яке зображує фактичні дані, тобто дає найменшу

суму квадратів відхилень фактичних значень результативної ознаки від вирівняних (теоретичних, розрахункових) значень:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min. \quad (5.2)$$

Іншими словами, кожній точці відповідає відхилення, пов'язане з її відстанню до лінії регресії. Тому, за допомогою підбору коефіцієнтів a і b необхідно отримати таке рівняння, щоб сума квадратів відхилень, пов'язаних з усіма точками, стала мінімальною.

У SQL Server 2012 під час вибору алгоритму лінійної регресії викликається особливий варіант алгоритму дерева рішень з параметрами, які обмежують поведінку алгоритму і вимагають використання певних типів даних на вході.

Лінійна регресія є корисним і широко відомим методом моделювання, особливо для випадків, коли відомий базовий фактор, що призводить до змін, і є підстави очікувати лінійний характер залежності. Існують також і інші типи регресії, в тому числі – нелінійні.

5.2 Методи аналізу часових рядів

Суспільні явища безперервно змінюються. Протягом певного часу – місяць за місяцем, рік за роком – змінюються кількість населення, обсяг і структура суспільного виробництва, рівень продуктивності праці тощо. Аналіз соціально-економічного розвитку – одне з важливих завдань статистики. Інформаційною базою його слугують динамічні (часові, хронологічні) ряди.

Динамічний ряд – це послідовність чисел, які характеризують зміну того чи іншого соціально-економічного явища. Елементами динамічного ряду є перелік хронологічних дат (моментів) або інтервалів часу і конкретні значення відповідних статистичних показників, які називаються рівнями ряду.

Під час вивчення динаміки важливі не лише числові значення рівнів, а і їхня послідовність. Як правило, часові інтервали між рівнями однакові (доба, декада, календарний місяць, квартал, рік тощо). Узявши будь-який інтервал за одиницю, послідовність рівнів записуємо так: $y_1, y_2, y_3, \dots, y_n$. Основною умовою для одержання правильних висновків під час аналізу рядів динаміки є порівнянність його елементів.

Величини тимчасових інтервалів повинні відповідати інтенсивності досліджуваних процесів. Чим більше варіація рівнів у часі, тим частіше варто робити виміри. Відповідно для стабільних процесів інтервали можна збільшити.

Так, перепис населення досить проводити один раз у десять років; облік національного доходу, урожаю ведеться один раз у рік; щодня реєструються курси покупки й продажу валют тощо.

Числові рівні рядів динаміки повинні бути впорядкованими в часі. Не допускається аналіз рядів із пропусками окремих рівнів, якщо ж такі пропуски неминучі, то їх заповнюють умовними розрахунковими значеннями.

Залежно від статистичної природи показника-рівня розрізняють первинні й похідні динамічні ряди, ряди абсолютних, середніх і відносних величин. За ознакою часу динамічні ряди поділяються на інтервальні та моментні. Рівень моментного ряду фіксує стан явища на певний момент часу t , наприклад кількість працюючих на початок року, студентів – на 1 вересня тощо. В інтервальному ряді рівень – це агрегований результат процесу й залежить від тривалості часового інтервалу: виробництво електроенергії за рік, вилов риби за сезон. Зауважимо, що й похідні показники, обчислені на основі інтервальних рядів, на відміну від моментних залежать від протяжності часу (середньодобове чи середньорічне виробництво електроенергії на душу населення).

У загальному випадку, часовий ряд – це набір числових значень, зібраних в послідовні моменти часу (в більшості випадків – через рівні проміжки часу). За приклад можна взяти котирування іноземних валют або інших біржових товарів, результати серії експериментів тощо. Метою аналізу часового ряду може бути виявлення наявних залежностей поточних значень параметрів від попередніх, з подальшим використанням їх для прогнозування нових значень. Виділяють два основних формати представлення часових рядів – стовпчиковий і чередуванням (або змішаний, від англ. mixed) [1].

Розглянемо торгівлю велосипедами, як предметну область. Необхідно подати дані продажів по місяцях двох моделей велосипедів під назвами M200 і R200. В даному випадку маємо два часових ряди. У табл. 5.1 вони подані в стовпчиковому форматі, в табл. 5.2 – використовується чередування.

Таблиця 5.1 – Приклад подання часових рядів у форматі стовпчиків

Рік і місяць	Кількість M200	Кількість R200
2012 04	100	50
2012 05	120	20
2012 06	110	60
...	...	...

Таблиця 5.2 – Приклад подання часових рядів у форматі чередування

Рік і місяць	Модель	Кількість
2012 04	M200	100
2012 04	R200	50
2012 05	M200	120
2012 05	R200	20
2012 06	M200	110
200906	R200	60
...	...	...

З даними в стовпчиковому форматі дещо простіше працювати, але цей формат менш гнучкий. У наведеному прикладі, якщо з якогось моменту з'явилася у продажу третя модель велосипеда, додати дані про неї, у разі використання формату чередування, буде простіше.

Фактичні значення елементів ряду використовуються для навчання моделі, після чого робиться спроба прогнозу для вказаної кількості нових елементів. Приклад подібного аналізу поданий на рисунку 5.2, де пунктирною лінією зображено значення, що передбачаються.

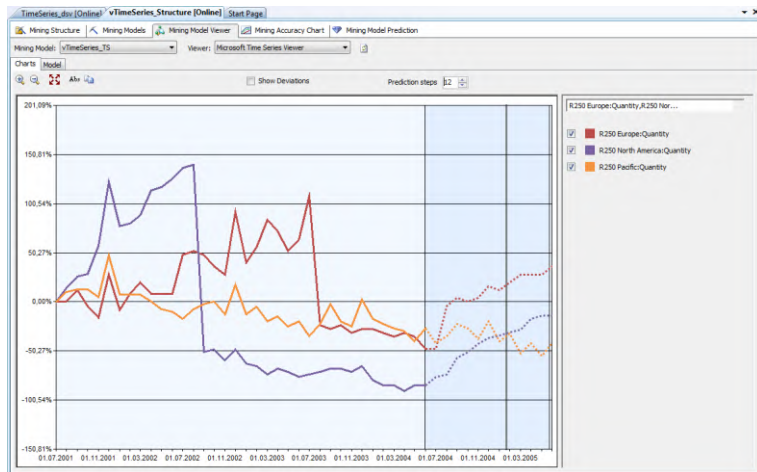


Рис. 5.2 – Вікно з графіками історичних та прогнозних значень часового ряду

Розглянемо тепер деякі особливості реалізації алгоритму в SQL Server 2012. Алгоритм часових рядів є сукупністю двох алгоритмів регресії, оптимізованих для прогнозу рядів неперервних числових значень. Ними є:

- алгоритм «дерево авторегресії з перехресним прогнозуванням» (ARTxp), який оптимізований для прогнозування наступного значення в ряду; він з'явився в SQL Server 2005;

- алгоритм «інтегровані ковзаючі середні авторегресії» (ARIMA), що є галузевим стандартом в даній області; доданий в SQL Server 2012, щоб підвищити точність довгострокового прогнозування.

За замовчуванням служби Analysis Services для навчання моделі використовують кожен алгоритм окремо, а потім об'єднують результати, щоб отримати найбільш точний прогноз. В залежності від наявних даних і вимог до прогнозів, можна вибрати для використання тільки один алгоритм.

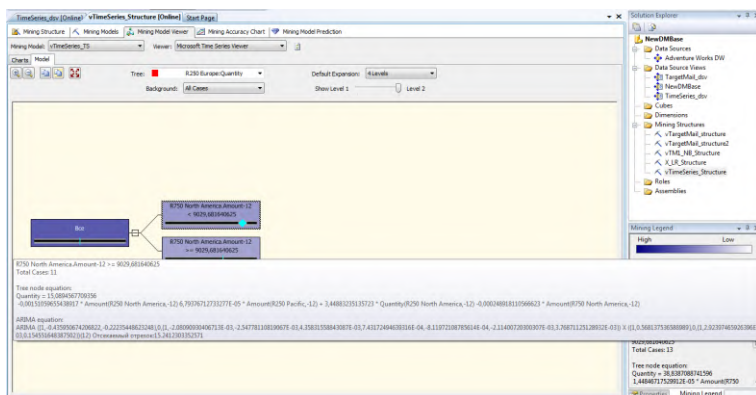


Рис. 5.3 – Вікно перегляду параметрів моделі часового ряду

На рисунку 5.3 подані параметри для одного з вузлів прогнозувальної моделі, яка базується на алгоритмі часових рядів. З рисунку 5.3 видно, що вузол насправді містить параметри для двох алгоритмів. Також у розглянутого часового ряду 'R250 Europe: Quantity' (кількість проданих в Європі велосипедів марки R250) виявлена кореляція з іншим рядом 'R750 North America: Amount' (продажу в Північній Америці велосипедів марки R750).

Облік кореляцій між розглянутими рядами або, інакше кажучи, перехресного впливу рядів, можна відзначити в якості особливості реалізації алгоритму часових рядів в SQL Server.

Це може бути важливо, якщо одночасно аналізуються пов'язані ряди, наприклад, описують ціни на нафту і котирування валют нафтовидобувних країн. Дану корисну можливість підтримує тільки алгоритм ARTxр. Якщо для прогнозу використовується тільки алгоритм ARIMA, перехресний вплив рядів не враховується.

Точність прогнозу для часового ряду може підвищити вказівку відомої періодичності. Наприклад, у даних про продажі магазину спорттоварів по місяцях, швидше за все, буде присутня періодичність 12 (по кількості місяців у році).

Коротко розглянемо основну ідею методу авторегресії. Авторегресія відрізняється від звичайної регресії тим, що поточне значення параметра x_t виражається через його значення в попередні моменти часу. Якщо використовувати лінійні залежності, то алгоритм шукає рішення у вигляді:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \dots + a_n x_{t-n} + \varepsilon_t, \quad (5.3)$$

де ε_t – похибка, яку потрібно мінімізувати шляхом підбору коефіцієнтів $a_1, \dots, a_n$, в чому і полягає навчання моделі. Як видно з рисунку 2.7, наявність кореляції може враховуватися шляхом включення в цю формулу членів з іншого ряду. Періодичність враховують, вводячи в розгляд додаткові члени послідовності. Наприклад, для періодичності 12 це буде x_{t-12} , x_{t-24} тощо. А використання аналітичними службами SQL Server дерева авторегресії дозволяє змінювати формулу шляхом розбиття в точках нелінійності.

Самостійна робота №6. Використання інструментів DMC Excel 2007 для створення моделі інтелектуального аналізу даних

Розглянуті в попередніх самостійних роботах “Надбудови інтелектуального аналізу даних для Microsoft Office” – “Використання інструментів” Prediction Calculator “i” ShoppingbasketAnalysis “” “Засоби аналізу таблиць для Excel” (TableAnalysisTools) для кінцевого користувача багато в чому представляються “чорним ящиком”, який виконує аналіз, але що не дає інформації про те, як отриманий результат. Якщо таке рішення не влаштовує, можна перейти з вкладки Analyze на вкладку Data Mining і скористатися інструментами Data Mining Client для Excel (рис. 5.4).

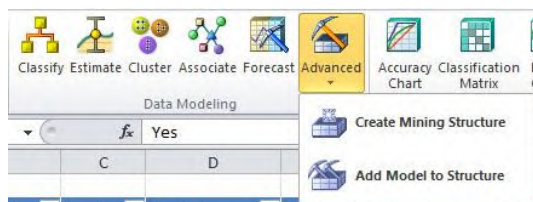


Рис. 5.4 – Вікно з інструментами Data Modeling

В “Використання інструментів Data Mining Client для Excel 2007 для підготовки даних” ми розглянули інструменти, що дозволяють підготувати дані для аналізу.

Класифікація (Classify) створює модель класифікації на основі існуючих даних таблиці Excel, діапазону Excel або зовнішнього джерела даних (Analysis Services Data Source). На основі даних, що обробляються формуються шаблони, які під час використання дозволяють віднести даний приклад до одного з можливих класів. За

замовчуванням використовується алгоритм Decision Trees, але також доступні Logistic Regression, Naive Bayes, Neural Networks.

Оцінка (Estimate) дозволяє створити модель оцінки значення цільового параметра (він повинен бути числовим) на основі даних з таблиці або діапазону комірок Excel або зовнішнього джерела даних. За замовчуванням використовується алгоритм Decision Trees, також доступні Linear Regression, Logistic Regression, Neural Networks.

Кластер (Cluster) запускає майстер, що дозволяє побудувати модель кластеризації на основі даних з таблиці або діапазону Excel, або зовнішнього джерела даних. Модель визначає групи рядків з подібними характеристиками, для чого використовується алгоритм Microsoft Clustering. Дане завдання аналогічна розв'язуваній засобом Detect Categories з набору Table Analysis Tools.

Пошук взаємозв'язків (Associate) допомагає створити модель, що описує взаємозв'язок об'єктів (товарів, що купуються та ін.), яких торкається однієї транзакцією, для чого використовується алгоритм Association Rules. З подібним завданням ми стикалися, використовуючи інструмент ShoppingBasketAnalysis з TableAnalysisTools. Для побудови моделі аналізу необхідно, щоб вихідні дані містили стовпець з ідентифікатором транзакцій і були по ньому відсортовані. Як джерело даних може використовуватися тільки таблиця або діапазон комірок Excel.

Прогноз (Forecast) Даний майстер дозволяє побудувати модель для прогнозування нових значень в числовій послідовності, аналогічно інструменту Forecast в TableAnalysisTools. Використовується алгоритм TimeSeries, для роботи якого потрібна, щоб стовпець (або стовпці), щодо якої буде виконуватися прогноз, риси безперервні числові значення. Також може бути присутнім стовпець з відміткою часу (в цьому випадку, рядки в таблиці повинні бути по ньому відсортовані).

Додатково (Advanced) дозволяє створити структуру інтелектуального аналізу даних або додати в існуючу структуру нову модель (наприклад, для порівняння результатів, які видаються різними алгоритмами аналізу).

Використовуємо інструмент Classify. У поставляється з надбудовами наборі даних (**Файл -> Надбудови інтелектуального аналізу даних -> Зразки даних Excel**) виберемо таблицю **TrainingData**, що містить випадкову вибірку 70% даних з таблиці **SourceData**. Запустимо майстер Classify, в першому вікні якого буде коментар щодо застосування інструменту, а друге вікно дозволить вказати джерело даних для аналізу (таблиця **TrainingData**). Далі буде потрібно описати мета аналізу.

Нехай нас цікавить, чи зробить даний клієнт покупку. У цільовому стовпці вказуємо параметр BikeBuyer (рис. 5.5, вікно зліва), скидаємо в переліку вхідних стовпців позначки навпроти ID (порядковий номер клієнта в базі ніяк не впливає на його рішення про покупку).

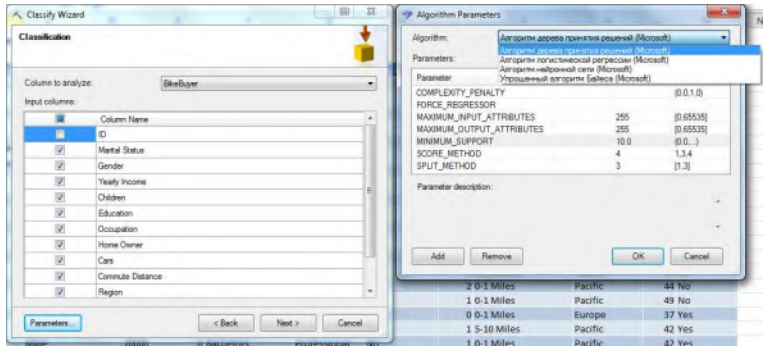


Рис. 5.5 – Вікно із вказівкою параметрів для аналізу

Якщо ID залишити серед аналізованих параметрів, то підсумкова модель може його врахувати. Зокрема, на рисунку 5.6 показано дерево рішень, що враховує значення поля ID в процесі класифікації, що однозначно неправильно.

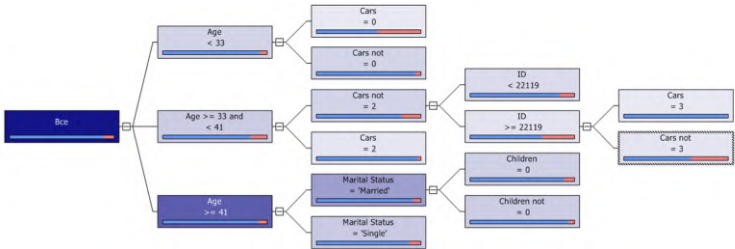


Рис. 5.6 – Приклад невдалого варіанту дерева рішень

Якщо потрібно більш точне налаштування, можна відкрити вікно Parameters і явно вказати використовуваний алгоритм і його параметри (рис. 5.7).

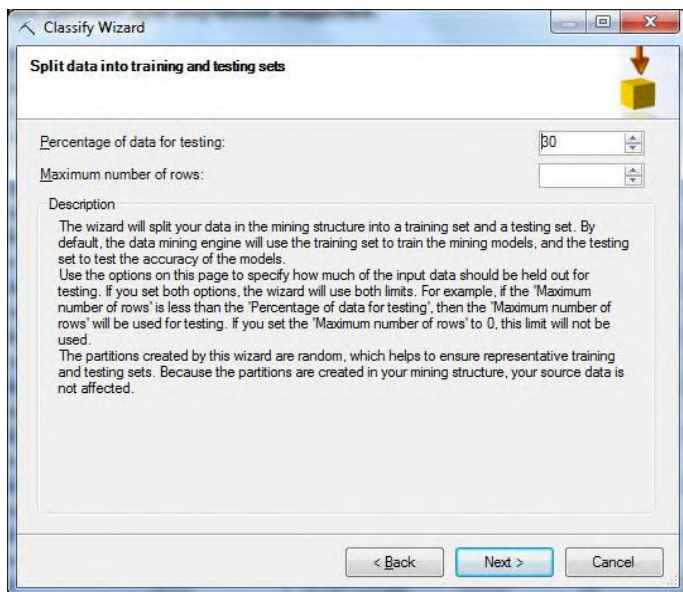


Рис. 5.7 – Вікно розбиття даних на тренувальні та тестові набори даних

Далі майстер запропонує розділити наявні дані на набір для навчання моделі і для її тестування (рис. 5.8). За замовчуванням на набір для тестування виділяється 30% рядків вихідного набору.

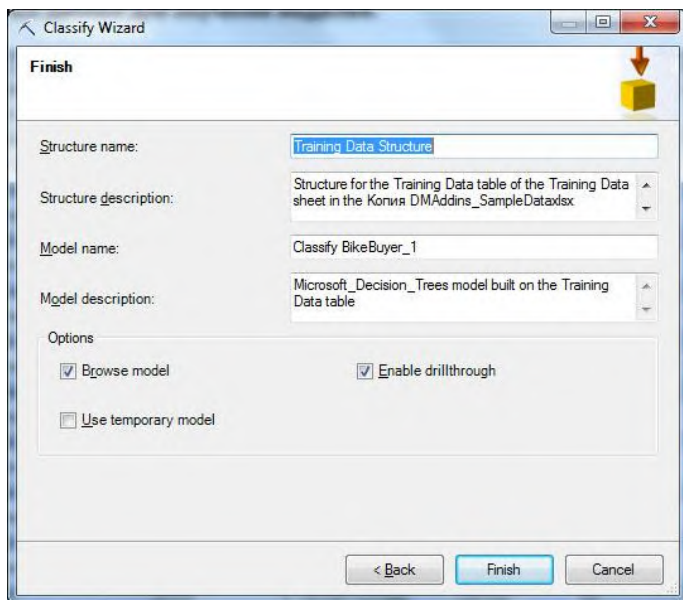


Рис. 5.8 – Вікно розбиття даних і вказівка назв моделі і структури

Останній етап роботи майстра – вказівка імені створюваної структури і моделі (див. рис. 5.8). У нашому прикладі структура буде назватися TrainingDataStructure, а модель Classify BikeBuyer_1. Ці назви нам знадобляться згодом для роботи з моделлю.

Якщо виконує аналіз користувач не має прав адміністратора в базі Аналітичних Служб, то створити постійною модель інтелектуального аналізу на сервері він не зможе. В цьому випадку можна використовувати тимчасову модель, для чого поставити галочку навпроти пункту Use temporary model. Часова модель буде автоматично видалена з сервера по завершенню сеансу роботи користувача.

Зазначена за замовчуванням настройка Brose model вказує на те, що після створення моделі буде відкрито вікно перегляду. Для моделі, створеної з використанням алгоритму DecisionTrees, відображається побудоване дерево рішень і діаграма залежностей. Представлене на рисунку 5.9 дерево рішень дозволяє оцінити побудовану модель.

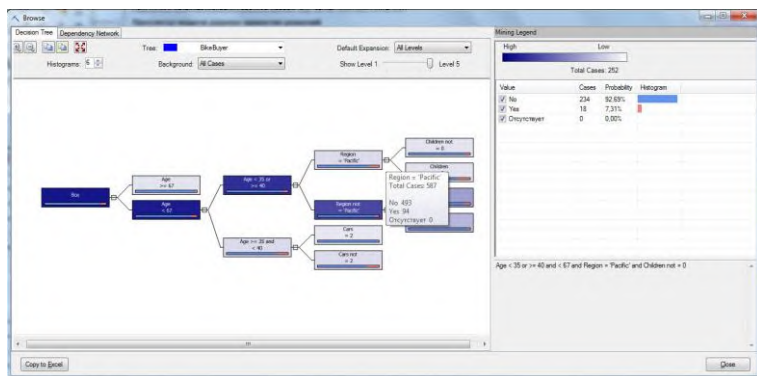


Рис. 5.9 – Вікно схеми побудованого дерева рішень

Розташовані у верхній частині екрану “повзунок” і випадас дозволяють встановити число відображуваних рівнів дерева. Якщо навести курсор миші на точку розгалуження, можна побачити підказку із зазначенням того, скільки і яких випадків в навчальному наборі їй відповідає. Для виділеного вузла в правій частині екрану відображається його опис та гістограма з розподілом значень. Кнопкою **Copy to Excel** можна перенести результат з вікна перегляду на новий лист Excel (для дерева рішень в Excel буде перенесено його растрове зображення).

Клацнувши по вузлу дерева правою клавішею миші і вибравши в контекстному меню **Drill Through Model Columns** ми отримаємо нову таблицю Excel, що містить набір рядків з навчальної вибірки, які відповідають даному вузлу (рис. 5.10).

	Marital Status	Gender	Yearly Income	Children	Education	Home Owner	Car	Commuting Distance	Region	Age	BikeBuyer	HouseIndex
1	Single	Female	60000	2	Bachelors	Professional	Yes	2-5-10 Miles	Pacific	43	Yes	12
2	Single	Female	80000	5	Bachelors	Professional	Yes	4-1-2 Miles	Pacific	41	No	13
3	Single	Female	70000	4	Partial Colleg Skilled Manual	Yes	Yes	2-5-10 Miles	Pacific	44	No	14
4	Married	Male	70000	4	Partial Colleg Skilled Manual	Yes	Yes	2-0-1 Miles	Pacific	44	No	15
5	Married	Male	70000	4	Partial Colleg Skilled Manual	Yes	Yes	3-10-1 Miles	Pacific	46	No	22
6	Single	Female	20000	1	Partial High-S Clerical	Yes	Yes	2-5-10 Miles	Pacific	58	Yes	27
7	Married	Male	100000	1	Bachelors	Management	Yes	3-2-5 Miles	Pacific	47	Yes	69
8	Single	Male	120000	1	Bachelors	Management	Yes	4-2-5 Miles	Pacific	48	Yes	85

Рис. 5.10 – Результат виконання аналітичної служби Drill Through Model Columns

На рисунку 5.11 представлена діаграма залежностей, що показує виявлені взаємозв'язку між параметрами. Її також можна скопіювати в Excel. Виділяючи на діаграмі вузол, можна побачити все що впливає на нього.

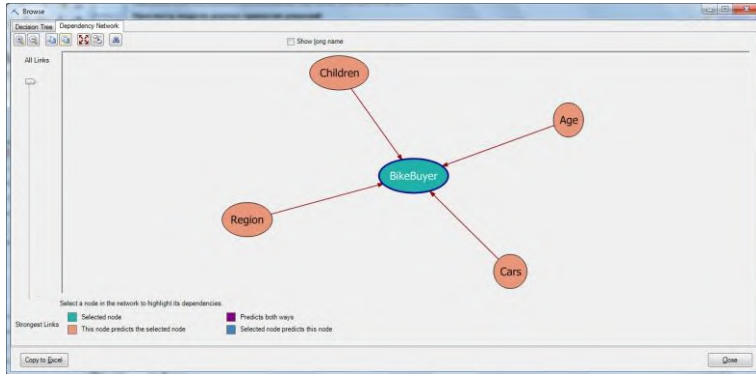


Рис. 5.11 – Вікно діаграми залежностей

Для того щоб управляти наявними на сервері структурами і моделями інтелектуального аналізу, можна скористатися відповідним майстром, запускаються після натискання кнопки **Manage Models** на вкладці **Data Mining** (рис. 5.12). Він дозволяє переглянути наявні структури і моделі, перейменувати їх, видалити непотрібні, виконати інші дії на сервері прямо з **Data Mining Client**.

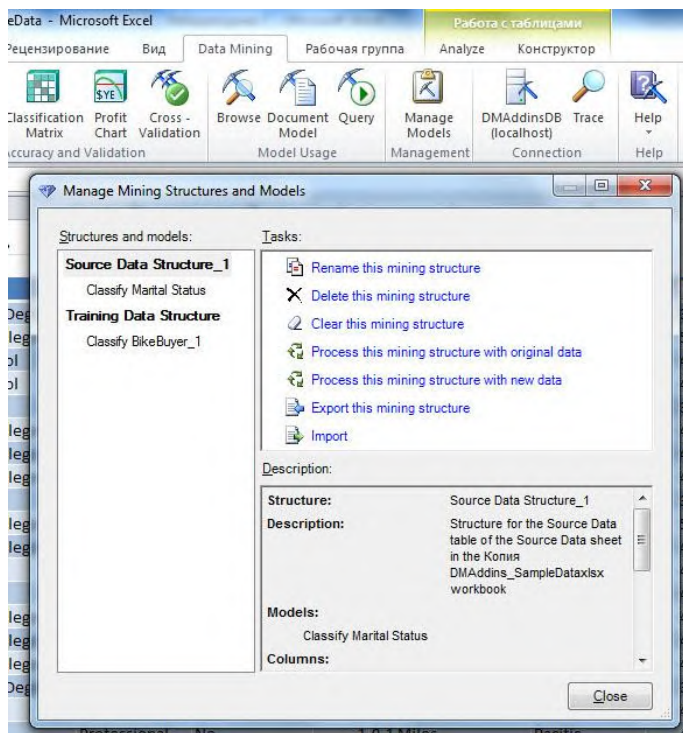


Рис. 5.12 – Вікно управління моделями

Завдання 1. Створіть модель інтелектуального аналізу, що аналогічна описаній в самостійній роботі.

Завдання 2. Скористайтесь набором даних в таблиці Associate і однойменною майстром для створення моделі, яка описує взаємозв'язки між категоріями товарів в одному замовленні. В разі необхідності, скористайтесь довідковою системою по інструменту. Проаналізуйте виявлені правила і діаграму залежностей. Порівняйте з результатами, отриманими в “Використання інструментів” Prediction Calculator “i” ShoppingbasketAnalysis (розділ “Аналіз купівельної поведінки”).

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Опишіть суть алгоритму лінійної регресії. Наведіть приклади.

2. Що таке часовий ряд. Опишіть суть аналізу часових рядів. Які існують формати представлення часових рядів? У чому особливості реалізації часових рядів в MS SQL Server? Наведіть приклади завдань.

ЛІТЕРАТУРА

1. Witten I.H. Data Mining: Practical Machine Learning Tools and Techniques / I.H. Witten, E.Frank. – Morgan Kaufmann, 2005. – 558 p.

РОЗДІЛ 6.

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: МЕТОДИ КЛАСТЕРИЗАЦІЇ, АЛГОРИТМ ВЗАЄМОЗВ'ЯЗКІВ ТА КЛАСТЕРИЗАЦІЯ ПОСЛІДОВНОСТЕЙ

6.1 Методи кластеризації

Під кластеризацією зазвичай мають на увазі процес перевірки набору «точок» та групування точок у «кластери» згідно якоїсь міри довжини. Суть полягає у тому, що точки з одного кластера матимуть невелику відстань одна з одною, а точки з різних кластерів навпаки – велику. Під «точками» мається на увазі n -вимірний вектор характеристик. Приклад кластерів подано на рисунку 6.1.

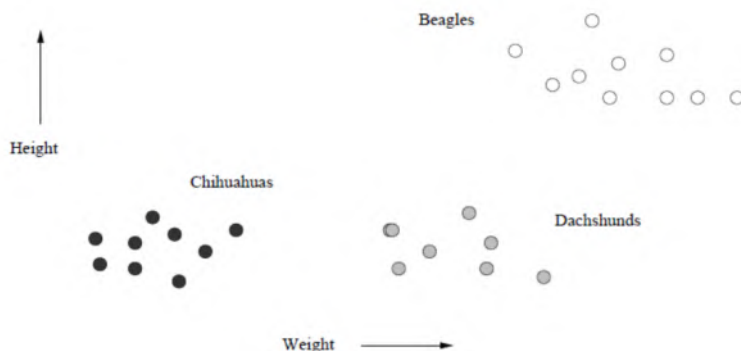


Рис. 6.1 – Приклад кластеризації трьох порід собак за висотою та вагою

Кластеризація (об'єднання в групи схожих об'єктів) є однією із фундаментальних задач Data Mining. Список прикладних областей, де вона застосовується, широкий: сегментація зображень, маркетинг, боротьба з шахрайством, прогнозування, аналіз текстів тощо. Задачу кластеризації в тому чи іншому вигляді формували в таких наукових напрямках, як статистика, розпізнавання образів, оптимізація, машинне навчання. Звідси різноманіття синонімів поняттю кластер – клас, таксон, згущення.

Кластеризація, яку іноді називають «сегментацією», має на увазі виділення з початкової множини даних груп об'єктів зі схожими властивостями і часто виступає першим кроком під час аналізу даних. Розбиття на групи дозволяє спростити роботу з даними. Після

кластеризації застосовуються інші методи, для кожної групи будується окрема модель.

Приклад застосування. Сегментація і побудова профілів клієнтів. За допомогою кластеризації можна виділити сегменти з групами “схожих” об’єктів. Даний алгоритм дає можливість виділити характерні ознаки і персональні переваги клієнтів, оцінити найбільш і найменш дохідні або активні сегменти. Це дозволяє вирішити завдання розробки маркетингових акцій, спрямованих на певні сегменти клієнтів, підвищує ефективність роботи з ними.

У попередніх розділах уже зазначалося, що кластеризація дозволяє знизити розмірність задачі аналізу предметної області, шляхом «природного» групування варіантів у кластери. Таким чином, кластер буде об’єднувати близькі за сукупністю параметрів елементи, і в деяких випадках, його можна розглядати як єдине ціле. Наприклад, описуючи якийсь курортний регіон, можна сказати, що в ньому 250 днів на рік гарна погода, не вказуючи подробиці щодо температури, атмосферного тиску тощо. Таким чином, не вдаючись у характеристики окремих елементів (днів), ми подали в цілому кластер «дні з гарною погодою».

Наданий аналітичними службами SQL Server 2012 алгоритм кластеризації (Microsoft Clustering), використовує ітераційні методи для групування варіантів з такими характеристиками в кластери. Алгоритм спочатку визначає наявні зв’язки в наборі даних і на основі цієї інформації формує кластери.

Ідею можна проілюструвати за допомогою діаграм на рисунку 6.2. На першому етапі (рис. 6.2.а) є множина варіантів, далі (рис. 6.2.б) відбувається ітераційний процес формування кластерів, і в підсумку отримано відносно невеликий набір кластерів, яким можна поставити ідентифікатори і продовжити аналіз.

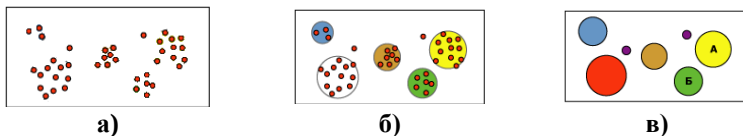


Рис. 6.2 – Перехід від окремих варіантів до кластерів

Microsoft Clustering містить реалізацію двох алгоритмів кластеризації. Перший з них, алгоритм *k*-середніх (англ. *k-means*), реалізує, так звану жорстку кластеризацію. Це означає, що варіант

може належати тільки одному кластеру. Ідея алгоритму полягає в наступному [1]:

1. Вибирається кількість кластерів k .
2. Із вхідної множини даних випадковим чином вибираються k записів, які будуть служити початковими центрами кластерів.
3. Для кожного запису вхідної вибірки визначається найближчий до неї центр кластера. Водночас записи, «притягнуті» певним центром, утворюють початкові кластери.
4. Обчислюються центроїди – центри тяжкості кластерів. Кожен центроїд – це вектор, елементи якого є середніми значеннями ознак, обраховані по всіх записах кластера. Центр кластера зміщується в його центроїд.

Кроки 3 і 4 ітеративно повторюються, Водночас може відбуватися зміна меж кластерів та зміщення їхніх центрів. В результаті мінімізується відстань між елементами всередині кластерів. Зупинка алгоритму робиться тоді, коли межі кластерів і розташування центроїдів не перестануть змінюватися від ітерації до ітерації, тобто на кожній ітерації в кожному кластері буде залишатися один і той же набір записів.

Інший метод, реалізований у Microsoft Clustering, це максимізація очікувань (англ. Expectation-maximization, EM). Він відноситься до методів м'якої кластеризації, тобто варіант в цьому випадку належить до декількох кластерів, а для всіх можливих поєднань варіантів з кластерами обчислюються ймовірності.

Під час кластеризації методом EM [2] алгоритм ітеративно уточнює початкову модель кластеризації, підганяючи її до даних, і визначає ймовірність приналежності точки даних кластеру. Цей алгоритм закінчує роботу, коли ймовірнісна модель відповідає даним. Функція, що використовується для встановлення відповідності, логарифм-функції правдоподібності даних, що вводяться в модель.

Якщо в процесі формуються порожні кластери або кількість елементів в одному або декількох кластерах виявляється менше заданого мінімального значення, нечисленні кластери заповнюються повторно за допомогою нових точок і алгоритм EM запускається знову. Результати роботи алгоритму максимізації очікування є ймовірносними: кожна точка даних належить усім кластерам, але з різною ймовірністю. Оскільки метод допускає перекриття кластерів, сума елементів усіх кластерів може перевищувати кількість елементів навчального набору.

Реалізація Microsoft надає два режими: масштабовану і немасштабовану максимізацію очікування. За замовчуванням, у разі

масштабованої максимізації, очікування проглядається 50000 записів. У разі успіху модель використовує тільки ці дані. Якщо модель не вдається підігнати на підставі 50000 записів, зчитуються ще 50000 записів. Під час немасштабованої максимізації очікування зчитується весь набір даних, незалежно від його розміру. Цей метод створює кластери більш точно, але пред'являє значні вимоги до обсягу пам'яті.

6.2 Алгоритм взаємозв'язків

Алгоритм взаємозв'язків або асоціативних правил (Association Rules) дозволяє виявити поєднання елементів даних, які часто зустрічаються і використовувати виявлені закономірності для побудови прогнозу. Класичний приклад – це аналіз купівельної корзини, коли проводиться пошук товарів, які найбільш часто зустрічаються в одному замовленні (чеку, транзакції), після чого, на основі виявлених закономірностей стає можливою видача рекомендацій.

Ще одним прикладом такого правила, служить твердження, що покупець, що придбає «Хліб», придбає і «Молоко» з імовірністю 75%. Перший алгоритм пошуку асоціативних правил, що називався AIS був розроблений в 1993 році співробітниками дослідницького центру IBM Almaden.

Моделі взаємозв'язків побудовані на наборах даних, що містять ідентифікатори для окремих варіантів і елементів цих варіантів. Група елементів у варіанті називається набором елементів. Модель взаємозв'язків складається з рядів наборів елементів і правил, що описують, як ці елементи групуються в варіантах. Правила, які визначаються алгоритмом, можуть використовуватися для прогнозування ймовірних майбутніх покупок покупців на основі елементів вже наявних в кошику покупця.

Приклад набору правил, які формуються подібним алгоритмом, наведено на рисунку 6.3. Предметна область – торгівля велосипедами та пов'язаними спортивними товарами. На рисунку 6.3 перше правило говорить про те, що якщо в замовленні присутній утримувач для велосипедної фляги та кепка, то з високою ймовірністю буде придбана і сама фляга.

Коли подібний набір правил сформований, його можна використовувати для формування рекомендацій. Наприклад, покупцеві утримувача фляги та кепки, запропонують звернути увагу і на наявні фляги. Питання полягає в тому, як сформувати подібні правила.

Правило
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Рис. 6.3 – Приклад набору правил, що побудовані за алгоритмом Association Rules

Алгоритм взаємозв'язків переглядає набір даних для пошуку елементів, які знаходяться разом у варіанті. Потім алгоритм об'єднує в набори пунктів будь-які пов'язані елементи, що з'являються, як мінімум, у кількості випадків, визначених параметром *MINIMUM_SUPPORT*. Наприклад, набір елементів може бути “Mountain 200 = Existing, Sport 100 = Existing”, і може мати підтримку 710. Потім алгоритм генерує правила з наборів елементів. Ці правила використовуються для прогнозування наявності елемента в базі даних на основі наявності інших специфічних елементів, які алгоритм визначає, як значущі. Наприклад, правило може бути “якщо Touring 1000 = Existing і Road bottle cage = Existing, тоді Water bottle = Existing”, і може мати вірогідність 0,812. У цьому прикладі алгоритм визначає, що наявність у кошику шини Touring 1000 та контейнер для пляшки передбачає, що пляшка для води також, ймовірно, буде в кошику.

6.3 Алгоритм кластеризації послідовностей

Алгоритм кластеризації послідовностей – це унікальний алгоритм, який поєднує в собі аналіз послідовностей і кластеризації. Даний алгоритм можна використовувати для перегляду даних, що містять події, які можуть бути пов'язані в послідовність. Алгоритм знаходить найпоширеніші послідовності і виконує кластеризацію для пошуку ідентичних послідовностей.

Цей алгоритм багато в чому нагадує алгоритм кластеризації. Однак замість пошуку кластерів варіантів, що містять схожі атрибути, алгоритм кластеризації послідовностей знаходить кластери варіантів, що містять схожі шляхи в послідовності.

Принцип роботи алгоритму. Однією з особливостей алгоритму кластеризації послідовностей є те, що він використовує дані послідовностей. Такі дані зазвичай представляють ряд подій або

переходів між станами в наборі даних, наприклад ряд придбань продуктів або клацань мишею на веб-сайті для конкретного користувача. Алгоритм вивчає ймовірність переходів і вимірює відмінності, або відстані, між усіма можливими послідовностями в наборі даних, щоб визначити, які послідовності найкраще використовувати в якості вхідних даних для кластеризації. Після створення алгоритмом списку ймовірних послідовностей він використовує ці відомості в якості вхідних даних для кластеризації методом максимізації очікувань (ЕМ).

Використовуваний аналітичними службами SQL Server 2012 алгоритм Microsoft Sequence Clustering – це гібридний алгоритм, що поєднує методи кластеризації з аналізом марковських ланцюгів. Аналізована множина варіантів формується з використанням вкладених таблиць. У табл. 6.1 поданий умовний приклад подібного варіанту інтелектуального аналізу. Важливо, щоб вкладена таблиця містила власний ідентифікатор, який дозволив би визначити послідовність елементів.

Таблиця 6.1 – Приклад формування варіантів для аналізу звернень до сайту

Ідентифікатор користувача	Розташування	Ідентифікатор послідовності	Тематика
1	Київ	1	Головна сторінка
		2	Велосипеди
		3	Запчастини
		4	Велосипеди

За допомогою марківських моделей аналізується орієнтований граф, який зберігає переходи між різними станами. Алгоритм послідовної кластеризації використовує марківські ланцюги n -го порядку. Число n вказує на те, скільки станів використовувалося для визначення ймовірності поточного стану. У моделі першого порядку ймовірність поточного стану залежить тільки від попереднього стану. У марківського ланцюга другого порядку ймовірність поточного стану залежить від двох попередніх станів, і так далі. Ймовірності переходу між станами зберігаються в матриці переходів. У міру подовження марківського ланцюга, розмір матриці зростає експоненціально, відповідно зростає і час обробки, що необхідно враховувати під час вирішення практичних завдань.

Далі алгоритм вивчає відмінності між усіма можливими послідовностями, щоб визначити, які послідовності краще всього

використовувати в якості вхідних даних для кластеризації. Створений алгоритмом список ймовірних послідовностей використовується в якості вхідних даних для застосовуваного за замовчуванням ЕМ-методу кластеризації (англ. Expectation Maximization, максимізації очікування). Цілями кластеризації є як пов'язані, так і не пов'язані з послідовностями атрибути. У кожного кластера є марковський ланцюг, що представляє повний набір шляхів, і матриця, що містить переходи і ймовірності послідовності станів. На основі початкового розподілу використовується правило Баєса для обчислення ймовірності будь-якого атрибуту, в тому числі – послідовності, в конкретному кластері.

Самостійна робота №7. Побудова моделі кластеризації, трасування і перехресна перевірка

Розглянемо ще ряд можливостей, що надаються надбудовами інтелектуального аналізу даних.

Нехай необхідно провести сегментацію клієнтів Інтернет магазину, список яких знаходиться в файлі Excel. Якщо використовувати TableAnalysisTools, для вирішення цього завдання треба застосувати інструмент DetectCategories (див. “Використання інструментів” AnalyzeKeyInfluencers “i” DetectCategories “”). Також можна скористатися засобами Data Mining Client for Excel, де вибрати інструмент Cluster (рис. 6.4).

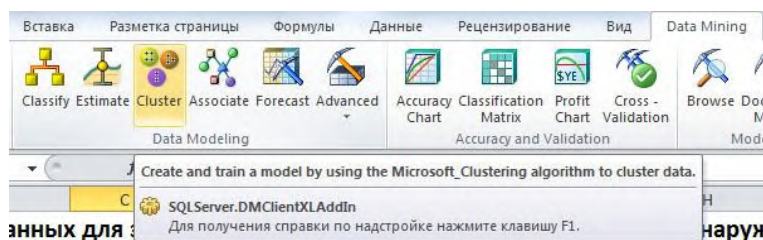


Рис. 6.4 – Вікно з аналітичною службою Cluster

Отже, відкриємо файл зі зразками даних, що йде з надбудовами інтелектуального аналізу, перейдемо на лист TableAnalysisToolsSample (або можна з першого листа зі змістом перейти по посиланню “Зразки даних для засобів аналізу таблиць”) і запустимо інструмент Cluster.

Перше вікно коротко описує суть завдання кластеризації і вказує на те, що для роботи майстра необхідно підключення до Microsoft SQL Server (яке у нас було налаштоване раніше). Наступне

вікно (рис. 6.5) дозволяє вказати джерело даних – в нашому випадку це електронна таблиця Excel, після чого можна вибрати число кластерів (рис. 6.6) або вказати автоматичне визначення, а також використовувані стовпці вхідних даних. Тут скинемо прапорці поруч зі стовпцями ID і PurchasedBike.

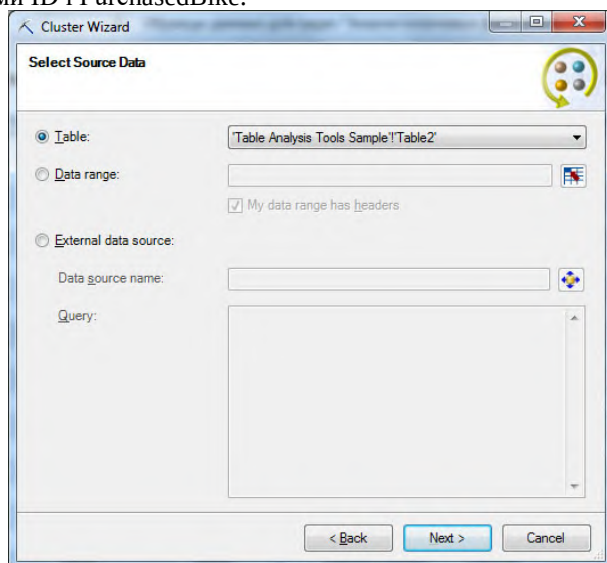


Рис. 6.5 – Вікно для вибору джерел даних майстра кластеризації

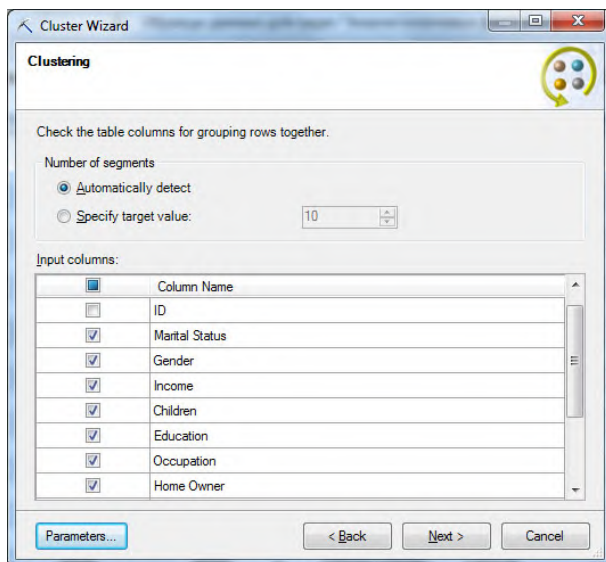


Рис. 6.6 – Діалогове вікно для вибору числа кластерів та використовуваних стовпців вхідних даних

Описаний вище вибір вхідних параметрів обумовлений тим, що стовпець з унікальним ідентифікатором покупця може тільки перешкодити алгоритму кластеризації, а купив клієнт велосипед чи ні, нас зараз не цікавить. Крім того, натиснувши в цьому вікні кнопку **Parameters...** можна отримати доступ до налаштування параметрів алгоритму кластеризації (рис. 6.7) і, наприклад, поміняти буде використовуватися під метод кластеризації.

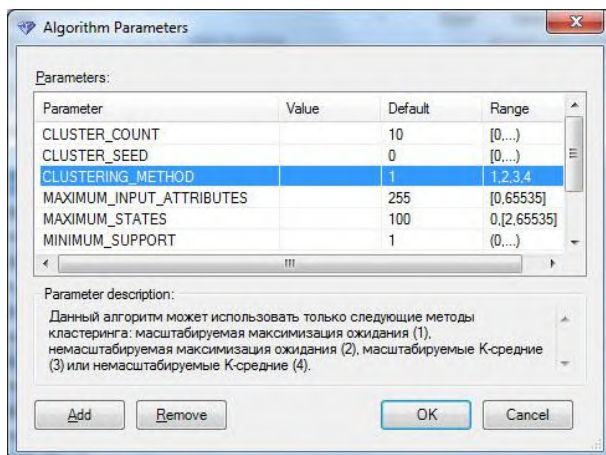


Рис. 6.7 – Діалогове вікно вибору параметрів моделі кластеризації

Більш детально настройки алгоритму кластеризації обговорюються в теоретичній частині курсу. Наступне вікно майстра дозволяє вказати відсоток даних, що резервуються для задач тестування. І нарешті, в останньому вікні майстра (рис. 6.8) можна задати ім'я структури і моделі, вказати, чи відкривати перегляд моделі, чи дозволити деталізацію, чи використовувати тимчасові моделі (за замовчуванням – немає).

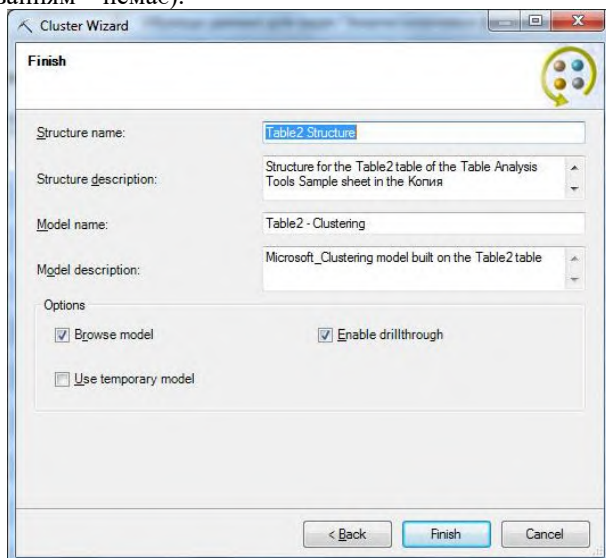


Рис. 6.8 – Діалогове вікно визначення імен структури і моделі кластеризації

Після натискання кнопки **Finish** буде створена структура і модель, після чого модель буде оброблена і відкрита для перегляду у вікні **Browser** (рис. 6.9).

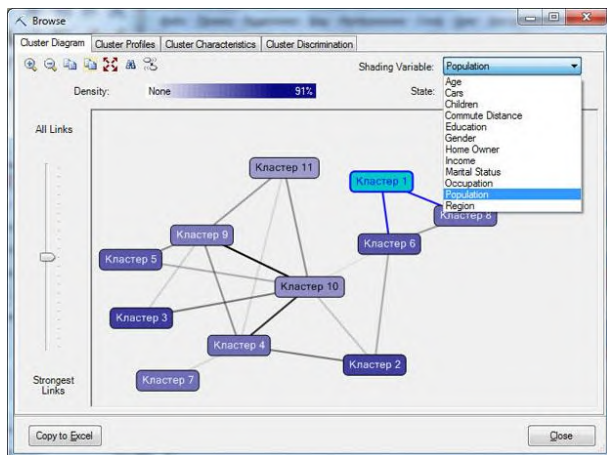


Рис. 6.9 – Вікно з результатами аналітичної служби Cluster

Діаграма кластерів відображає всі кластери в моделі, в нашому прикладі їх 11. Заливка ліній, що з'єднує кластери, показує ступінь їх схожості. Світла або відсутня заливка означає, що кластери не надто схожі. Можна вибрати аналіз за окремим атрибутом або по всій сукупності (Population). Натиснувши кнопку **Copy to Excel** можна отримати зображення на окремий лист таблиці Excel.

Вікно ClusterProfile дозволяє переглянути розподіл значень атрибутів в кожному кластері. Наприклад, на рисунку 6.9 видно, що більша частина клієнтів, віднесених до кластеру 1, проживають в регіоні Europe, а більшість з кластера 3 відноситься до регіону NorthAmerica. Дискретні атрибути представлені у вигляді кольорових ліній, безперервні атрибути у вигляді діаграми ромбів, що представляє середнє значення і стандартне відхилення в кожному кластері. Параметр Histogram bars ("Стовпці гістограми") управляє кількістю стовпців, видимих на гістограмі. Якщо є більше стовпців, ніж вибрано для відображення, то найбільш важливі стовпці зберігаються, а що залишилися групуються в сегмент сірого кольору.

У заголовку під назвою кожного кластера вказує число варіантів, які до нього віднесені. Клацнувши правою клавішею миші на заголовку стовпчика, можна викликати контекстне меню, що дозволяє зокрема перейменувати відповідний кластер. Крім того, з контекстного меню, вибравши опцію Drill Through Model Column, можна отримати деталізацію моделі (результати виводяться на окремий лист Excel). Наприклад, на рисунку 6.10 показані всі варіанти, віднесені до кластеру 1.

	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	RowIndex
4	Single	Male	30000	0	Bachelors	Clerical	No		0 0-1 Miles	Europe	36	5
5	Married	Female	40000	0	Graduate De	Clerical	Yes		0 0-1 Miles	Europe	36	21
6	Single	Male	40000	2	Partial Colleg	Clerical	Yes		0 1-2 Miles	Europe	35	23
7	Single	Male	40000	2	Partial Colleg	Clerical	No		1 0-1 Miles	Europe	34	25
8	Single	Male	30000	0	Partial Colleg	Clerical	No		1 0-1 Miles	Europe	29	27
9	Single	Female	20000	0	Partial High S	Manual	No		2 0-1 Miles	Europe	32	30
10	Married	Male	10000	0	Partial Colleg	Manual	No		1 0-1 Miles	Pacific	26	32
11	Single	Female	20000	0	High School	Manual	No		1 5-10 Miles	Europe	31	33
12	Single	Female	10000	3	Partial High S	Manual	No		2 0-1 Miles	Europe	41	36
13	Single	Female	30000	0	Partial Colleg	Clerical	No		1 2-5 Miles	Europe	30	38
14	Single	Male	20000	0	High School	Manual	No		1 2-5 Miles	Europe	28	39
15	Single	Female	10000	4	Partial High S	Manual	Yes		2 0-1 Miles	Europe	40	40
16	Single	Female	30000	2	Partial Colleg	Clerical	No		0 0-1 Miles	Europe	43	41
17	Married	Female	20000	3	High School	Manual	Yes		0 0-1 Miles	Europe	41	45
18	Single	Female	30000	0	Partial Colleg	Clerical	No		1 0-1 Miles	Europe	28	51

Рис. 6.10 – Результати деталізації моделі

Але повернемося до вікон Modelbrowser. Вікно ClusterCharacteristics дозволяє переглянути найбільш імовірні значення атрибутів для всієї множини варіантів (Population) і для кожного кластера (якщо вибрати кластер в випадковому списку). В останньому випадку, стовпці упорядковано відповідно до ступеня важливості даного атрибута для кластера. Наприклад, в розглянутому вище кластері 1 на першому місці буде знаходитися атрибут Region зі значенням Europe. Водночас, ймовірність того що клієнт, віднесений алгоритмом до цієї категорії, проживає саме в Європі оцінюється як дуже висока.

Тепер перейдемо до аналізу того, що ж відбувається на сервері. У цьому допоможе інструмент **Trace**, розташований в стрічці **Data Mining** в розділі **Connection**. Якщо натиснути на цю кнопку, відкриється вікно, в якому відображається вміст відправляються на сервер запитів (рис. 6.11).

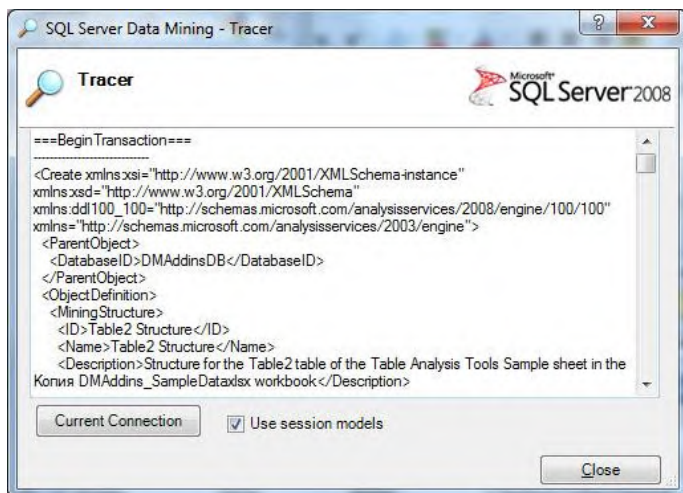


Рис. 6.11 – Діалогове вікно з результатами трасування

Якщо проаналізувати текст запитів, видно, що перша частина транзакції – це опис на XML створюваної структури і моделі, друга частина, яка наводиться нижче – це DMX запит на заповнення структури (і, відповідно, обробку моделі).

```
INSERT INTO MINING STRUCTURE [Table2
Structure] (__RowIndex,
[Marital Status],
[Gender],
[Income],
[Children],
[Education],
[Occupation],
[Home Owner],
[Cars],
[Commute Distance],
[Region],
[Age]) @ParamTable
ParamTable = Microsoft.SQL Server.Data
Mining.Office.Excel.ExcelDataReader
```

Використання трасування дозволяє глибше розібратися в особливостях роботи надбудов інтелектуального аналізу і у разі виникнення помилок виявити їх причини.

Завдання 1. За аналогією з розглянутим прикладом створіть модель кластеризації. Вивчіть і проаналізуйте отримані результати. Відкрийте вікно трасування, проаналізуйте відправляються на сервер запити.

Тепер розглянемо інструмент перехресної перевірки Cross-Validation (треба відзначити, що перехресна перевірка доступна під час використання SQL Server версії Enterprise або Developer). Суть її полягає в тому, що множина варіантів, які використовує модель, розбивається на підмножини (розділи), для кожного з яких проводиться обробка моделі і отримані результати порівнюються з тими, що були на початковій множині варіантів. Якщо результати близькі, можна говорити про вдалу моделі інтелектуального аналізу (вихідних даних вистачило, результат аналізу / прогнозу досить стабільний).

У розділі Accuracy and Validation виберемо інструмент Cross-Validation. Перше вікно майстра повідомляє про суть виконуваної перевірки. У другому вікні (рис. 6.12) проводиться вибір моделі для перехресної перевірки. Зазначимо нашу модель кластеризації – Table2-Clustering.

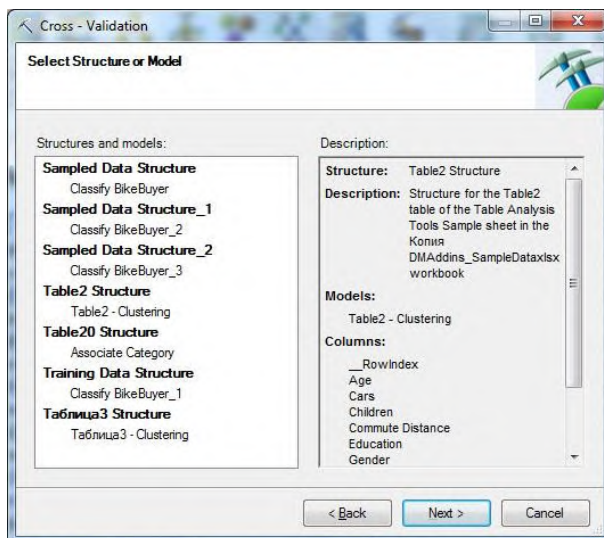


Рис. 6.12 – Вибір моделі для перехресної перевірки

Після вибору моделі потрібно вказати параметри проведеної перехресної перевірки (рис. 6.13).

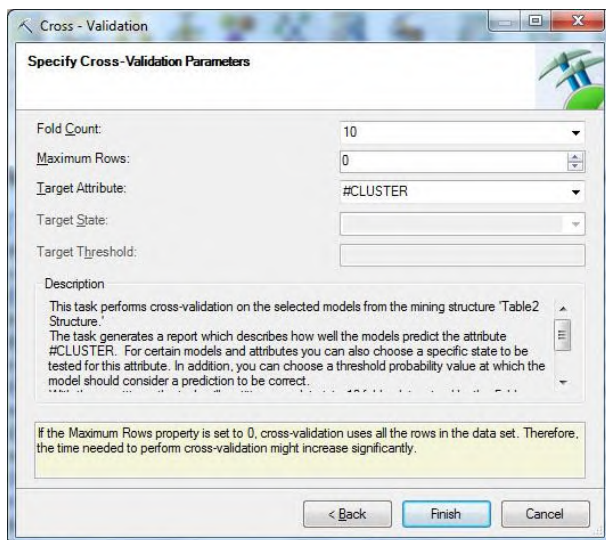


Рис. 6.13 – Діалогове вікно для вибору параметрів перехресної перевірки

Зокрема, вказується число розділів з даними для перехресної перевірки (FoldCount, за замовчуванням 10), максимальне число варіантів, які використовуються під час перевірки (значення MaximumRows = 0 вказує на те, що будуть використовуватися всі; якщо вихідних даних багато, під час використання всіх даних перехресна перевірка може зайняти тривалий час), цільової атрибут (Target Attribute). На рисунку 6.13 зображено Target Attribute # Cluster, тобто номер кластера, до якого належить варіант. Суть перевірки буде полягати в тому, що виконується кластеризація в рамках окремого розділу і отриманий номер кластера, до якого віднесено варіант, буде порівнюватися з номером кластера, отриманим під час обробки моделі з використанням всієї множини варіантів. Збіг говорить про те, що модель хороша (правильно визначені наявні шаблони).

За результатами виконання перехресної перевірки формується звіт (рис. 6.14).

Cross-Validation Report for 'Table2 Structure'
For Target '#CLUSTER'

Models	Table2 - Clustering		
Fold Count	10		
Maximum Rows	0		
Rows Used	700		
Target Attribute	#CLUSTER		

Cross-Validation Summary for Row Likelihood

Model Name	Mean	Standard Deviation
Table2 - Clustering	0,8208	0,0277

Cross-Validation Details

Model Name	Partition Index	Partition Size	Measure	Value
Table2 - Clustering	1	70	Row Likelihood	0,7849
Table2 - Clustering	2	70	Row Likelihood	0,8523
Table2 - Clustering	3	70	Row Likelihood	0,8097
Table2 - Clustering	4	70	Row Likelihood	0,8499
Table2 - Clustering	5	70	Row Likelihood	0,8345
Table2 - Clustering	6	70	Row Likelihood	0,8213
Table2 - Clustering	7	70	Row Likelihood	0,7626
Table2 - Clustering	8	70	Row Likelihood	0,8447
Table2 - Clustering	9	70	Row Likelihood	0,8125
Table2 - Clustering	10	70	Row Likelihood	0,8357
Table2 - Clustering	All	700	Mean (Row Likelihood)	0,8208
Table2 - Clustering	All	700	Standard Deviation (Row Likelihood)	0,0277

Рис. 6.14 – Звіт за результатами перевірки

У ньому показується, скільки варіантів використовувалося для перевірки (на рисунку – 700), які розділи були сформовані (в нашому прикладі 10 розділів по 70 рядків даних), результати проведеного аналізу. Звіт (рис. 6.16) показав, що в середньому, результати, отримані під час аналізу по розділах, більш ніж в 82% випадків збігаються з результатами вихідної моделі. Невеликий розкид значень для різних розділів, вказує на стабільність одержуваного результату, тобто побудована модель інтелектуального аналізу може бути визнана вдалою.

Завдання 2. Виконайте перехресну перевірку для створеної моделі інтелектуального аналізу. Опишіть і проаналізуйте отримані результати.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Опишіть суть алгоритму кластеризації. Які алгоритми кластеризації використовуються в MS SQL Server? Опишіть їхні ідеї? Наведіть приклади завдань.

2. Опишіть алгоритм взаємозв'язків. Що таке підтримка, достовірність і важливість. Наведіть формули. Наведіть приклади завдань.

3. Опишіть суть кластеризації послідовностей. Який алгоритм використовується в MS SQL Server? У чому його суть? Наведіть приклади завдань.

4. В чому основна ідея алгоритму k-найближчих сусідів?

5. Сформулюйте задачу кластеризації

6. В чому різниця між класифікацією і кластеризацією?

7. Які алгоритми KNN ви знаєте і чим вони відрізняються?

8. Коротко опишіть алгоритм взаємозв'язків.

9. Коротко опишіть алгоритм кластеризації послідовностей.

10. Вкажіть інструменти СКБД Microsoft SQL Server 2012 для побудови алгоритму кластеризації та взаємозв'язків.

ЛІТЕРАТУРА

1. Бармак О.В. Інтелектуальний аналіз даних. Методичні вказівки до самостійного вивчення курсу «Інтелектуальний аналіз даних» для студентів напрямку підготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.

2. Башмаков А.И. Интеллектуальные информационные технологии: учеб. пособ. / А.И. Башмаков, И.А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

РОЗДІЛ 7.

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: ШТУЧНІ НЕЙРОННІ МЕРЕЖІ ТА ЛОГІСТИЧНА РЕГРЕСІЯ

7.1 Штучні нейронні мережі

Деякі задачі ІАД, зокрема, задача класифікації, можуть вирішуватися за допомогою різних алгоритмів. У разі наявності в даних складних залежностей між атрибутами, «швидкі» алгоритми інтелектуального аналізу, такі як спрощений алгоритм Баєса, можуть давати недостатньо точний результат. Поліпшити ситуацію може застосування нейромережевих алгоритмів.

Нейронні мережі – це клас моделей, побудованих за аналогією з роботою людського мозку [1-2]. Існують різні типи мереж, зокрема, в SQL Server алгоритм нейронної мережі використовує мережу у вигляді багатшарового перцептрона, до складу якого може входити до трьох шарів нейронів, або перцептронів. Такими шарами є вхідний шар, необов'язковий прихований шар і вихідний шар (рис. 7.1).

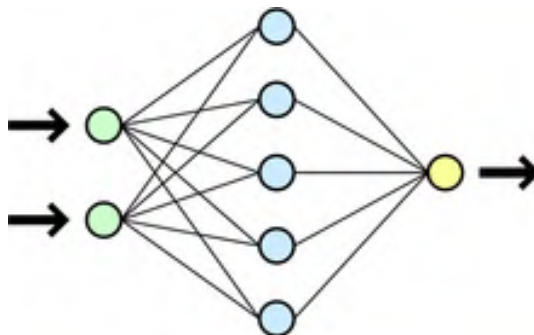


Рис. 7.1 – Схема штучної нейронної мережі; зліва направо: вхідний, прихований та вихідний шар

Незважаючи на велику різноманітність варіантів нейронних мереж, всі вони мають спільні риси. Так, всі вони, так само, як і мозок людини, складаються з великого числа зв'язаних між собою однотипних елементів – нейронів, які імітують нейрони головного мозку. На рисунку 7.2 показана схема нейрона.

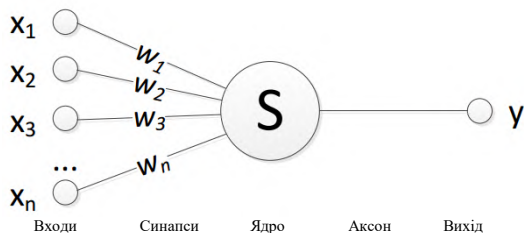


Рис. 7.2 – Схема штучного нейрону

З рисунку 7.2 видно, що штучний нейрон, так само, як і живий, складається із синапсів, що пов'язують входи нейрону з ядром; ядро нейрону, яке здійснює обробку вхідних сигналів і аксона, що пов'язує нейрон з нейронами наступного шару. Кожен синапс має вагу, яка визначає, наскільки відповідний вхід нейрону впливає на його стан. Стан нейрону визначається за формулою:

$$S = \sum_{i=1}^n x_i w_i, \quad (7.1)$$

де n – число входів нейрона;

x_i – значення i -го входу нейрона;

w_i – вага i -го синапса.

Далі визначається значення аксона за формулою:

$$Y = f(S), \quad (7.2)$$

де f – деяка функція, що називається активаційною.

У використовуваному аналітичними службами SQL Server 2012 алгоритмі Microsoft Neural Network, який має більше двох станів дискретний вхідний атрибут моделі інтелектуального аналізу призводить до створення одного вхідного нейрона для кожного стану і одного вхідного нейрона для відсутнього стану (якщо навчальні дані містять які-небудь значення NULL). Неперервний вхідний атрибут «створює» два вхідних нейрона: один нейрон для відсутнього стану і один нейрон для значення самого безперервного атрибуту. Вхідні нейрони забезпечують входи для одного або декількох прихованих нейронів.

Початкові нейрони є значеннями прогнозованих атрибутів для моделі ІАД. Дискретний прогнозований атрибут, який має більше двох

станів, «створює» один вихідний нейрон для кожного стану і один вихідний нейрон для відсутнього або існуючого стану. Неперервні прогнозовані стовпці «створюють» два початкових нейрона: один нейрон для відсутнього або існуючого стану і один нейрон для значення самого неперервного стовпця.

Кожному нейрону зіставлена проста нелінійна функція, яка називається функцією активації, яка описує релевантність або важливість певного нейрона для цього шару нейронної мережі. В якості функції активації в алгоритмі Microsoft Neural Network приовані нейрони використовують функцію гіперболічного тангенса (*Tanh*), а початкові нейрони – сигмоїдальну (логістичну) функцію. Обидві функції є нелінійними і неперервними, що дозволяє нейронній мережі моделювати нелінійні зв'язки між вхідними та вихідними нейронами.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}; \quad (7.3)$$

$$\tanh(Ax) = \frac{e^{Ax} - e^{-Ax}}{e^{Ax} + e^{-Ax}}. \quad (7.4)$$

Навчання моделі ІАД проводиться за наступною схемою. Алгоритм спочатку оцінює навчальні дані і резервує певний відсоток з них для використання у випадку визначення точності мережі.

Потім алгоритм визначає кількість і складність мереж, що включаються в модель ІАД. Визначається кількість нейронів у кожному шарі. Процес навчання будується за наступною схемою [1].

1. На початковій стадії випадковим чином присвоюються значення всім ваговим коефіцієнтам всіх входів в мережі. Значення зазвичай беруться з інтервалу (-1,1).

2. Для кожного навчального варіанту обчислюються виходи.

3. Обчислюються помилки виходів. В якості функції помилки може використовуватися квадрат залишку (квадрат різниці між спрогнозованим і фактичним значенням).

Кроки 2,3 повторюються для всіх варіантів, використовуваних у якості зразків. Після цього, вагові коефіцієнти в мережі оновлюються таким чином, щоб мінімізувати помилки.

У процесі навчання може виконуватися декілька ітерацій. Після припинення росту точності моделі навчання завершується.

7.2 Метод логістичної регресії

Логістична регресія є відомим статистичним методом для визначення впливу декількох чинників на логічну пару результатів. Наприклад, задача може бути наступною. Припустимо, що прогнозований стовпець містить тільки два стани, і необхідно провести регресійний аналіз, зіставляючи вхідні стовпці з ймовірністю того, що прогнозований стовпець міститиме конкретний стан [3]. Результати, отримані методами лінійної та логістичної регресії, подані на рисунку 7.3.a і 7.3.b відповідно. Лінійна регресія не обмежує значення функції діапазоном від 0 до 1, незважаючи на те, що вони повинні бути мінімальним і максимальним значеннями цього стовпця. Крива, формована алгоритмом логістичної регресії, в цьому випадку більш точно описує досліджувану характеристику.

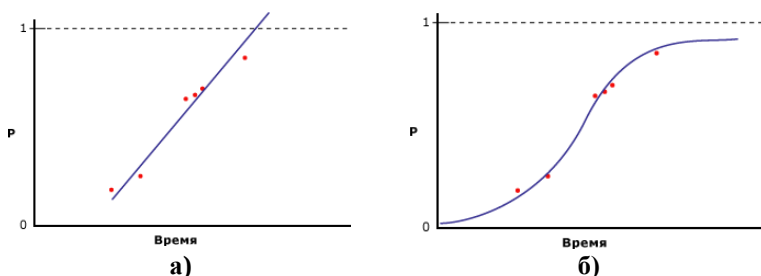


Рис. 7.3 – Порівняння результатів, що одержані за методами лінійної (а) і логістичної регресії (б)

У реалізації Microsoft для моделювання зв'язків між вхідними та вихідними атрибутами застосовується видозмінена нейронна мережа, в якій відсутній прихований шар. Вимірюється внесок кожного вхідного атрибута, і в закінченій моделі різні входи забезпечуються ваговими коефіцієнтами. Назва «логістична регресія» відображає той факт, що крива даних стискається шляхом застосування логістичного перетворення, щоб знизити ефект екстремальних значень.

Самостійна робота №8. Аналіз точності прогнозу і використання моделі інтелектуального аналізу

У попередній самостійній роботі ми створили модель для класифікації клієнтів магазину з метою визначити, чи зробить даний клієнт покупку чи ні. Наступне завдання – оцінити точність

побудованої моделі інтелектуального аналізу. Для цього можна використовувати інструменти з групи Accuracy and Validation (в російській версії – Точність і Правильність).

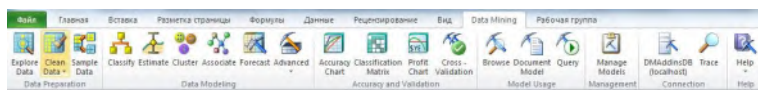


Рис. 7.4 – Інструменти Data Mining Client

Діаграма точності (Accuracy Chart) дозволяє, застосувавши модель на тестовій вибірці даних, оцінити результати її роботи. Під час виконання “Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних” була створена структура Training Data Structure і модель класифікації Classify BikeBuyer_1. Під час створення моделі ми резервували 30% даних для цілей тестування (рис. 7.5).

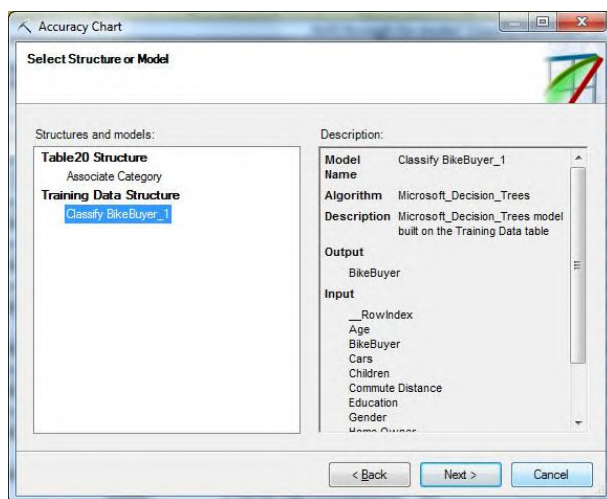


Рис. 7.5 – Діалогове вікно аналітичної служби Accuracy Chart (1)

Запустимо інструмент AccuracyChart. Первое вікно майстра містить короткий опис інструменту, в наступному – треба вказати структуру або модель (рис. 7.6). Якщо для однієї структури визначені кілька моделей, по діаграмі можна буде провести їх порівняльний аналіз.

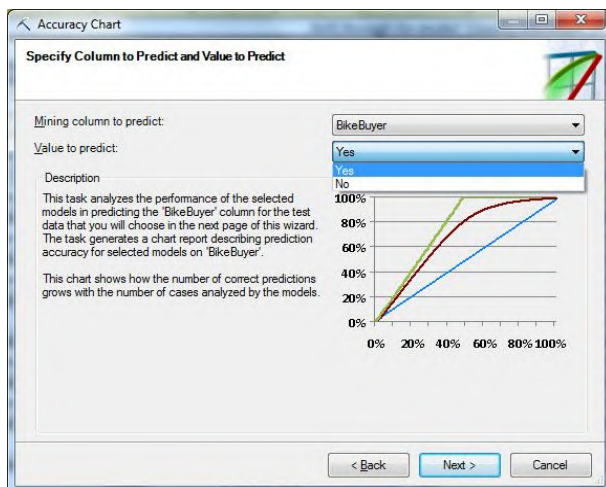


Рис. 7.6 – Діалогове вікно аналітичної служби Accuracy Chart (2)

Наступне вікно (рис. 7.7) служить для вибору прогнозного параметра і його значення.

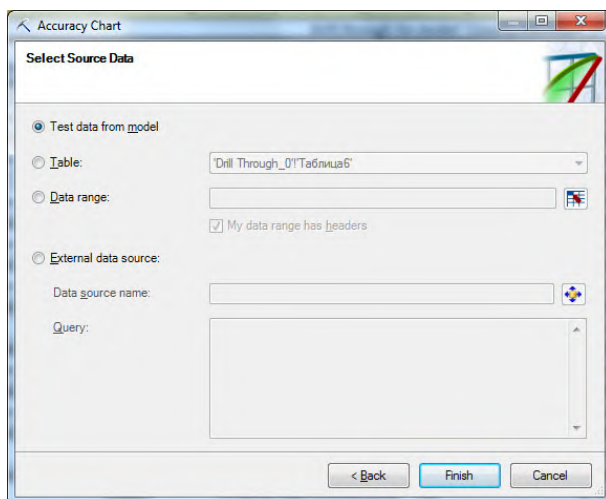


Рис. 7.7 – Діалогове вікно аналітичної служби Accuracy Chart (3)

У нашому випадку параметр – BikeBuyer, а оцінювати будемо точність передбачення значення “Yes”. Далі потрібно вказати джерело даних для тестування. Це можуть бути зарезервовані під час створення

моделі дані, дані з таблиці або діапазону комірок Excel, або з зовнішнього джерела. Зараз виберемо дані з моделі. У разі зазначення таблиці Excel треба описати відповідність стовпців в моделі і використовуваної для тестування таблиці (рис. 7.8).

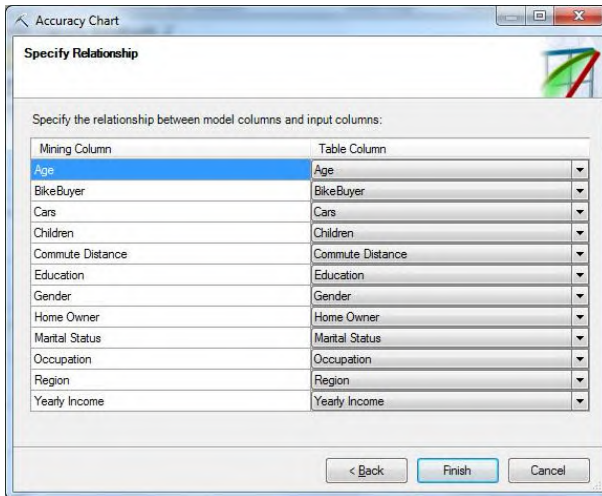


Рис. 7.8 – Діалогове вікно аналітичної служби Accuracy Chart (4)

Після цього будуть сформовані і поміщені на новий лист Excel діаграма точності і таблиця зі значеннями, представленими на діаграмі (рис. 7.9).

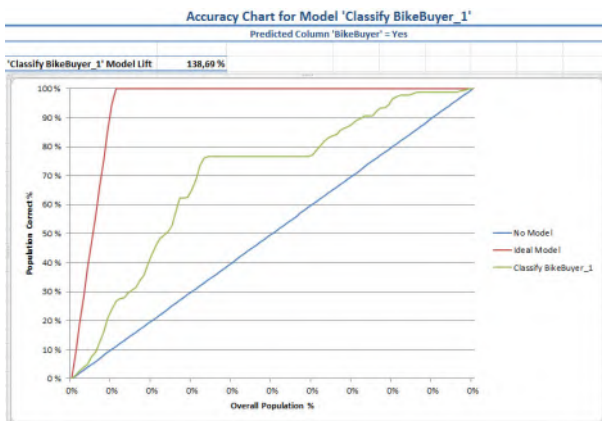


Рис. 7.9 – Результат аналітичної служби Accuracy Chart

На діаграмі червона лінія відповідає ідеальній моделі, світло-зелена – нашій моделі, нижня (синя) лінія – лінія випадкового вибору, завжди йде під кутом 45 градусів.

Дані на діаграмі і в таблиці можна інтерпретувати в такий спосіб. Нехай нам необхідно вибрати всіх клієнтів, які зроблять покупки. Формована ідеальною моделлю вибірка обсягом в 11% від числа вихідних записів буде включати всі 100% потрібних записів (в тестовій множині їх трохи менше 11%). Випадкова вибірка обсягом в 11% містить 11% потрібних записів, а вибірка такого ж обсягу, яка формується нашою моделлю – 26,58%. До вибірки в 25% від загального обсягу даних, наша модель помістить 52,7% “правильних” клієнтів та ін. Якість прогнозу падає (горизонтальну ділянку зеленого графіка) після виявлення 76% цікавлять випадків. Під час візуального аналізу – чим ближче графік оцінюваної моделі до ідеального, тим точніший прогноз вона видає.

Percentile	Ideal Model	Classify BikeBuyer_1
0 %	0,00 %	0,00 %
1 %	9,46 %	1,35 %
2 %	18,92 %	2,70 %
3 %	28,38 %	4,05 %
4 %	37,84 %	4,95 %
5 %	47,30 %	7,66 %
6 %	56,76 %	9,46 %
7 %	66,22 %	12,61 %
8 %	75,68 %	16,22 %
9 %	85,14 %	20,72 %
10 %	94,59 %	23,87 %
11 %	100,00 %	26,58 %
12 %	100,00 %	27,48 %
13 %	100,00 %	27,93 %
14 %	100,00 %	29,28 %
15 %	100,00 %	30,63 %
16 %	100,00 %	31,53 %
17 %	100,00 %	33,78 %
18 %	100,00 %	35,59 %
19 %	100,00 %	39,19 %
20 %	100,00 %	42,79 %
21 %	100,00 %	45,95 %
22 %	100,00 %	48,20 %
23 %	100,00 %	49,55 %
24 %	100,00 %	50,45 %
25 %	100,00 %	52,70 %

Рис. 7.10 – Фрагмент таблиці з оцінками точності прогнозу

Завдання 1. Побудувати діаграму точності аналогічну тій, що представлена вище (використовуваний файл – “Зразки даних Excel”).

Додатково побудуйте діаграму для BikeBuyer = “No”. Поясніть відмінність у зовнішньому вигляді графіків.

Завдання 2. У попередньому завданні для цілей тестування використовувалися дані з моделі. Модель формувалася на даних з таблиці TrainingData. У таблиці TestingData знаходяться 30% даних з вихідного набору SourceData. Перевірте точність моделі на наборі TestingData.

Проблема, з якою ми зіткнулися, могла бути виявлена і раніше, якщо уважно подивитися на побудоване дерево рішень (рис. 7.11). Проте в такому випадку не вдалося б продемонструвати можливості Data Mining за оцінкою точності моделі.

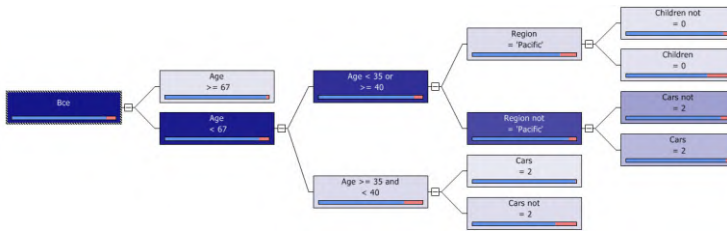


Рис. 7.11 – Схематична модель дерева рішень

З рисунку 7.11 видно, що всі кінцеві вузли дерева дають рішення BikeBuyer = “No” (йому відповідає синя смужка на діаграмі характеризує розподіл відповідей в навчальній вибірці). Відповіді “Yes” відповідає більш коротка червона смужка, що говорить про те, що підтримують такий результат прикладів було менше. Очевидно це пов’язано з тим, що таких прикладів і взагалі меншість в даному наборі (близько 10%).

Спробуємо використовувати навчальний набір більшого обсягу і з більш часто зустрічається значенням BikeBuyer = “Yes”. Відкриємо таблицю SourceData, де даних більше. Але відсоток цікавлять нас записів залишається таким же (це можна визначити за допомогою інструменту ExploreData). Тому скористаємося інструментом SampleData (“Використання інструментів Data Mining Client для Excel 2007 для підготовки даних”), щоб сформувати “надлишкову” вибірку з 2000 року рядків, де в 30% випадків BikeBuyer = “Yes”. У отриманого набору залишимо автоматично призначений назву SampledData. За допомогою інструменту Classify побудуємо модель аналогічна тій, як це було зроблено в “Використання інструментів Data Mining Client для Excel 2007 для створення моделі

інтелектуального аналізу даних” (алгоритм – DecisionTrees, цільової параметр BikeBuyer, стовпець ID під час аналізу не враховуємо, інші настройки за замовчуванням). Отримане дерево рішень представлено на рисунку 7.12.

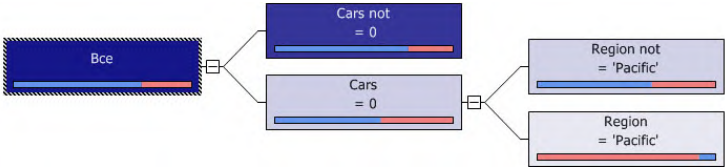


Рис. 7.12 – Нове дерево рішень

Воно простіше попереднього, але в залежності від значень параметрів може давати як прогноз “Yes”, так і “No”. “Yes” буде в тому випадку, якщо у клієнта 0 машин і він з регіону “Pacific”.

На основі нового набору даних також створимо модель для класифікації на сонові алгоритму Neural Networks (нейронних мереж). Якщо побудувати для них матриці класифікації (рис. 7.13) буде видно, що модель на основі нейронних мереж дає більш точний прогноз.

Counts of correct/incorrect classification for model 'Classify BikeBuyer'			
Predicted Column 'BikeBuyer'			
Columns correspond to actual values			
Rows correspond to predicted values			
Model name:	Classify BikeBuyer	Classify BikeBuyer	
Total correct:	71,40 %	1428	
Total misclassified:	28,60 %	572	
Results as Percentages for Model 'Classify BikeBuyer'			
	▼ No(Actual)	▼ Yes(Actual)	▼
No	99,86 %	95,00 %	
Yes	0,14 %	5,00 %	
Correct	99,86 %	5,00 %	
Misclassified	0,14 %	95,00 %	
Results as Counts for Model 'Classify BikeBuyer'			
	▼ No(Actual)	▼ Yes(Actual)	▼
No	1398	570	
Yes	2	30	
Correct	1398	30	
Misclassified	2	570	

Рис. 7.13 – Матриця класифікації для дерева рішень

З огляду на відносно високий відсоток помилок, ні та, ні інша модель, напевно, не може бути визнана вдалою (рис. 7.14).

Counts of correct/incorrect classification for model 'Classify BikeBuyer_3'			
Predicted Column 'BikeBuyer'			
Columns correspond to actual values			
Rows correspond to predicted values			
Model name:	Classify BikeBuyer_3	Classify BikeBuyer_3	
Total correct:	75,67 %	7567	
Total misclassified:	24,33 %	2433	
Results as Percentages for Model 'Classify BikeBuyer_3'			
	▼ No(Actual)	▼ Yes(Actual)	▼
No	78,70 %	51,60 %	
Yes	21,30 %	48,40 %	
Correct	78,70 %	48,40 %	
Misclassified	21,30 %	51,60 %	
Results as Counts for Model 'Classify BikeBuyer_3'			
	▼ No(Actual)	▼ Yes(Actual)	▼
No	7083	516	
Yes	1917	484	
Correct	7083	484	
Misclassified	1917	516	

Рис. 7.14 – Матриця класифікації для нейронних мереж

Завдання 3. Проведіть описані в роботі дії. Прокоментуйте результати.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Опишіть суть нейромережових алгоритмів. Як відбувається процес навчання нейронної мережі. Наведіть приклади завдань.
2. Що таке штучні нейронні мережі?
3. Яка будова перцептрона?
4. Що таке активаційна функція?
5. Функції якого типу найчастіше застосовуються у якості активаційної функції і чому?
6. Яким чином штучні нейронні мережі застосовуються для розв'язання задачі класифікації?
7. Метод зворотного поширення помилок. Основна ідея. Основні недоліки та способи їх усунення.
8. Опишіть суть логістичної регресії. Який алгоритм використовується в MS SQL Server? Наведіть приклади завдань.
9. Як буде виглядати процес навчання логістичної регресії?
10. Вкажіть інструменти СКБД Microsoft SQL Server 2012 для побудови баєсівського класифікатора та алгоритму дерева прийняття рішень.

ЛІТЕРАТУРА

1. Хайкин С. Нейронные сети : полный курс, 2-е издание : пер. с англ / С.Хайкин. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
2. Любунь З.М. Основи теорії нейромереж / З.М. Любунь // Текст лекцій. – Львів: Видавничий центр ЛНУ імені Івана Франка, 2007. – 142 с.
3. Lohr S. Data-Ism: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else / S. Lohr. – HarperBusiness, 2015. – 489 p.

РОЗДІЛ 8.

КОНЦЕПЦІЇ МОВИ DATA MINING EXTENSIONS

8.1 Базові поняття мови Data Mining Extensions

У попередніх розділах курсу ми познайомилися з деякими важливими поняттями ІАД. Тепер прийшов час зібрати їх разом і розглянути, як їх опис може бути зроблено за допомогою мови ІАД – *Data Mining Extensions (DMX)*, яка використовується для запитів до служби Analysis Services [1].

Найменшою логічною одиницею роботи з даними під час інтелектуального аналізу є атрибут, який містить деяку «елементарну» інформацію про приклад, що аналізується. Наприклад, вік клієнта. Для алгоритмів Data Mining існує два основних типи атрибутів [2]:

- *категоріальні* (дискретні), приймаючі значення з деякого фіксованого кінцевого набору значень;
- *неперервні* числові атрибути.

Додаткові типи атрибутів засновані на базових. До них, зокрема, відноситься *впорядкований* тип (англ. ordered). Такий атрибут є категоріальним, але для нього заданий певний порядок значень (наприклад, розміри одягу). *Дискретизованні* атрибути – це спеціальний варіант категоріального типу, отриманий з неперервного шляхом розбиття на діапазони. Наприклад, спрощений алгоритм Баєсса не може обробляти неперервні атрибути, тому буде потрібна дискретизація. Термін «тип вмісту» далі використовуватиметься як синонім терміну «тип атрибута».

З кожним категоріальним атрибутом пов'язаний набір його значень (або *станів*). Наприклад, атрибут «Місто проживання» може приймати значення: «Київ», «Хмельницький» тощо. На етапах підготовки і вивчення даних важливо провести аналіз множини станів атрибутів і, за потреби, внести корективи.

Наприклад, якщо в одних випадках місто записано як «Кам'янець-Подільський», а в інших «К-Подільський», потрібно привести все до єдиного формату (інакше алгоритм інтелектуального аналізу буде розглядати це як різні стани). Крім того, можливо для аналізу не потрібна настільки подібна деталізація по містах, а інтерес представляє, чи є людина жителем великого міста чи ні.

У випадку аналітичних служб Microsoft SQL Server, атрибути також можуть мати стани «Missing» і «Existing». Перший вказує на те,

що в рядку даних атрибут відсутній (або стан не визначено), інший – що атрибут присутній.

Варіант визначається як окремий приклад, що надається алгоритмом ІАД. Він складається з набору атрибутів з відповідними значеннями і у багатьох випадках описує об'єкт чи подію. Нерідко варіант можна подати рядком в таблиці, стовпці якої – атрибути. Нехай, наприклад, вирішується завдання класифікації потенційного банківського позичальника під час ухвалення рішення про видачу кредиту. У цьому випадку, варіант містить відомі дані про цього клієнта (розмір заробітку, професія, місце роботи тощо), що цілком представимо у вигляді рядка таблиці.

У той же час, Microsoft SQL Server і мова DMX дозволяють використовувати вкладені таблиці, що дозволяє описувати більш складні за структурою варіанти. Це можна проілюструвати на наступному прикладі [3]. Нехай потрібно провести аналіз товарів, куплених кожним з клієнтів. Водночас є можливість зв'язати дані про покупки і про покупця. Початкові дані беруться з двох реляційних таблиць, пов'язаних зовнішнім ключем (рис. 8.1): таблиця Customers (Замовники) з полями CustomerKey (Ключ Замовника), FirstName (Ім'я), LastName (Прізвище), CommuteDistance (Відстань до роботи) і таблиця Products (Товари) з полями CustomerKey (зовнішній ключ, посилається на Customers) і ModelName (Назва Моделі). Як видно на рисунку, сформований варіант може містити багато товарів, куплених одним клієнтом, і відрізняється за виглядом від рядка реляційної таблиці.

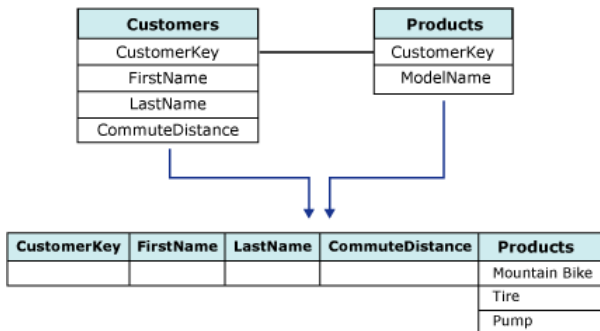


Рис. 8.1 – Схема формування варіантів під час використання вкладених таблиць

Ключ варіанта використовується для ідентифікації варіанта. Найчастіше в цій якості може використовуватися вихідний ключ

таблиці, з якої беруться дані для аналізу. На рисунку 8.1 ключем може виступати стовпець CustomerKey.

Вкладений ключ дозволяє ідентифікувати об'єкт, описуваний у вкладеній таблиці. У нашому прикладі це буде назва товару (якщо б там були ще якісь атрибути – ціна, колір тощо – вони відносились б саме до цього товару).

Атрибут може розглядатися алгоритмом інтелектуального аналізу в якості входу, виходу або входу і виходу одночасно. Мова DMX дозволяє це вказати в процесі опису моделі. На стадії навчання алгоритмом надаються як вхідні, так і початкові дані. На стадії прогнозування – алгоритм отримує вхідні дані і повертає початкові.

Тепер розглянемо питання отримання вхідних даних для інтелектуального аналізу. Аналізувати можна дані з реляційних таблиць та інших джерел, якщо вони спеціальним чином описані в якості *представлення джерела даних* в службах Analysis Services. Спочатку визначається *джерело даних* (Data Source), а потім його *уявлення* (Data Source View). Подання джерела даних дозволяє поєднувати різні джерела даних і працювати з вкладеними таблицями. Один із способів визначити джерело даних – використання відповідного майстра в середовищі BI Dev Studio.

Служби Analysis Services зчитують дані з джерела в спеціальний кеш. Вміщені в кеш дані можна зберегти і використовувати під час створення інших моделей інтелектуального аналізу або видалити, щоб звільнити місце в сховищі.

Дані можуть братися не тільки з реляційних баз даних, але і з аналітичних кубів (англ. cube).

Наступний крок – створення структури. Це поняття вже траплялося раніше, але розглянемо його ще раз більш докладно. *Структура ІАД* може бути подана як сукупність вхідних даних і опису способів їх обробки. Структура містить моделі, що використовуються для аналізу даних.

Під час створення структури (або моделі одночасно зі структурою) в Microsoft SQL Server Analysis Services, необхідно визначити типи даних для кожного стовпця в структурі [4]. Тип даних повідомляє модулю інтелектуального аналізу, чи є дані в джерелі числовими або текстовими, і як їх обробляти. Типи даних перераховані в табл. 8.1: текстовий (Text), дійсний числовий (Long), «довгий» дійсний числовий (Double), логічний (Boolean) і дата (Date). Вони стандартні і особливих коментарів не потребують. Єдине, хотілося б звернути увагу на те, що типів даних в DMX менше, ніж в SQL, і в процесі підготовки даних з реляційних таблиць до аналізу, часто

робиться перетворення до найбільш властивого типу. Кожен тип даних підтримує кілька типів вмісту. Задаючи тип вмісту, можна налаштовувати метод, яким дані в стовпці обробляються (або обчислюються) в моделі інтелектуального аналізу.

Таблиця 8.1 – Типи даних і типи вмісту мови DMX

Тип даних	Підтримувані типи вмісту
Text	Cyclical, Discrete, Key, Key Sequence, Ordered, Sequence
Long	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered, Sequence, Time Classified
Boolean	Cyclical, Discrete, Ordered
Double	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered, Sequence, Time Classified
Date	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered

У табл. 8.1 вказано, які типи вмісту можуть відповідати різним типам даних. Далі наведемо характеристики типів вмісту та вкажемо основні особливості їх використання [3, 4, 5].

Тип вмісту Discrete вказує на те, що атрибут категоріальний (дискретний). Одна з особливостей даного типу – до цих значень не можна застосувати впорядкування. Крім того, навіть якщо значення, що використовуються для заповнення дискретного стовпця, є числовими, то не передбачена можливість обчислення дробових значень.

Тип Continuous – неперервні числові значення. Інакше кажучи, в стовпці містяться значення, які представляють числові дані в масштабі, що допускає застосування проміжних значень. В якості неперервного значення може розглядатися, наприклад, обсяг продажів за вказаний період або значення середньодобової температури повітря. Якщо відомо розподіл значень неперервного показника, то потенційно можна збільшити точність аналізу, вказавши його під час визначення структури ІАД. Даний параметр буде застосовний до всіх моделей, що базуються на структурі.

Алгоритми ІАД Microsoft можуть працювати з наступними типами розподілів:

– NORMAL – на основі значень стовпця, що містить неперервні дані, може бути побудована гістограма з нормальним Гаусівським розподілом;

– LOG NORMAL – на основі значень стовпця, що містить неперервні дані, може бути побудована гістограма з нормальним розподіленою функцією логарифма значень;

– UNIFORM – рівномірний розподіл значень стовпця (всі значення є рівномірними).

Тип Discretized – вказує на те, що це дискретизоване значення. Дискретизація – це процес розподілу значень неперервного набору даних по сегментах так, щоб вийшла обмежена кількість допустимих значень. Дискретизувати можна тільки числові дані. Отже, дискретизований тип вмісту показує, що стовпець містить значення, що представляють групи або сегменти значень, отриманих з неперервного стовпця. Сегменти сприймаються як впорядковані дискретні значення.

Дискретизацію даних можна провести вручну, щоб отримати необхідні сегменти, або можна використовувати методи дискретизації, що надаються службами SQL Server Analysis Services. У деяких алгоритмах дискретизація виконується автоматично.

Що стосується кількості сегментів, то за замовчуванням система намагається створити 5 сегментів, а якщо даних на 5 сегментів не вистачає, робиться спроба створити меншу кількість. У випадку явного задання параметрів опис дискретизованого стовпця під час створення структури може бути приблизно таким:

```
[SalaryDisc] LONG DISCRETIZED (EQUAL_AREAS,4)
```

Тип вмісту Key (ключ) означає, що стовпець унікально визначає рядок (варіант). Ключ варіанту не використовується для аналізу, а потрібен для відстеження записів.

8.2 Мова DMX: створення структури

Розглянемо конструкції мови DMX, що дозволяють створювати структури. Для цього використовується оператор CREATE MINING STRUCTURE. В узагальненому вигляді його формат поданий нижче:

```
CREATE [SESSION] MINING STRUCTURE <structure>  
(
```

```

[(<column Definition list>)]
)
[WITH HOLDOUT (<holdout-specifier> [OR
<holdout-specifier>])]
[REPEATABLE (<holdout seed>)]
де
<holdout-specifier> :: = <holdout-maxpercent>
PERCENT |
<holdout-maxcases> CASES

```

Необов'язкове ключове слово SESSION показує, що структура є тимчасовою і її можна використовувати лише протягом поточного сеансу роботи з SQL Server. Після завершення сеансу структура і будь-які моделі на її основі видаляються. Щоб створити тимчасові структури і моделі ІАД, необхідно спочатку задати властивість бази даних AllowSessionMiningModels. Треба відзначити, що під час використання для аналізу інструментів Table Analysis Tools з надбудов ІАД для Microsoft Excel створюються саме такі структури.

Для визначення стовпця використовується наступний формат:

```

<column name> <data type> [<Distribution>]
[<Modeling Flags>] <Content Type> [<column
relationship>]

```

де обов'язково вказуються:

<column name> – ім'я стовпця,

<data type> – тип даних,

<Content Type> – тип вмісту.

Необов'язкові параметри, до яких відноситься

<Distribution> – розподіл,

<Modeling Flags> – список прапорів моделювання,

<column relationship> – зв'язок зі стовпцем атрибутів (обов'язкова, тільки якщо застосовна; визначається пропозицією RELATED TO), будуть розглянуті нижче.

Для визначення стовпця з вкладеною таблицею використовується наступний синтаксис:

```

<column name> TABLE (<column definition
list>)

```

В якості прикладу розглянемо створення структури ІАД з ім'ям New Mailing, що включає чотири стовпці: ідентифікатор клієнта

CustomerKey, стать клієнта Gender, число машин Number Cars Owned, ознака покупки велосипеда Bike Buyer. Якщо назви стовпців не включають прогалин, їх можна писати без квадратних дужок (хоча для одноманітності, дужки можна і залишити). У структурі буде створюватися тестувальний набір, що складається з 25% від загального числа входних варіантів. В якості початкового значення під час створення секції використовується число 5000. Після вказівки початкового значення ці ж варіанти будуть вибиратися для тестового набору кожного разу в процесі обробки структури ІАД (за умови, що базові дані не зміняться).

```
CREATE MINING STRUCTURE [New Mailing]
(
CustomerKey LONG KEY,
Ge nder TEXT DISCRETE,
[Number Cars Owned] LONG DISCRETE,
[Bike Buyer] LONG DISCRETE
)
WITH HOLDOUT (25 PERCENT) REPEATABLE (5000)
```

Якщо потрібно використати дані з вкладки таблиці Goods, що містить перелік куплених клієнтом товарів, де будуть стовпці Product (назва товару) і Quantity (кількість), то команда створення структури може виглядати наступним чином.

```
CREATE MINING STRUCTURE [New Mailing1]
(
CustomerKey LONG KEY,
Gender TEXT DISCRETE,
[Number Cars Owned] LONG DISCRETE,
[Bike Buyer] LONG DISCRET E,
[Goods] TABLE
(
[Product] TEXT KEY,
[Quantity] LONG CONTINUOUS
)
)
WITH HOLDOUT (25 PERCENT)
```

Тепер згадаємо про додаткові прапорці. Прапорець Distribution дозволяє вказати на розподіл для стовпця з числовим значенням (це питання розбиралося в попередньому розділі). Допустимі значення – NORMAL, LOG NORMAL, UNIFORM.

До будь-якого визначення стовпця можна додати пропозицію, що описує зв'язок між двома стовпцями. Робиться це за допомогою ключового слова `RELATED TO`, що показує ієрархію значень. Призначенням стовпця `RELATED TO` може бути ключовий стовпець вкладеної таблиці, стовпець з дискретними значеннями з рядка варіантів або який-небудь інший стовпець з пропозицією `RELATED TO`, що вказує на більш глибокий рівень ієрархії.

8.3 Мова DMX: створення моделі

Створення моделі ІАД можна здійснити одним із таких способів:

1) Після створення структури ІАД можна додавати в неї моделі за допомогою інструкції `ALTER MINING STRUCTURE`.

2) Можна використовувати інструкцію `CREATE MINING MODEL`, в результаті виконання якої створюється модель і автоматично формується структура ІАД, що лежить в її основі. Ім'я структури ІАД формується шляхом додавання рядка «_structure» до імені моделі.

Перший спосіб є кращим, особливо коли планується створити на основі однієї структури кілька моделей (що використовують різні набори стовпців, алгоритми тощо). Формат оператора поданий нижче.

```
ALTER MINING STRUCTURE <structure>
ADD MINING MODEL <model>
(
  <column definition list>
  [(<nested Column definition list>) [WITH
FILTER (<nested filter criteria>)]]
)
USING <algorithm> [(<parameter list>)]
[WITH DRILLTHROUGH]
[, FILTER (<filter criteria>)]
```

де `<structure>` – ім'я структури ІАД, до якої буде додана модель;

`<model>` – унікальне ім'я моделі ІАД; `column definition list` – список визначень стовпців з роздільниками-комами;

`<nested column definition list>` – отісок з роздільниками-комами стовпців вкладеної таблиці (якщо застосовно);

`<nested filter criteria>` – визначення фільтра, що застосовується до стовпців вкладеної таблиці;

`<algorithm>` – назва використовуваного моделлю алгоритму ІАД;

`<parameter list>` – отісок параметрів алгоритму (через кому);

`<filter criteria>` – визначення фільтра, що застосовується до стовпців таблиці варіантів.

Якщо структура ІАД містить складові ключі, то модель ІАД повинна включати в себе всі ключові стовпці, визначені в структурі.

Якщо моделі не потрібний прогнозований стовпець (наприклад, під час кластеризації), то в інструкцію не потрібно включати визначення стовпців. Всі атрибути в створюваній моделі будуть розглядатися як входи. Наприклад, наступне визначення моделі MyCluster1 для вирішення задачі кластеризації:

```
ALTER MINING STRUCTURE [New Mailing]
ADD MINING MODEL [MyCluster1]
USING Microsoft_Clustering
```

У загальному випадку, визначення стовпця виконується у відповідності з наступним форматом:

```
<structure column name> [AS <model column
name>] [<modeling flags>] [<prediction>]
```

де `<structure column name>` – ім'я стовпця у відповідності з визначенням структури;

`<model column name>` – псевдонім (необов'язковий параметр, що дозволяє в моделі використовувати назву, відмінну від тієї, що використовується в структурі);

`<modeling flags>` – прапорці моделювання, про які йшлося вище: значення REGRESSOR вказує, що алгоритм регресії може використовувати заданий стовпець у формулі регресії; значення MODEL_EXISTENCE_ONLY вказує, що сама присутність атрибуту важливіша, ніж конкретне його значення;

`<prediction>` – прапорець прогнозування, який вказує, що даний стовпець містить прогнозовану величину. Може приймати значення PREDICT (стовпець є як входом, так і виходом) або PREDICT_ONLY (стовпець є тільки виходом). Стовпець без прапорця вважається входом.

Якщо у визначенні моделі вказано WITH DRILLTHROUGH, то користувачам дозволяється проводити деталізацію (тобто

переглядати не тільки параметри моделі, але й дані варіантів у цій моделі).

Визначення фільтрів дозволяє використовувати в процесі обробки моделі тільки варіанти, що відповідають умовам фільтрації.

Далі наводиться приклад створення моделі ІАД Naive Bayes Women в структурі ІАД New Mailing. Модель використовує спрощений алгоритм Баєса і обмежується тільки тими варіантами зі структури, в яких клієнтами є жінки старше 50 років (значення стовпця Gender, що містить інформацію щодо статі клієнта, дорівнює F, від англ. Female – жінка). Прогнозуватися буде значення атрибута Bike Buyer, тобто чи купить ця клієнтка велосипед чи ні.

```
ALTER MINING G STRUCTURE [New Mailing]
ADD MINING MODEL [Naive Bayes Women]
(
  CustomerKey,
  Gender,
  [Number Cars Owned],
  [Bike Buyer] PREDICT
)
USING Microsoft_Naive_Bayes
WITH FILTER ([Gender] = 'F' AND [Age]> 50)
```

Як уже зазначалося, модель можна створити і за допомогою інструкції CREATE MINING MODEL. Її формат наступний:

```
CREATE [SESSION] MINING MODEL <model>
(
  [(<column Definition list>)]
)
USING <algorithm> [(<parameter list>)]
[WITH DRILLTHROUGH]
```

Всі параметри вже розбиралися вище, тому складнощів в розумінні синтаксису тут бути не повинно.

Крім того, цей оператор можна використовувати для створення моделі, описаної за допомогою заснованої на XML мови розмітки прогнозуючих моделей (Predictive Model Markup Language, PMML). Тоді його формат буде наступним:

```
CREATE MINING MODEL <model> FROM PMML <xml
string>
```

де XML string – опис моделі, закодований мовою PMML. Рядок повинен бути вкладеним в одиначні лапки (''). Цей варіант опису рекомендується тільки для досвідчених користувачів.

Приклад використання першого варіанта запису оператора CREATE MINING MODEL для створення моделі MyAssociationModel наведено нижче. Модель заснована на алгоритмі взаємозв'язків (Association Rules), і в ній використовується вкладена таблиця Products. Значення MINIMUM_PROBABABILITY (мінімальна ймовірність) і MINIMUM_SUPPORT (мінімальна підтримка) є параметрами алгоритму Association Rules (докладніше, див. розділ, присвячений даному алгоритму).

```
CREATE MINING MODEL MyAssociationModel (  
  OrderNumber TEXT KEY,  
  [Products] TABLE PREDICT ([Model] TEXT KEY)  
)  
USING Microsoft_Association_Rules  
  (Minimum_Probability = 0.1, MINIMUM_SUPPORT =  
0.01)
```

Тут хотілося б звернути увагу на одну особливість, яка з'являється під час створення моделі, що використовує тільки ключовий стовпець з вкладеної таблиці. У наведеному прикладі прапорець PREDICT встановлений для вкладеної таблиці Products, а не її ключа Model. Це вказує на те, що модель прогнозує множину рядків, які складають вкладену таблицю для варіанта.

В інших випадках, прапорець може ставитися і для заданого стовпця вкладеної таблиці, який не є вкладеним ключем.

Самостійна робота №9. Початок роботи в BI Dev Studio

Робота із засобами інтелектуального аналізу даних в середовищі BI Dev Studio може виконуватися в одному з двох режимів: immediate або offline.

Режим immediate передбачає роботу в процесі постійного підключення до сервера аналітичних служб. Водночас об'єкти відкриваються безпосередньо з сервера і під час внесення змін, після того як об'єкт збережений, він відразу ж зміниться на сервері. Необхідно враховувати: якщо в цьому режимі модифікується робоча модель інтелектуального аналізу, після збереження вона стане

необробленої, що призведе до збою запитів до цієї моделі. Також можуть виникнути складності, якщо одночасно з вами якийсь об'єкт БД змінював інший користувач. Тоді у разі збереження буде видано попередження, що збереження об'єкта “перезапише” зроблені в ньому зміни. Щоб побачити зроблені іншими зміни в об'єкті, треба його закрити і знову відкрити. Щоб побачити нові об'єкти, треба закрити і заново відкрити весь проект.

Для того, щоб почати роботу в режимі immediate потрібно:

- 1) Запустити BI Dev Studio;
- 2) В меню **File** вибрати **Open -> Analysis Services Database** (рис. 22.1);
- 3) У вікні Connect to Database треба вказати сервер і вибрати існуючу базу або створити нову, а також вказати шлях для розміщення файлів проекту (рис. 8.2).

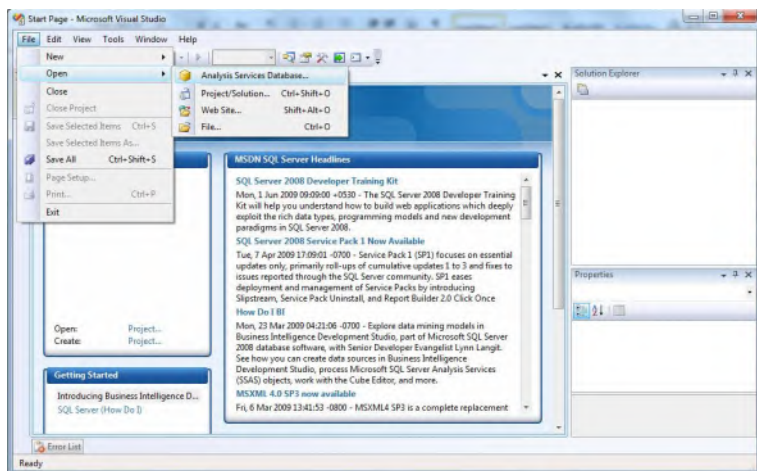


Рис. 8.2 – Діалогове вікно роботи в режимі Immediate

Під час роботи з існуючою базою інтелектуального аналізу в режимі immediate, для створення і модифікації в ній об'єктів треба володіти правами адміністратора цієї бази даних (рис. 8.3).

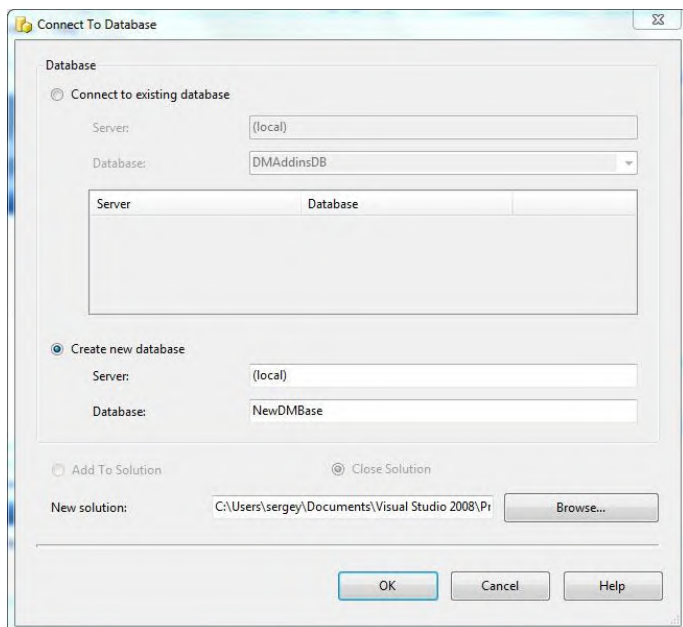


Рис. 8.3 – Діалогове вікно для вибору існуючої бази даних та створення нової

Особливість роботи в режимі offline полягає в тому, що створювані під час роботи над проектом об'єкти зберігаються на вашому комп'ютері у вигляді XML-файлів. І на сервері об'єкти будуть створені тільки після того, як ви “розгорнете” проект на сервер.

Одиницею розгортання є весь проект, який представляє собою базу даних аналітичних служб. У міру роботи над проектом можливо інкрементальное розгортання змін (на сервері створюються або змінюються тільки нові і змінені об'єкти БД). Водночас для розгортання проекту, треба мати права адміністратора примірника SQL Server, на який проводиться розгортання. Якщо на сервері вже є база даних з тим же ім'ям, що і в налаштуваннях вашого проекту, в процесі розгортання існуюча база буде видалена.

Цей режим може бути більш зручний, якщо потрібно створити проект інтелектуального аналізу даних, випробувати роботу створюваних об'єктів на тестовому сервері, а потім розгорнути рішення у виробничому середовищі.

Для того, щоб почати роботу в режимі offline потрібно:

- 1) Запустити BI Dev Studio.

2) В меню **File** вибрати **New → Project**.

3) У вікні (рис. 8.4) вибрати тип проекту (Project types: Business Intelligence Projects) і шаблон (Templates: Analysis Services Project), вказати ім'я проекту і місце розташування його файлів і натиснути кнопку **OK**.

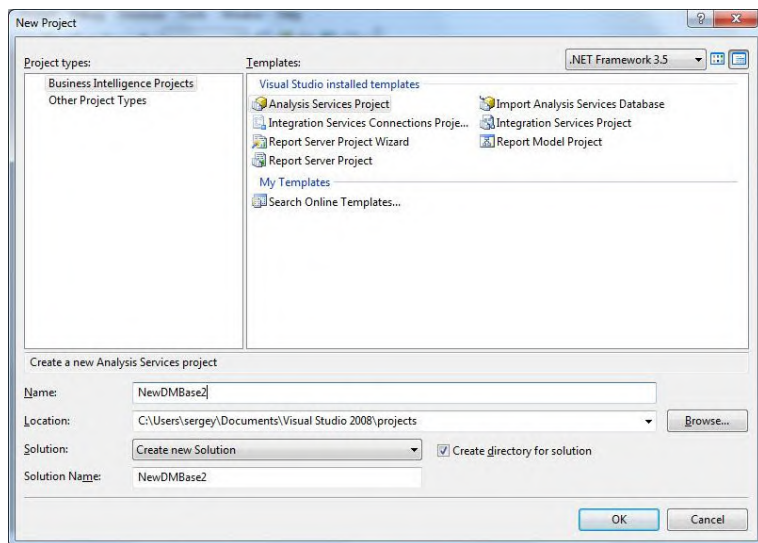


Рис. 8.4 – Діалогове вікно для створення нового проекту

Щоб розгорнути проект на SQL Server 2012, треба в меню **Build** вибрати пункт **Deploy** (рис. 8.5). За замовчуванням буде використовуватися неіменованого екземпляр SQL Server 2012, встановлений на тому ж комп'ютері (ім'я сервера буде localhost).

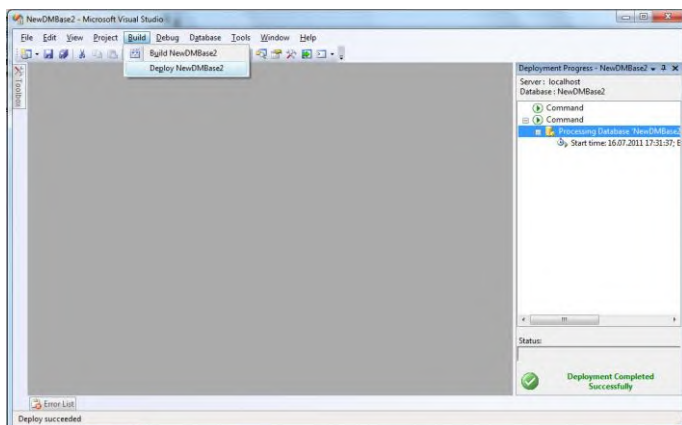


Рис. 8.5 – Вікно розгортання проекту на сервер

Налаштування використовуваного сервера аналітичних служб можна змінити через меню **Project -> <ім'я проекту> Properties ...** (рис. 8.6).

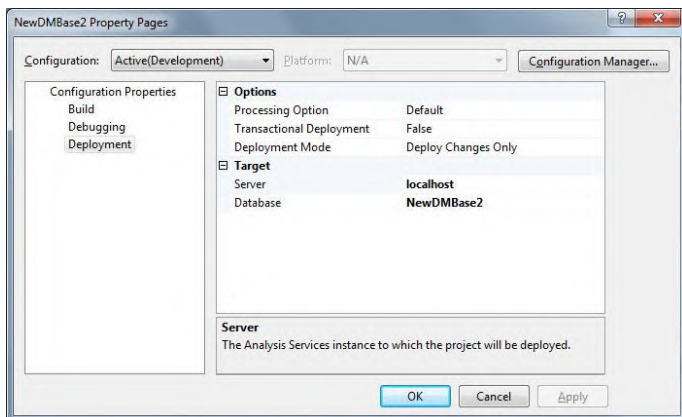


Рис. 8.6 – Вікно налаштування параметрів розгортання проекту

У розділі **ConfigurationProperties -> Deployment** можна вказати цільової сервер для обраної конфігурації (вузол Target настройка Server). А скориставшись кнопкою **Configuration Manager**, можна створити додаткові конфігурації, для яких вказати свої настройки сервера призначення і бази даних.

Якщо проект був створений в режимі Offline і розгорнуть на сервер, а далі з ним вирішено працювати в режимі Immediate, треба, відкривши BI Dev Studio, в меню **File** вибрати **Open -> AnalysisServicesDatabase** і вказати створену базу даних.

Завдання 1. Створіть проект інтелектуального аналізу даних (в режимі offline). Розгорніть його на сервер. Закрийте BI Dev Studio. Знову відкрийте середу. Почніть роботу з БД інтелектуального аналізу в режимі immediate.

Завдання 2. Для подальшого виконання самостійної роботи нам знадобляться дані з навчальної бази **AdventureWorksDW** для SQL Server 2012 (або SQL Server 2012 R2 – в залежності від встановленої версії). Якщо ця база даних у вас не встановлена, потрібно завантажати програму-інстальатор за посиланням <http://Microsoftftdbprodsamples.codeplex.com/>

Створення джерела даних

З BI Dev Studio в режимі immediate підключіться до створеної під час виконання Завдання 1 бази даних інтелектуального аналізу. У разі необхідності, відкрийте вікно **SolutionExplorer** (меню **View -> SolutionExplorer** або **Ctrl + Alt + L**). Виділіть вузол **DataSources** і в контекстному меню виберіть **NewDataSource** (рис. 8.7). Ми збираємося створити нове джерело даних (**DataSource**).

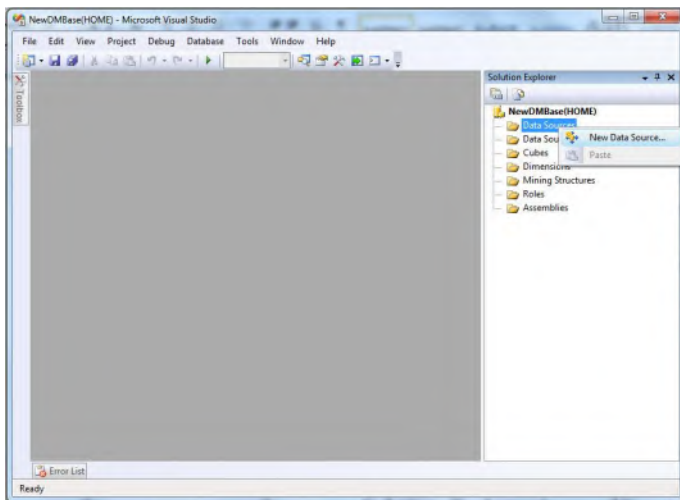


Рис. 8.7 – Вікно створення нового джерела даних

Виконані дії призводять до запуску майстра Data Source Wizard. Після вікна вітання, буде запропоновано визначити, яке підключення (рис. 8.8).

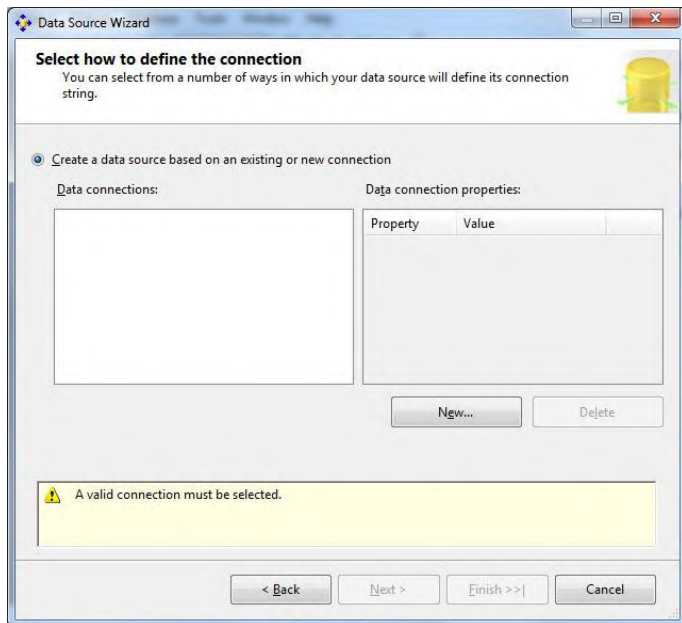


Рис. 8.8 – Діалогове вікно вибору підключення до бази даних

Якщо потрібного підключення немає, його треба створити, натиснувши кнопку **New**. Далі вказати провайдер даних (за замовчуванням обраний провайдер для Microsoft SQL Server 2012, що нам підходить), сервер, тип аутентифікації і базу даних, які будемо використовувати (рис. 8.9). Нам знадобиться база **Adventure Works DW**. Коли підключення створено і протестовано (кнопка **Test**), воно з'явиться в списку доступних.

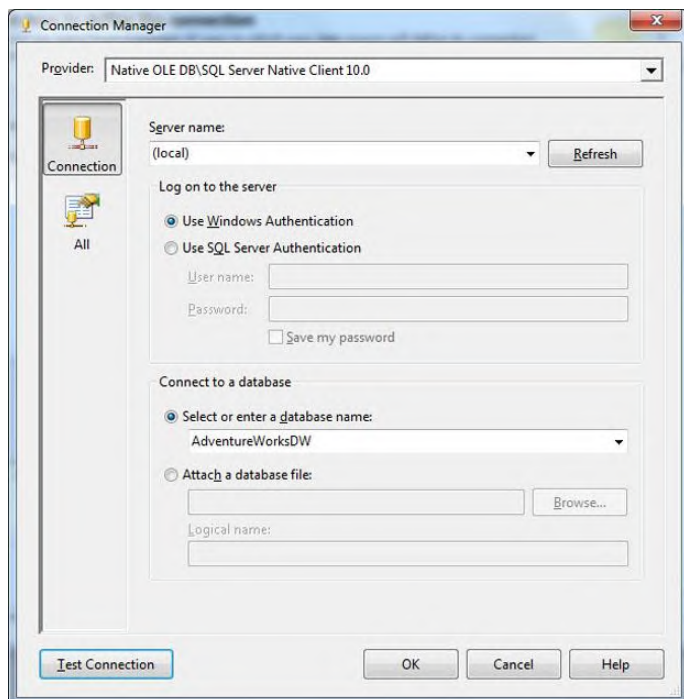


Рис. 8.9 – Діалогове вікно створення нового підключення

Далі потрібно вказати обліковий запис, від імені якої аналітичні служби будуть звертатися до зазначеної бази даних (рис. 8.10). Можна вказати окрему обліковий запис або, наприклад, вибрати використання облікового запису поточного користувача (якщо він має права на доступ до бази Adventure Works DW). До цього питання ми ще повернемося в “Створення структури і моделі інтелектуального аналізу. Завдання кластеризації”.

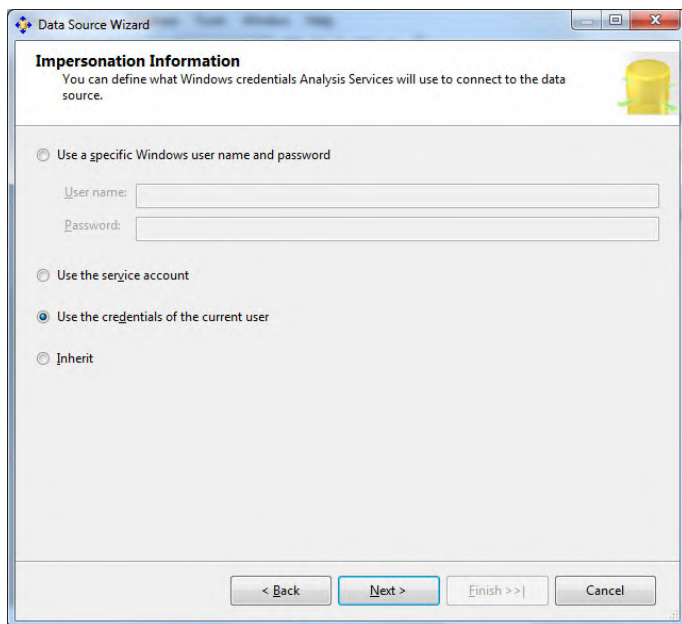


Рис. 8.10 – Діалогове вікно з вказівкою облікового запису для доступу до даних

В останньому вікні буде відображена сформована майстром рядок, що описує джерело даних.

Завдання 3. Створіть джерело даних, який вказує на реляційну базу Adventure Works DW.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Які типи розподілів для неперервного показника використовують алгоритми інтелектуального аналізу Microsoft?
2. Які дані описує тип Discretized? Наведіть приклади.
3. Наведіть способи розбиття (параметри дискретизації) для типу вмісту Discretized.
4. Що означає тип вмісту Key (ключ)?
5. Що означає тип вмісту Cyclical?
6. Приведіть та опишіть узагальнений формат конструкції мови DMX для створення структури інтелектуального аналізу.
7. Наведіть способи створення моделі інтелектуального аналізу

8. Що є наступним кроком після визначення структур і моделей?

9. За допомогою якої інструкції можна видалити модель або структуру інтелектуального аналізу даних із бази даних?

10. За допомогою якої інструкції у DMX проводиться робота з даними та побудова прогнозів?

ЛІТЕРАТУРА

1. DMX Tutorials (Analysis Services – Data Mining) [Електронний ресурс] // SQL Docs. – 2018. – Режим доступу до ресурсу: <https://docs.microsoft.com/en-us/sql/dmx/dmx-tutorials-analysis-services-data-mining?view=sql-server-2017>.

2. Lohr S. Data-Ism: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else / S. Lohr. – HarperBusiness, 2015. – 489 p.

3. Службы SQL Server Analysis Services – интеллектуальный анализ данных. – [Електронний ресурс]. – Режим доступу: <https://docs.microsoft.com/ru-ru/sql/analysis-services/analysis-services-tutorials-ssas?view=sql-server-2017>

4. Службы Analysis Services – видеоролики по интеллектуальному анализу данных (SQL Server 2012). URL: [http://Microsoftdn.microsoft.com/ru-ru/library/dd776389\(v=SQL.100\).aspx](http://Microsoftdn.microsoft.com/ru-ru/library/dd776389(v=SQL.100).aspx)

5. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.

РОЗДІЛ 9.

ЗАПИТИ МОВИ DATA MINING EXTENSIONS

9.1 Мова DMX: обробка, очищення, видалення та відновлення структур і моделей

Наступним кроком після визначення структур і моделей, є їх обробка, що включає заповнення структури ІАД, секціонування даних (якщо це було визначено під час створення структури), застосування щодо отриманих даних алгоритмів інтелектуального аналізу (в процесі обробки моделі) [1]. Це робиться за допомогою інструкції INSERT INTO, формат якої наведено нижче:

```
INSERT INTO [MINING MODEL] | [MINING  
STRUCTURE] <model> | <structure> (<mapped model  
columns>) <source data query>
```

або

```
INSERT INTO [MINING MODEL] | [MINING  
STRUCTURE] <model> | <structure>. COLUMN_VALUES  
(<mapped model columns>) <source data que ry>
```

де <model> – назва моделі;

<structure> – назва структури;

<mapped model columns> – список через кому з назвами стовпців, в тому числі і вкладених таблиць з їх стовпцями;

<source data query> – запит, що описує завантажуваний набір вхідних даних.

Якщо в операторі вказана структура ІАД, то обробляється ця структура і всі пов'язані з нею моделі. Якщо задана модель, інструкція обробляє тільки цю модель. У випадку, коли не вказаний аргумент MINING MODEL або MINING STRUCTURE, служби Analysis Services проводять пошук типу об'єкта на основі імені, і потім обробляють коректний об'єкт. Якщо сервер містить структуру і модель ІАД з однаковими іменами, повертається помилка.

Форма INSERT INTO <об'єкт>. COLUMN_VALUES, дозволяє робити вставку даних безпосередньо в стовпці моделі без її навчання. Під час використання цього методу, дані стовпців поставляються моделі в стислому і впорядкованому вигляді, що корисно під час роботи з наборами даних, що містять ієрархії або упорядковані стовпці.

Елементи списку `<mapped model columns>` можуть бути подані у вигляді:

```
<column identifier> | SKIP | <table  
identifier> (<column identifier> | SKIP)
```

де `<column identifier>` – назва стовпця;

`<table identifier>` – назва вкладеної таблиці;

У нашому курсі ми найчастіше будемо використовувати перший варіант у випадку відсутності в структурі вкладених таблиць і третій – за їхньої наявності. Розглянемо ці формати більш докладно, почавши з першого.

```
OPENQUERY (<named datasource>, <query  
syntax>)
```

де `<named datasource>` – джерело даних, існуюче в базі служб Analysis Services (наприклад, визначений за допомогою конструктора в BI Dev Studio);

`<query syntax>` – текст запиту, що повертає набір рядків (наприклад, запиту на SQL).

У наступному прикладі інструкція OPENQUERY застосовується для навчання моделі спрощеного алгоритму Баєса на основі даних про цільову розсилку з бази даних AdventureWorksDW (те, що це модель, а не структура, SQL Server визначить в результаті аналізу доступних моделей і структур).

```
INSERT INTO NBSample (CustomerKey, Gender,  
[Number Cars Owned], [Bike Buyer])  
OPENQUERY ([Adventure Works DW], 'Select  
CustomerKey, Gender, [NumberCarsOwned], [BikeBuyer]  
FROM [vTargetMail]')
```

Інший приклад [20] показує варіант запису з явною вказівкою, що обробляється структура People1 зі стовпчиками CustID, Name тощо. Для її заповнення використовується джерело даних Chapter3Data, в апострофи вкладений текст SQL-запиту, виконання якого призведе до отримання необхідного набору рядків.

```
INSERT INTO MINING STRUCTURE [People1]  
([CustID], [Name], [Gender], [Age],  
[CarMake], [CarModel])
```

```
OPENQUERY (Chapter3Data,
'SELECT [Key], Name, Gender, Age, CarMake,
CarModel FROM People')
```

Інструкція SHAPE має більш складний у порівнянні з OPENQUERY формат:

```
SHAPE {<master query>}
APPEND ({<child table query>}
RELATE <master column> TO <child column>)
AS <column table name>
[
({<child Table query>}
RELATE <master column> TO <child column>)
AS <column table name>
]
```

де <master query> – запит, який повертає батьківську таблицю (таблицю варіантів);

<child table query> – запит, який повертає вкладену таблицю;

<master column> – стовпець в батьківській таблиці, з яким пов'язані рядки з результату запиту дочірньої таблиці (часто це потенційний ключ, на який посилається зовнішній ключ з вкладеної таблиці);

<child column> – стовпець у дочірньої таблиці для визначення батьківського рядка з результату головного запиту (як правило-зовнішній ключ у таблиці, використовуваної в якості вкладеної);

<column table name> – ім'я стовпця вкладеної таблиці в батьківській таблиці.

Інструкція SHAPE вимагає, щоб результати запитів були впорядковані по стовпцях, що використовуються для зв'язку між батьківською і вкладеними таблицями. Для цього в SQL використовується інструкція ORDER BY.

У наступному прикладі структури ІАД Targeted Mailing і Forecasting, а також модель ІАД Association експортуються в певний файл. Нехай модель Association є частиною структури інтелектуального аналізу Market Basket. Тоді також екпортується структура Market Basket. Будь-які інші моделі ІАД, які можуть існувати як частина структури Market Basket, не будуть експортуватися, тому що модель Association була експортована за допомогою інструкції MINING MODEL, а не MINING STRUCTURE.

```
EXPORT MINING STRUCTURE [Targeted Mailing],  
[Forecasting] MINING MODEL Association TO 'C: \\  
TM_NEW.abf'
```

У наступному прикладі модель ІАД Association експортується в певний файл. Внаслідок того, що в інструкції вказується WITH DEPENDENCIES, об'єкти джерела даних і представлення джерела даних також включаються до ABF-файла.

```
EXPORT MINING MODEL [Association] TO 'C: \\  
Association_NEW.abf' WITH DEPENDENCIES
```

У наступному прикладі виконується імпорт всього вмісту файлу з моделлю Association, структурою та додатковими об'єктами на поточний сервер.

```
IMPORT FROM 'C: \ Association_NEW.abf'
```

Також можна захищати експортовану копію паролем:

```
EXPORT MINING STRUCTURE People2 TO 'C: \ temp  
\ People2.abf' WITH PASSWORD = 'People2'
```

Той же пароль знадобиться під час імпорту:

```
IMPORT FROM 'C: \ temp \ People2.abf' WITH  
PASSWORD = 'People2'
```

9.2 Мова DMX: запити

У попередньому розділі було розглянуто порядок обробки структур і моделей ІАД. Наступне завдання – це робота з даними та побудова прогнозів. Для цього використовується оператор SELECT.

SELECT мовою DMX нагадує аналогічний оператор в SQL, але треба розуміти, що і перелік вирішуваних завдань, та особливості застосування мають суттєві відмінності.

Для вирішення перших двох завдань використовується наступний формат запису оператора:

```
SELECT [FLATTENED] [TOP <n>] <select list>  
FROM <model/structure> [. Aspect]  
[WHERE <condition expression>]
```

```
[ORDER BY <expression> [DESC | ASC]]
```

Інструкція **FLATTENED** вказує на необхідність перетворення результатів, що повертаються запитом **SELECT** в «плоский» набір рядків (тобто перетворення до звичайної таблиці). Вона виконується, коли подання варіантів з вкладеними таблицями у використовуваному за умовчанням ієрархічному форматі неприйнятно.

Інструкції **ORDER BY** і **TOP <n>** дозволяють упорядкувати набір, що повертається за вказаним параметром і повернути тільки перші *n* значень. Це може бути корисно в сценаріях аналогічних цільовим розсилкам, де результати потрібно відправляти тільки найбільш імовірним одержувачам (подібний приклад розглядається в самостійних роботах). Для цього можна впорядкувати результати прогнозного запиту для цільової розсилки по ймовірності, а потім повернути тільки верхні *n* результатів.

У списку вибору допустимо використовувати псевдоніми і прості вирази. Наприклад, нижче для всіх присутніх в моделі **MyModel** варіантів (на це вказує ключове слово **CASES**, див. далі) виводиться ідентифікатор клієнта і його прізвище та ім'я у вигляді одного стовпця **FullName**:

```
SELECT [CustomerID], [Last Name] + ', ' +  
[FirstName]  
AS FullName  
FROM MyModel. CASES
```

Інструкція **WHERE** дозволяє обмежити перелік результатів, що повертаються лише тими, що відповідають зазначеній логічній умові. Синтаксис буде наступний:

```
WHERE <condition expression>
```

Наприклад:

```
SELECT [CustomerID], [Last Name] + ', ' +  
[FirstName]  
AS FullName  
FROM MyModel.CASES  
WHERE [CustomerID]> 100
```

Деталізація структури

Розглянемо тепер інструкцію FROM. Якщо в ній стоїть <structure>.CASES, де <structure> – ім'я структури інтелектуального аналізу, то будуть повертатися варіанти, використані для створення структури. Якщо деталізація для структури не включена, виконання даної інструкції завершиться збоєм. Але за умовчанням деталізація включена. Явно вказати, що ми хочемо працювати зі структурою, можна за допомогою ключових слів MINING STRUCTURE. Наприклад, вибрати всі варіанти зі структури MyStr1 з перетворенням результату в плоский набір рядків можна виконавши команду:

```
SELECT  FLATTENED * FROM MINING STRUCTURE
MyStr1.CASES
```

Тепер згадаємо про те, що під час визначення структури могло бути задане секціонування варіантів на навчальний і тестувальний набір. Щоб дізнатися, до якого набору відноситься варіант, використовуються функції IsTestCase() і IsTrainingCase(), що повертають істину для «перевірочного» і «навчального» варіанта відповідно.

Нижче наведено приклад отримання списку стовпців Customer Key («ключ клієнта»), Gender («стать»), Age («вік») для варіантів, зарезервованих у тестувальному наборі структури Targeted Mailing («цільова розсилка»):

```
SELECT  [Customer Key], Gender, Age FROM
[Targeted Mailing]. Cases WHERE IsTestCase ()
```

Деталізація моделі

Аналогічним чином можуть бути отримані варіанти з моделі (проведена деталізація моделі). Формат оператора виглядає наступним чином:

```
SELECT  [FLATTENED] [TOP <n>] <expression
list>
FROM <model>. CASES
[WHERE <condition expression>]
[ORDER BY <expression> [DESC | ASC]]
```

Якщо деталізація для моделі ІАД не включена, виконання даної інструкції завершиться помилкою. Для розширень ІАД активувати деталізацію можна тільки під час створення моделі (за допомогою інструкції WITH DRILLTHROUGH). У середовищі BI Dev

Studio можна додати деталізацію і в існуючу модель, але перш ніж можна буде переглядати варіанти, необхідно виконати повторну обробку моделі.

Якщо деталізація включена як для моделі, так і для структури ІАД, користувачі, які є членами ролі, що володіє дозволом на деталізацію моделі та структури, можуть звертатися до стовпців в структурі ІАД, які не включені в модель.

Під час використання функції IsInNode в пропозиції WHERE видаються тільки варіанти, пов'язані з зазначеним вузлом (номер вузла задається в стовпці NODE_UNIQUE_NAME, про нього див. далі).

Нижче наведено приклад, який повертає всі стовпці для всіх варіантів, зарезервованих у тестувальному наборі моделі «Цільова розсилка» (TM Decision Tree).

```
SELECT * FROM [TM Decision Tree]. Cases WHERE  
IsTestCase ()
```

У наступному прикладі повертаються тільки ті варіанти з навчального набору, які використовувалися в кластері 2 моделі TM_Clustering. Вузол кластера 2 має значення «002» в стовпці NODE_UNIQUE_NAME, і він вказується в якості аргументу функції IsInNode. Крім всіх стовпців моделі, повертається і стовпець структури Customer Key, який не входив до моделі ІАД. Цьому стовпцю присвоюється псевдонім CustomerID. Робиться це за допомогою функції StructureColumn. Ім'я стовпця структури передається як строкове значення, тому його слід брати в лапки.

```
SELECT StructureColumn ('Customer Key') AS  
CustomerID,  
*  
FROM [TM_Clustering]. Cases  
WHERE IsTrainingCase () AND IsInNode ('002 ')
```

Запит значень стовпця

Нижче наведено синтаксис оператора, що дозволяє отримати значення зазначених стовпців моделі.

```
SELECT [FLATTENED] DISTINCT [TOP <n>]  
<expression list>  
FROM <model>  
[WHERE <condition list>] [ORDER BY  
<expression>]
```

Водночас, для дискретного стовпця будуть виведені всі можливі значення, для безперервного – середнє значення, для дискретизованного – середнє значення для кожного із визначених у процесі дискретизації інтервалів.

Розглянемо два приклади. У першому будуть виведені всі значення дискретного стовпця Gender (Стать) моделі TM Decision Tree (результат виконання запиту представлений в табл. 9.1).

```
SELECT DISTINCT [Gender]
FROM [TM Decision Tree]
```

Таблиця 9.1 – Результат виконання запиту для дискретного стовпця

Gender
Дружина
Чоловік

У другому прикладі для стовпця Age (вік), визначеного як неперервний, буде виведено середнє значення, а також за допомогою функцій RangeMin і RangeMax, відповідно, мінімальний і максимальний вік (табл. 9.2).

```
SELECT DISTINCT [Age] AS [Midpoint Age],
RangeMin ([Age]) AS [Minimum Age],
RangeMax ([Age]) AS [Maximum Age]
FROM [TM Decision Tree]
```

Таблиця 9.2 – Результат виконання запиту для неперервного стовпця

Midpoint Age	Minimum Age	Maximum Age
62	26	97

Запит вмісту моделі

Оператор SELECT може не тільки повернути використовувані моделлю варіанти, але і вивести її вміст. Для цього використовується синтаксис:

```
SELECT [FLATTENED] [TOP <n>] <expression list>
FROM <model>. CONTENT
```

```
[WHERE <condition expression>]
[ORDER BY <expression> [DESC | ASC]]
```

Аргументи команди були вже розглянуті вище, зараз важливо зрозуміти, що ж поверне запит.

9.3 Мова DMX: запити для прогнозу

Зараз ми підійшли до одного з найцікавіших етапів інтелектуального аналізу – побудові прогнозів. Як вже зазначалося в попередніх розділах, завдання прогнозування мовою DMX також вирішується за допомогою оператора SELECT. Водночас найчастіше використовується конструкція прогнозного з'єднання – PREDICTION JOIN. З її допомогою шаблонам моделі зіставляється набір даних із зовнішнього джерела, що дозволяє визначити значення для прогнозованого стовпця.

Спрощено це можна подати таким чином. Модель містить деяку множину шаблонів, що описують залежність початкових параметрів від вхідних. Подібний приклад наведено нижче.

1. Якщо Вхід1 = x, то Вихід = «так».
2. Якщо Вхід1 = y і Вхід2 = z, то Вихід = «ні».
3. ...

Під час прогнозування модель отримує вхідні дані із зовнішнього джерела, визначає найбільш відповідний шаблон і видає результат. Наприклад, вхідний варіант має атрибут «Вхід1 = x», застосовуючи перше правило отримуємо «Вихід = «так»».

Розглянемо тепер синтаксис оператора SELECT, що використовується для прогнозного з'єднання.

```
SELECT      [FLATTENED]      [TOP    <n>]    <select
expression list>
FROM        <model>      |    <sub select>    [NATURAL]
PREDICTION JOIN <source data query> [ON <join
mapping list>]
[WHERE <condition expression>]
[ORDER BY <expression> [DESC | ASC]]
```

де n – ціле число, яке вказує кількість рядків, що повертається;
 <select expression list> – розділений комами список стовпців і/або виразів;

<model> – назва моделі;

<sub select> – впроваджена (вкладена) інструкція select;

<source data query> – «вихідний» запит, що описує набір даних, для якого робиться прогнозне з'єднання. він може бути таблицею або одноелементним запитом;

<join mapping list> – логічний вираз, що включає порівняння стовпців моделі зі стовпчиками, повернутими запитом;

<source data query> – визначає умови з'єднання;

<condition expression> – логічний вираз, що обмежує множину значень, що повертаються, тільки тими, що відповідають умові; умову where можна застосовувати тільки до прогнозованих стовпців або до пов'язаних стовпців;

<expression> – вираз, що визначає упорядкування результату який повертається. Order by може прийняти в якості аргументу тільки один стовпець, сортувати по декількох стовпцях – не можна.

Пропозиція ON дозволяє зіставити стовпці початкового запиту зі стовпчиками моделі ІАД (щоб визначити, які значення розглядати в якості вхідних під час прогнозування). Стовпці в списку join mapping list співвідносяться з допомогою знаку рівності, як показано в наступному прикладі:

```
[MiningModel]. ColumnA = [source data query].  
Column1 AND  
[MiningModel]. ColumnB = [source data query].  
Column2 AND ...
```

Інструкція NATURAL PREDICTION JOIN автоматично зіставляє імена стовпців початкового запиту, співпадаючих з іменами стовпців моделі. У разі використання інструкції NATURAL PREDICTION пропозицію ON можна пропустити.

У списку SELECT EXPRESSION LIST і виразі CONDITION EXPRESSION можна вказувати прогнозуючі функції і не повертають табличного виразу.

Розглянемо тепер кілька прикладів побудови прогнозів.

Одноелементний запит (прогноз в реальному часі)

Одноелементним будемо називати запит, в якому вхідні дані для прогнозування не вибираються з джерела, а явно вказані в тексті запиту. Такий запит може формуватися «нальоту», наприклад, на основі введених користувачем даних, і тут же видавати результати прогнозу.

Нижче наведено приклад подібного прогнозного запиту до моделі TM Decision Tree для оцінки того, чи стане покупцем

велосипеда людина у віці 35 років, яка проживає в 5-10 милях від місця роботи, яка володіє одним будинком, двома автомобілями, з двома дітьми в сім'ї (значення перераховані в тому порядку, як йдуть в лістингу). Запит повертає прогнозоване логічне значення стовпця Bike Buyer і набір табличних значень, повернутих функцією PredictHistogram, що описують, як був зроблений прогноз.

```
SELECT
    [TM Decision Tree].      [Bike Buyer],
PredictHistogram ([Bike Buyer])
FROM
    [TM Decision Tree]
NATURAL PREDICTION JOIN (SELECT 35 AS [Age],
    '5 -10 Miles 'AS [Commute Distance],
    '1 'AS [House Owner Flag],
    2 AS [Number Cars Owned],
    2 AS [Total Children]) AS t
```

Тут потрібно відзначити, що за рахунок використання конструкції NATURAL PREDICTION JOIN визначати відповідності між стовпцями моделі і набору вхідних даних не потрібно.

Використання функції Predict

Функція Predict повертає прогнозоване значення або набір значень для заданого стовпця. Вона є поліморфною, і її поведінка залежить від того, чи підставляється в якості аргументу ім'я звичайного стовпця або вкладеної таблиці. Синтаксис буде наступним:

```
Predict      (<scalar      column      reference>,
[option1],
    [Option2], [option n], [INCLUDE_NODE_ID], n)
або
Predict (<table column reference>, [option1],
    [Option2], [option n], [INCLUDE_NODE_ID], n)
```

У той же час, просто вибір прогнозованого стовпця в заголовку SELECT, аналогічний виклику функції PREDICT з параметрами за замовчуванням. Наприклад, якщо під час створення моделі зазначено що стовець Gender (стать) – прогнозований, то SELECT [Gender] буде аналогічно Predict ([Gender], EXCLUDE_NULL). Де опція EXCLUDE_NULL вказує на те, що порожні значення виключаються.

Більш докладно з можливостями функції PREDICT можна ознайомитися в довіднику по функціях розширень ІАД.

Прогнозування найбільш вірогідного значення

Якщо використовувати наступний синтаксис оператора SELECT:

```
SELECT <expression list> [TOP <n>] FROM
<model>
[WHERE <condition list>]
[ORDER BY <expression> [DESC | ASC]]
```

і в списку expression list вказати тільки прогнозовані стовпці, то оператор виконає порожнє прогнозне з'єднання і поверне найбільш імовірні значення для зазначених стовпців. Для створення прогнозу використовується тільки вміст моделі ІАД. Наприклад:

```
SELECT ([Bike Buyer]) FROM [TM_Naive_Bayes]
```

видасть найбільш часто зустрічається значення стовпця Bike Buyer.

Створення копії моделі – оператор SELECT INTO

На закінчення огляду можливостей оператора SELECT, розглянемо створення копії існуючої моделі ІАД. Для цього використовується синтаксис:

```
SELECT INTO <new model>
USING <algorithm> [(<parameter list>)]
[WITH DRILLTHROUGH [,] [FILTER
(<expression>)]]
FROM <existing model>
```

де <new model> – ім'я для нової створюваної моделі;

<algorithm> – назва використовуваного новою моделлю алгоритму ІАД;

<parameter list> – список через кому параметрів алгоритму;

<expression> – вираз, що визначає фільтр для значень, що попадають в нову модель;

<existing model> – ім'я існуючої моделі для копіювання.

Якщо відповідна модель є навченою, нова модель автоматично обробляється в процесі виконання цієї інструкції. В іншому випадку, нова модель залишається необробленою.

Наприклад, наведений нижче фрагмент коду створює нову модель New_Clustering на базі існуючої моделі ІАД TM Clustering. В обох випадках використовується алгоритм кластеризації. Але в новій моделі параметр CLUSTER_COUNT змінюється так, щоб існувало максимум п'ять кластерів, тоді як значення за замовчуванням дорівнює 10.

```
SELECT * INTO [New_Clustering]
USING [Microsoft_Clustering] (CLUSTER_COUNT =
5)
FROM [TM Clustering]
```

Самостійна робота №10. Створення уявлення джерела даних

У середовищі BI Dev Studio в режимі immediate відкрийте створену базу аналітичних служб (в меню **File** вибрати **Open -> AnalysisServicesDatabase**, далі вказати використовуваний сервер і вибрати створену під час виконання попередньої самостійної роботи базу). У вікні SolutionExplorer переконайтеся, перевірте, що в базі є джерело даних, що посилається на реляційну базу даних AdventureWorksDW (рис. 9.1, вузол DataSources).

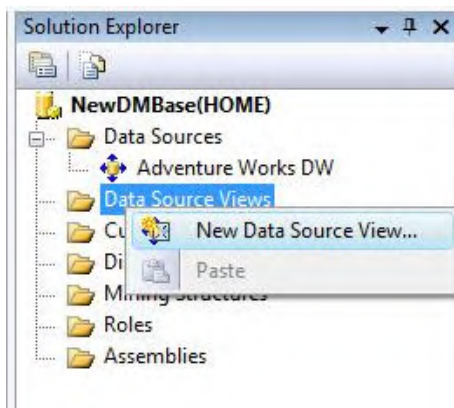


Рис. 9.1 – Діалогове вікно створення джерела даних

Подання джерела даних (DataSourceView, DSV) визначається як абстрактне уявлення, яке дозволяє модифікувати спосіб розгляду джерела даних (DataSource), або описати схему і надалі змінювати

фактичний джерело даних. Останнє може знадобитися, якщо в процесі розробки і тестування використовуються дані з однієї реляційної БД, а в подальшому використовуватимуться інші бази аналогічної структури.

Для створення DSV треба вибрати відповідний вузол у вікні Solution Explorer, клацанням правої клавіші викликати контекстне меню і вибрати пункт “NewDataSourceView”.

У реляційній базі даних AdventureWorksDW є уявлення dbo.vTargetMail, що дозволяє отримати інформацію про клієнта (ідентифікатори, ім'я, прізвище, регіон, розмір річного доходу та ін.) і про те, купив даний клієнт велосипед чи ні. Ці дані нам знадобляться для подальшого аналізу.

Запустивши майстер створення DSV, в першому вікні вкажемо використовуваний джерело даних – AdventureWorksDW (рис. 9.2).

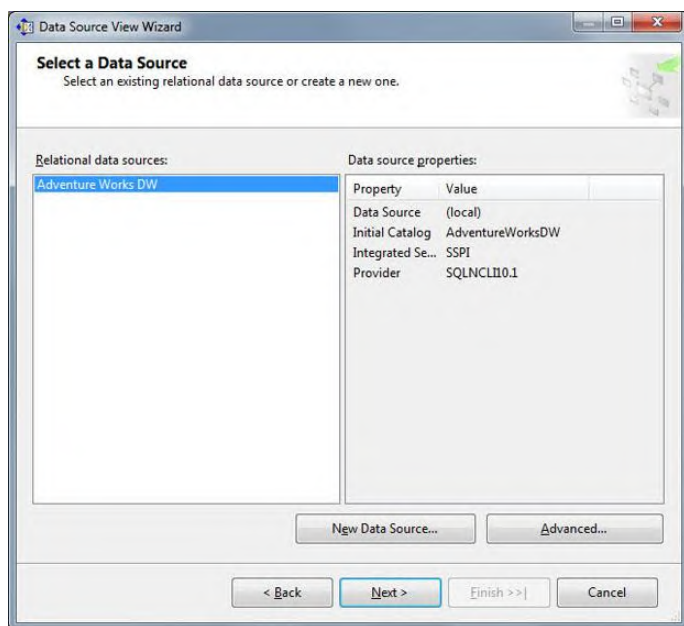


Рис. 9.2 – Діалгове вікно створення джерела даних: вибір джерела даних

У зв'язку з тим, що обраний джерело даних вказує на реляційну БД, наступне вікно буде містити перелік таблиць і уявлень,

які можна включити в уявлення джерела даних, тут треба вибрати уявлення vTargetMail і таблицю DimGeography (рис. 9.3).

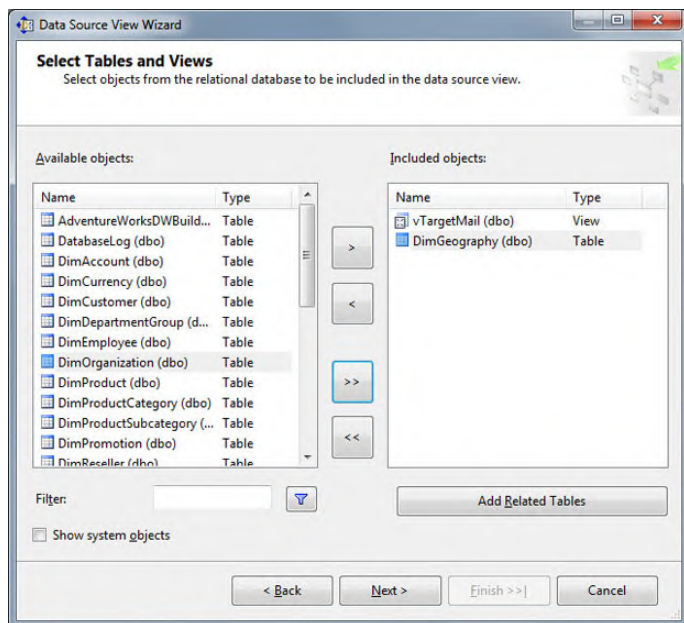


Рис. 9.3 – Діалогове вікно створення джерела даних: вибір об'єктів, що включаються в уявлення

Якщо у вихідній БД відкрити в конструкторі уявлення vTargetMail>, ми побачимо наступний код мовою SQL:

```
SELECT      c.CustomerKey,          c.GeographyKey,
c.CustomerAlternateKey, c.Title,
           c.FirstName,          c.MiddleName,          c.LastName,
c.NameStyle, c.BirthDate,
           c.MaritalStatus,          c.Suffix,          c.Gender,
c.EmailAddress, c.YearlyIncome,
           c.TotalChildren,          c.NumberChildrenAtHome,
c.EnglishEducation,
           c.SpanishEducation,          c.FrenchEducation,
c.EnglishOccupation,
           c.SpanishOccupation,          c.FrenchOccupation,
c.HouseOwnerFlag, c.NumberCarsOwned,
```

```

        c.AddressLine1,    c.AddressLine2,    c.Phone,
c.DateFirstPurchase,
        c.CommuteDistance, x.Region, x.Age, CASE x.
[Bikes] WHEN 0 THEN 0 ELSE 1 END AS BikeBuyer
        FROM dbo.DimCustomer AS c INNER JOIN
        (SELECT CustomerKey, Region, Age, SUM (CASE
[EnglishProductCategoryName] WHEN 'Bikes' THEN 1
ELSE 0 END) AS Bikes
        FROM dbo.vDATA MININGPrep
        GROUP BY CustomerKey, Region, Age) AS x ON
c.CustomerKey = x.CustomerKey

```

Звідси видно, що частина інформації про клієнта береться з таблиці `dbo.DimCustomer`, зокрема це код регіону проживання клієнта (`GeographyKey`). На рисунку 9.4 наведено фрагмент діаграми, що показує зв'язок таблиці `DimCustomer` і таблиці-довідника `DimGeography`, а рисунку 9.5 демонструє вибірку даних з уявлення `vTargetMail` і таблиці `DimGeography`. Таким чином, якщо ми хочемо знати назви країн і міст перебування клієнтів, нам знадобиться таблиця `DimGeography`.

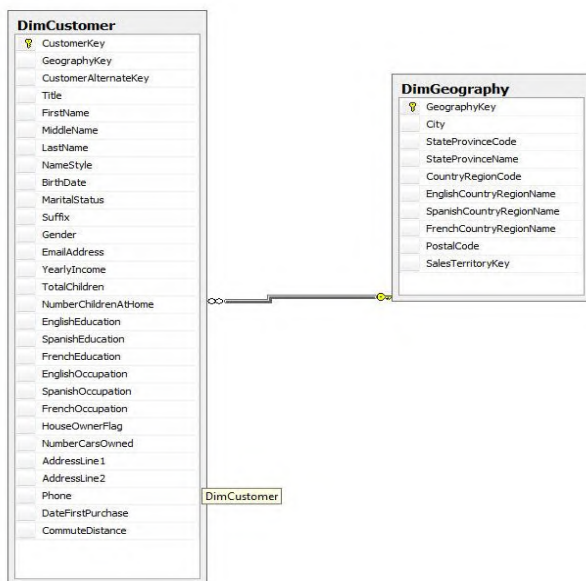


Рис. 9.4 – Діаграма, що показує зв'язок таблиць dbo.DimCustomer і dbo.DimGeography

	CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	MiddleName
1	11000	26	AW00011000	NULL	Jon	V
2	11001	37	AW00011001	NULL	Eugene	L
3	11002	31	AW00011002	NULL	Ruben	NULL
4	11003	11	AW00011003	NULL	Christy	NULL
5	11004	19	AW00011004	NULL	Elizabeth	NULL
6	11005	22	AW00011005	NULL	Julio	NULL
7	11006	8	AW00011006	NULL	Janet	G
8	11007	40	AW00011007	NULL	Marco	NULL
9	11008	32	AW00011008	NULL	Rob	NULL

Рис. 9.5 – Дані з vTargetMail (a) і DimGeography

Але повернемося до створюваного DSV. Останнє вікно майстра дозволяє ввести ім'я DSV. Назвемо його TargetMail_dsv (рис. 9.6)

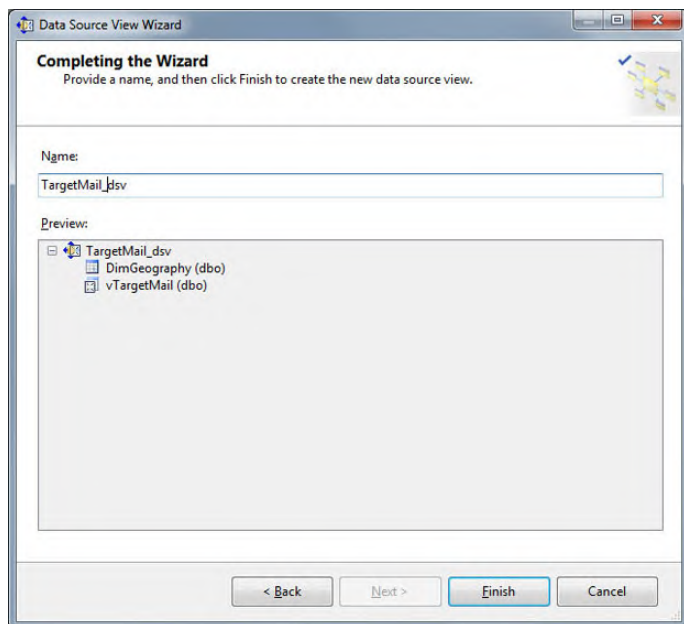


Рис. 9.6 – Діалогове вікно створення джерела даних: введення імені DSV

Після того, як уявлення джерела даних створено, відкривається візуальний редактор, що дозволяє зокрема, вказати зв'язок між таблицями. Спершу зв'язок не визначено (рис. 9.7), оскільки у вихідній БД vTargetMail є не таблицею, а поданням.

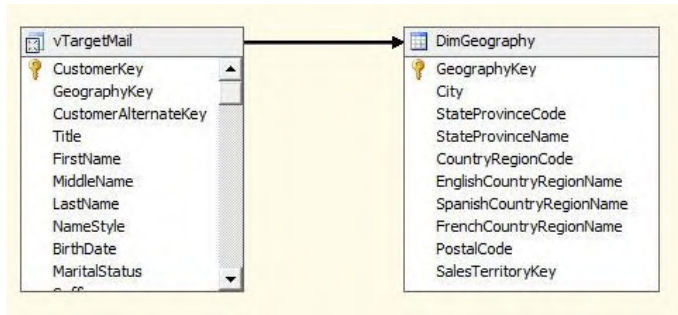


Рис. 9.7 – Схема зв'язку між таблицями бази даних

Але зв'язок можна визначити, “простягнувши” її мишею на діаграмі від стовпчика зовнішнього ключа до відповідного первинному ключу. Особливо важливо це зробити, якщо для аналізу використовуються варіанти з вкладеними таблицями.

Завдання. Відповідно до наведеного описом створіть уявлення джерела даних.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Що є найменшою логічною одиницею роботи з даними при інтелектуальному аналізі?
2. Які два основних типи атрибутів існують для алгоритмів інтелектуального аналізу?
3. Приведіть додаткові типи атрибутів
4. Про що вказують стани атрибутів «Missing» і «Existing»?
5. Дайте означення варіанту. Покажіть як варіант може подаватися у інтелектуальному аналізі.
6. Для чого використовується ключ варіанта?
7. Що таке структура інтелектуального аналізу?
8. Які можуть бути типи даних для стовпців структури інтелектуального аналізу?
9. На що вказує тип вмісту Discrete?

10. Які дані описує тип Continuous? Наведіть приклад таких даних.

ЛІТЕРАТУРА

1. Data Mining Extensions (DMX) Reference [Електронний ресурс] // SQL Docs. – 2018. – Режим доступу: <https://docs.microsoft.com/en-us/sql/dmx/data-mining-extensions-dmx-reference?view=sql-server-2017>.

РОЗДІЛ 10.

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ВБУДОВАНИХ АЛГОРИТМІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В СКБД MICROSOFT SQL SERVER 2012

10.1 DMX. Параметри алгоритмів інтелектуального аналізу даних: спрощений метод Баєса, алгоритм дерева рішень, метод лінійної регресії

Спрощений метод Баєса

У попередніх розділах було розглянуто формат команд DMX для створення структур і моделей ІАД. Зокрема, з використанням наступної команди можна створити структуру, яка містить одну модель [1]:

```
CREATE [SESSION] MINING MODEL <model>
(
  [(<column Definition list>)]
)
USING <algorithm> [(<parameter list>)]
[WITH DRILLTHROUGH]
```

Під час створення моделі з використанням спрощеного алгоритму Баєса, алгоритму дають Microsoft_Naive_Bayes.

Нижче наведений приклад створення моделі, яка, на основі значень атрибутів Age (вік), Number Cars Owned (число машин у власності) прогнозує значення атрибуту Bike Buyer (покупець велосипеда), ключовий атрибут – Customer Key.

```
CREATE MINING MODEL vTM1_NB (
  [Customer Key] LONG KEY,
  [Age] LONG DISCRETIZED,
  [Number Cars Owned] LONG DISCRETE,
  [Bike Buyer] LONG DISCRETE PREDICT_ONLY)
USING Microsoft_Naive_Bayes
```

На основі даної інструкції створена структура vTM1_NB_Structure і модель vTM1_NB з параметрами, які прийняті за замовчуванням [2]. Дані параметри представлені на рисунку 10.1 (перегляд параметрів алгоритму в середовищі BI Dev Studio).

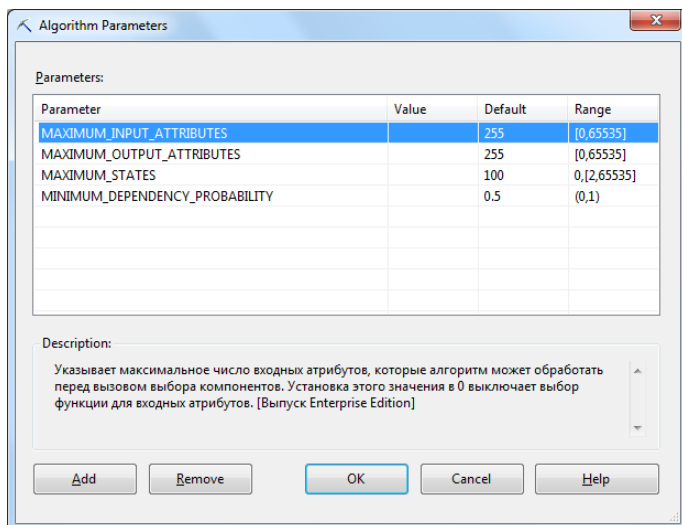


Рис. 10.1 – Діалогове вікно вибору параметрів для спрощеного алгоритму Баєса

У випадку, якщо значення параметрів за замовчуванням не підходять, можна, під час створення моделі або вказати їх явно, або згодом змінити їх [3].

Дерева рішень і лінійна регресія

Створення моделі за використанням алгоритму Microsoft Decision Trees для прогнозування дискретного атрибуту, застосовуючи мову DMX може мати наступний вигляд:

```
CREATE MINING MODEL vTM1_DT (
  [Customer Key] LONG KEY,
  [Age] LONG CONTINUOUS,
  [Number Cars Owned] LONG DISCRETE,
  [Bike Buyer] LONG DISCRETE PREDICT_ONLY)
USING Microsoft Decision Trees
```

Зауважимо, що в даному випадку може використовуватись тип вмісту Continuous, на відміну від спрощеного алгоритму Баєса, розглянутого раніше, і, під час створення структури і моделі необов'язково проводити дискретизацію неперервних числових параметрів. Проте, необхідно враховувати, що за дискретного

прогнозного атрибута та неперервних вхідних даних, автоматично дискретизується вхід неперервних стовпців [6].

По-іншому виглядатиме створення моделі для прогнозування значень неперервного атрибута. Алгоритму потрібно буде використовувати незалежну змінну-регресор. Це можна вказати явно під час створення моделі. Наступний приклад показує прогнозування річного доходу в залежності від віку і числа машин. В якості потенційного регресора можна явно вказати вік.

```
CREATE MINING MODEL vTM2_DT (  
  [Customer Key] LONG KEY,  
  [Age] LONG REGRESSOR CONTINUOUS,  
  [Number Cars Owned] LONG DISCRETE,  
  [Yearly Income] LONG CONTINUOUS PREDICT_ONLY)  
USING Microsoft_Decision_Trees
```

Якщо стовпець вказаний як регресор, це не означатиме, що в остаточній моделі він буде використаний в цій якості. В іншому випадку, алгоритм дерева прийняття рішень розбиватиме набір даних на області зі значимими шаблонами; тому для стовпця може бути не заданий прапорець REGRESSOR.

Для того, щоб забезпечити використання конкретного регресора може бути застосований параметр моделі FORCE_REGRESSOR. Можливе застосування даного параметра з алгоритмом дерева прийняття рішень та алгоритмом лінійної регресії.

Наступний приклад показує, як створити модель алгоритму лінійної регресії:

```
CREATE MINING MODEL X_LR (  
  [ID] LONG KEY,  
  [X] LONG REGRESSOR CONTINUOUS,  
  [Y] LONG CONTINUOUS PREDICT)  
USING Microsoft_Linear_Regression
```

В розділі 4.2 розглядалося, що до використання особливого варіанту алгоритму дерева рішень з параметрами, які накладають обмеження на поведінку алгоритму і визначають вимоги до використання певних типів даних на вході призводить вказівка алгоритму лінійної регресії. Нижче наведені основні вимоги:

- 1) У вміст кожної моделі повинен входити один числовий або текстовий стовпець, який ідентифікує кожен запис (тип key). Не допускаються складові ключі.

2) Повинен бути як мінімум один прогнозований стовпець. Можливе включення в модель декількох прогнозованих атрибутів, проте, їх типи даних повинні бути неперервні числові. Як прогнозований атрибут тип даних `datetime` не можна використовувати навіть тоді, коли власний формат зберігання даних є числовим.

3) Вхідні стовпці повинні містити неперервні числові дані, які повинні бути відповідного типу.

10.2 DMX. Параметри алгоритмів інтелектуального аналізу даних: часові ряди, алгоритм кластеризації

Microsoft Time Series

Опишемо особливості алгоритму Microsoft Time Series. Розглянемо процес створення структури і моделі. Нехай маємо часовий ряд, який чергується, що описує продажу (в штуках) велосипедів різних марок, з розбиттям по місяцях і регіонах. Подібний формат поданий в табл. 10.1.

Таблиця 10.1 – Приклад формату початкових даних для алгоритму часових рядів

DateSeries	Model Region	Quantity
2009 04	M200 Europe	30
2009 05	M200 Europe	40
...	...	...
2009 04	M200 Pacific	20
...	...	...

Нижче наведено приклад коду, що створює структуру інтелектуального аналізу:

```
CREATE MINING STRUCTURE TimeSeries1_structure
([DateSeries] DATE KEY TIME,
 [Model Region] TEXT KEY,
 [Quantity] LONG CONTINUOUS)
```

Під час створення структури необхідно зауважити, щоб ключ містив атрибут, що містить позначку часу (у прикладі це `DateSeries`). Також, для того, щоб алгоритм працював коректно, необхідно, щоб аналізований ряд був без пропусків. Тому, не слід резервувати дані в

структурі для тестового набору (виконуване за допомогою інструкції WITH HOLDOUT). Розглянемо приклад створення моделі:

```
ALTER MINING STRUCTURE TimeSeries1_structure
ADD MINING MODEL TimeSeries1_TS ([DateSeries],
    [Model Region],
    [Quantity] PREDICT)
USING Microsoft_Time_Series (PERIODICITY_HINT
=
    '{12}', MISSING_VALUE_SUBSTITUTION = 0)
```

Звернемо увагу на наступних два параметра. Параметр PERIODICITY_HINT – вказує на наявність періодичності 12 (за числом місяців у році). Параметр MISSING_VALUE_SUBSTITUTION вказує на необхідність замінення нулем пропущених значень елементів ряду. В даному випадку, це означає, що у випадку відсутності інформації про продаж в даному місяці даної моделі велосипедів вважатимемо, що продажів не було. Параметри алгоритму можливо переглядати та змінювати після створення моделі, наприклад, через графічний інтерфейс BI Dev Studio (рис.10.2).

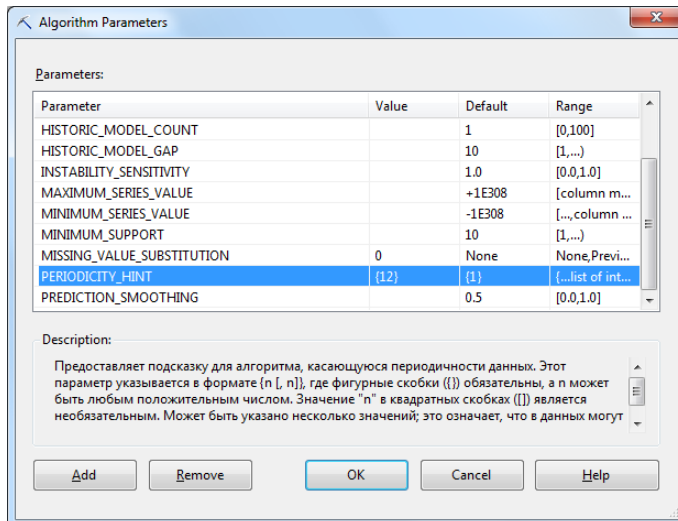


Рис. 10.2 – Діалогове вікно перегляду параметрів алгоритму Microsoft Time Series

В розділі 5.2 було зазначено, що алгоритм часових рядів Microsoft (Microsoft Time Series) складається з двох алгоритмів:

1) Алгоритм «Дерево авторегресії з перехресним прогнозуванням» (ARTхр), який використовується для прогнозування наступного значення ряду; з'явився в SQL Server 2012.

2) Алгоритм «Інтегровані ковзаючі середні авторегресії» (ARIMA), що в даній області є галузевим стандартом; доданий в SQL Server 2012, для підвищення точності довгострокового прогнозування.

Вважають, що в порівнянні з ARIMA алгоритм ARTхр, дає більш точний прогноз для найближчих значень ряду, але він має значно меншу точність в довгострокових прогнозах. За стандартних налаштувань використовуватимуться результати роботи обох алгоритмів. Водночас, використовуючи параметри моделі можна визначити ступінь впливу кожного з алгоритмів на підсумковий результат (рис.10.3), або задати використання тільки одного алгоритму

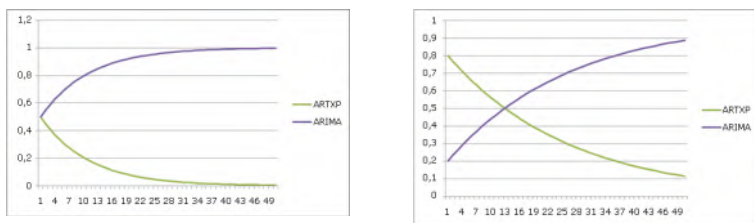


Рис. 10.3 – Коефіцієнти, за якими обраховуються результати прогнозування алгоритмів ARIMA і ARTхр
а) PREDICTION_SMOOTHING = 0,5;
б) PREDICTION_SMOOTHING = 0,2 [4]

На рисунку 10.3.а показано об'єднання результатів алгоритмів, у випадку, коли параметр PREDICTION_SMOOTHING має значення за замовчуванням – 0,5. На початку обоє з алгоритмів ARIMA і ARTхр отримують рівні вагові коефіцієнти, але вага алгоритму ARIMA зростає в залежності від збільшення кількості кроків прогнозування.

Microsoft Clustering

Розглянемо наступний приклад створення моделі ІАД на основі алгоритму Microsoft Clustering. Нехай, для клієнтів фірми з відомим віком, сімейним станом, числом дітей, необхідно провести сегментацію (кластеризацію). Спочатку створимо структуру:

```
CREATE MINING STRUCTURE Customer_structure
([CustomerKey] LONG KEY,
 [MaritalStatus] TEXT DISCRETE,
 [TotalChildren] LONG DISCRETE,
 [Age] LONG CONTINUOUS)
```

До створеної структури додамо модель, розроблену на основі алгоритму кластеризації. Атрибути MaritalStatus, TotalChildren, Age розглядатимуться як вхідні. Також зазначимо, скільки кластерів ми хочемо отримати:

```
ALTER MINING STRUCTURE Customer_structure ADD
MINING MODEL Customer CL
([CustomerKey],
 [MaritalStatus],
 [TotalChildren],
 [Age])
USING Microsoft_Clustering (CLUSTER_COUNT =
6) WITH DRILLTHROUGH
```

Нижче наведені вимоги до даних для моделі кластеризації:

- до складу кожної моделі необхідним є включення одного ключового числового або текстового стовпця, не допускається застосування складових ключів;

- для кожної моделі необхідним є вміст хоча б одного вхідного стовпця, який включає в себе значення, використовувані для формування кластерів;

- можливим є наявність необов'язкового прогнозованого стовпця: в даному алгоритмі для формування моделі немає потреби в прогнозованому стовпцю, проте, існує можливість додавання прогнозованого стовпця з даними майже будь-якого типу. Наприклад, у випадку, коли необхідно передбачити дохід замовника за допомогою кластеризації з такими показниками, як регіон і вік, можна вказати дохід як PredictOnly, і в якості вхідних даних використовувати стовпці з інформацією про регіон і вік.

Наступний приклад коду виконує обробку моделі і прогнозний запит, в якому, за допомогою функції Cluster() виводиться номер кластера, а функцією ClusterProbability() виводиться оцінка ймовірності того, що цей варіант належить даному кластеру.

```
// Заповнення структури та її обробка INSERT
INTO MINING STRUCTURE [Customer_structure]
```

```

([CustomerKey],          [Age],          [MaritalStatus],
[TotalChildren])
    OPENQUERY ([Adventure Works DW],
        'SELECT          [CustomerKey],          [Age],
[MaritalStatus],          [TotalChildren]          FROM
dbo.vTargetMail')
    / / Кінець обробки структури GO
    / / Запит, який повертає номер кластера та
ймовірність SELECT t. *, Cluster (),
ClusterProbability ()
FROM Customer_CL NATURAL PREDICTION JOIN
OPENQUERY ([Adventure Works DW],
    'SELECT          [CustomerKey],          [Age],
[MaritalStatus], [TotalChildren]
FROM dbo.vTargetMail ') as t

```

Нижче перераховані основні параметри алгоритму, які можна застосувати для моделі.

CLUSTERING_METHOD – число, яке вказує на те, який метод кластеризації використовує алгоритм.

10.3 DMX. Параметри алгоритмів інтелектуального аналізу даних: алгоритм взаємозв'язків, кластеризація послідовностей

Association Rules

Розглянемо код мови DMX, який створює структуру інтелектуального аналізу і модель, що використовує алгоритм Microsoft Association Rules. Зверніть увагу, що тут будуть використані вкладені таблиці, що містять перелік товарів в замовленні (як і в попередніх прикладах, мова йде про дані з бази Adventure Works DW).

```

CREATE MINING STRUCTURE SalesData ([Order
Number] TEXT KEY,
Products TABLE (Product TEXT KEY,
Category TEXT DISCRETE
)
)
/ / Структура створена GO
ALTER MINING STRUCTURE SalesData ADD MINING
MODEL Recommendations ([Order Number],
Products PREDICT (Product)

```

```

) USING
Microsoft_Association_Rules (MINIMUM_SUPPORT
= 10,
MINIMUM_PROBABILITY = 0.4)

```

В даному випадку необхідно звернути увагу на наступне. Варіант містить тільки один ключовий стовпець і один прогнозований стовпець. Як правило, прогнозується ключовий стовпець вкладеної таблиці, у нашому прикладі – перелік придбаних продуктів. Ці значення повинні бути дискретними або дискретизованими. Вхідні стовпці також повинні бути дискретними або дискретизованими.

Microsoft Sequence Clustering

Нижче наведено фрагмент коду мовою DMX, який створює модель інтелектуального аналізу, що ґрунтується на алгоритмі Microsoft Sequence Clustering. В даному випадку завдання – аналіз послідовностей переходу клієнтів по розділам web-сайту. Варіант формують поля ідентифікатор клієнта (CustomerGuid), інформація про його місце розташування (GeoLocation), вкладена таблиця з ідентифікатором посилання в послідовності переходів клієнта (SequenceID) і зазначенням категорії посилання (URLCategory).

```

CREATE MINING MODEL WebSequence (
CustomerGuid TEXT KEY,
GeoLocation TEXT DISCRETE,
ClickPath TABLE PREDICT
(SequenceID LONG KEY SEQUENCE,
URLCategory TEXT DISCRETE PREDICT
)
)
USING Microsoft_Sequence_Clustering

```

Після того, як модель створена та оброблена, вона може використовуватися для побудови прогнозів. Наприклад, може бути запропоновано найбільш ймовірне продовження послідовності, початок якої відомо.

Для формування прогнозу використовується функція PredictSequence(), формат якої передбачає використання різних комбінацій параметрів. PredictSequence (<table column reference>), де параметр в дужках – назва вкладеної таблиці, дасть найбільш ймовірну послідовність значень. PredictSequence (<table column reference>, n): прогнозування *n* наступних значень з вкладеної таблиці. Pre-

dictSequence(<table column reference>, n, m): прогнозування значень з n -го по m -й.

Повернувшись до розглянутого прикладу з web-сайтом [20], можна використати наступний запит для того, щоб дізнатися два найбільш вірогідних переходи клієнта, який дивився сторінки по розділах страхування (англ. insurance), кредити (англ. loan):

```
SELECT PredictSequence (ClickPath, 2) AS
Sequences FROM WebSequence NATURAL PREDICTION JOIN
(SELECT (SELECT 1 AS SequenceID, 'Insurance' AS
URLCategory UNION
SELECT 2 AS SequenceID, 'Loan' AS
URLCategory)
AS ClickPath) AS T
```

10.4 DMX. Параметри алгоритмів інтелектуального аналізу даних: алгоритми нейронних мереж, логістична регресія

Алгоритми нейронних мереж та логістичної регресії

Як було розібрано в розділі 2.9, алгоритм Microsoft Logistic Regression є варіантом алгоритму Microsoft Neural Network, в якому не використовується прихований шар нейронної мережі. Інакше кажучи, параметр HIDDEN_NODE_RATIO встановлений рівним 0 (докладніше про параметр – див далі).

Тому параметри алгоритмів і особливості їх використання в чомусь схожі. Модель, заснована на нейронній мережі, повинна містити, принаймні, один вхідний і один вихідний (прогнозований) стовпець. Вкладена таблиця в якості прогнозованого стовпця використовуватися не може, але допустимо використовувати вкладену таблицю в як вхідний атрибут.

Приклад коду DMX, що створює в існуючій структурі vTargetMail_structure2 модель на основі нейронних мереж, наведено нижче.

```
ALTER MINING STRUCTURE
[vTargetMail_structure2]
ADD MINING MODEL [vTargetMail2_NN]
([Customer Key],
[Commute Distance], [Age],
[Number Cars Owned],
[Yearly Income],
```

[Bike Buyer] PREDICT)
 USING Microsoft Neural Network

Якщо створюється модель на основі алгоритму логістичної регресії, це вказується ключовими словами USING Microsoft_Logistic_Regression.

Нижче перераховані параметри алгоритмів.

HIDDEN_NODE_RATIO – вказує співвідношення числа між вхідними, прихованими і вихідними шарами.

Наступна формула визначає початкову кількість нейронів в прихованому шарі:

$$HIDDEN_NODE_RATIO \times \sqrt{(k - \text{ть_вх\ddot{и}д.нейрон\ddot{и}в} \times k - \text{ть_вих.нейрон\ddot{и}в})}$$

Значення за замовчуванням – 4,0.

Вміст моделі

Будова моделі, що побудована на алгоритмі нейронних мереж, не можна назвати «інтуїтивно зрозумілим», тому потрібно зупинитися на його розгляді більш детально (рис. 10.4).



Рис. 10.4 – Схематична структура вмісту моделі, що базується на методі штучних нейронних мереж [3]

Кожна нейронна мережа має лише один батьківський вузол, який відображає модель та її метадані і також вузол граничної статистики (NODE_TYPE = 24), що містить описову статистику щодо

вхідних атрибутів. Під цими двома вузлами розташоване не менш двох інших вузлів.

Перший вузол (NODE_TYPE = 18) завжди є верхнім вузлом вхідного шару. Під цим верхнім вузлом знаходяться вхідні вузли (NODE_TYPE = 21), що містять фактичні вхідні атрибути та їхні значення (рис. 10.5).

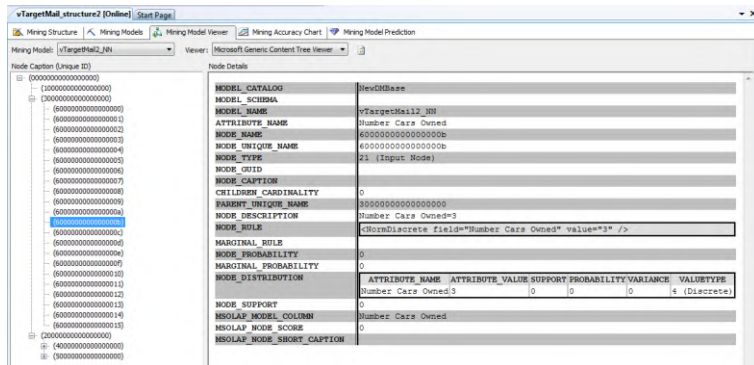


Рис. 10.5 – Вікно з вмістом «вхідного» вузла моделі

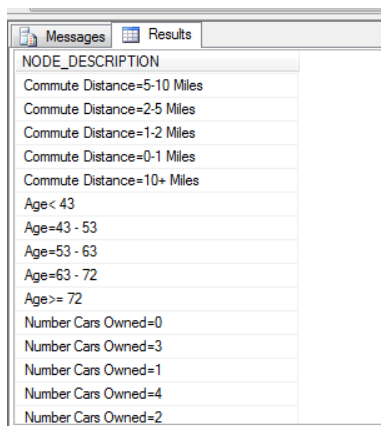
У кожному з наступних вузлів міститься окрема підмережа (NODE_TYPE = 17). Кожна підмережа містить власний приований шар (NODE_TYPE = 19) і вихідний шар (NODE_TYPE = 20).

Кожна підмережа (NODE_TYPE = 17) представляє аналіз впливу вхідного шару на дискретний прогнозований атрибут. Якщо прогнозованих виходів декілька, буде створено й кілька підмереж. Приований шар кожній підмережі містить кілька приованих вузлів (NODE_TYPE = 22), що містять дані про вагові коефіцієнти для кожного переходу який завершується в даному приованому вузлі.

Вихідний шар (NODE_TYPE = 20) містить початкові вузли (NODE_TYPE = 23), де знаходяться унікальні значення прогнозованого атрибута. Якщо прогнозований атрибут має неперервний числовий тип даних, то для нього буде тільки один вихідний вузол.

```
SELECT NODE_DESCRIPTION
FROM [vTargetMail2_NN]. CONTENT
WHERE NODE TYPE = 21
```

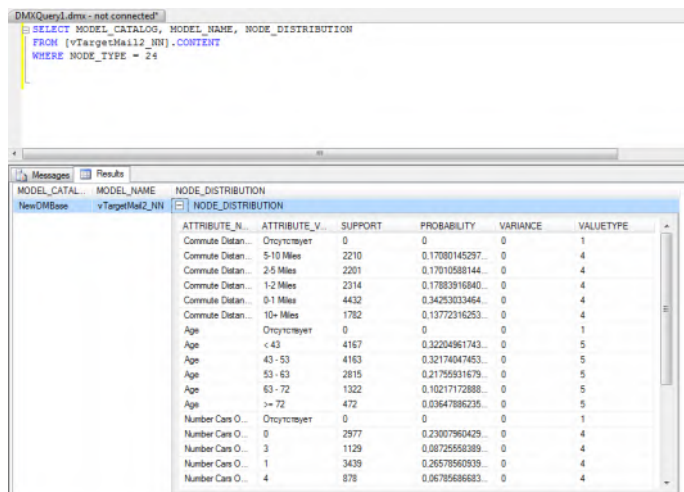
Опис вузлів можна розглянути за допомогою запиту до моделі. Далі наведено код запиту для вхідних вузлів (тип 21), а також результат його виконання представлений на рисунку 10.6.



NODE_DESCRIPTION
Commute Distance=5-10 Miles
Commute Distance=2-5 Miles
Commute Distance=1-2 Miles
Commute Distance=0-1 Miles
Commute Distance=10+ Miles
Age < 43
Age=43 - 53
Age=53 - 63
Age=63 - 72
Age >= 72
Number Cars Owned=0
Number Cars Owned=3
Number Cars Owned=1
Number Cars Owned=4
Number Cars Owned=2

Рис. 10.6 – Вікно з переліком вхідних вузлів

На рисунку 10.7 подано текст та результат виконання запиту до вузла граничної статистики (тип 24).



```

SELECT MODEL_CATALOG, MODEL_NAME, NODE_DISTRIBUTION
FROM [vTargetMail2_NN].CONTENT
WHERE NODE_TYPE = 24

```

MODEL_CATALOG	MODEL_NAME	NODE_DISTRIBUTION	ATTRIBUTE_N...	ATTRIBUTE_V...	SUPPORT	PROBABILITY	VARIANCE	VALUETYPE
NewCRBase	vTargetMail2_NN	1	Commute Distanc...	Отсутствует	0	0	0	1
NewCRBase	vTargetMail2_NN	2	Commute Distanc...	5-10 Miles	2210	0.17080145297...	0	4
NewCRBase	vTargetMail2_NN	3	Commute Distanc...	2-5 Miles	2201	0.17010588144...	0	4
NewCRBase	vTargetMail2_NN	4	Commute Distanc...	1-2 Miles	2314	0.17883916940...	0	4
NewCRBase	vTargetMail2_NN	5	Commute Distanc...	0-1 Miles	4432	0.34283033464...	0	4
NewCRBase	vTargetMail2_NN	6	Commute Distanc...	10+ Miles	1782	0.13772316253...	0	4
NewCRBase	vTargetMail2_NN	7	Age	Отсутствует	0	0	0	1
NewCRBase	vTargetMail2_NN	8	Age	< 43	4167	0.32204961743...	0	5
NewCRBase	vTargetMail2_NN	9	Age	43 - 53	4163	0.32174047453...	0	5
NewCRBase	vTargetMail2_NN	10	Age	53 - 63	2815	0.21755931679...	0	5
NewCRBase	vTargetMail2_NN	11	Age	63 - 72	1322	0.10217172888...	0	5
NewCRBase	vTargetMail2_NN	12	Age	>= 72	472	0.03647886235...	0	5
NewCRBase	vTargetMail2_NN	13	Number Cars O...	Отсутствует	0	0	0	1
NewCRBase	vTargetMail2_NN	14	Number Cars O...	0	2977	0.23007960425...	0	4
NewCRBase	vTargetMail2_NN	15	Number Cars O...	3	1129	0.08725558389...	0	4
NewCRBase	vTargetMail2_NN	16	Number Cars O...	1	3439	0.26578560939...	0	4
NewCRBase	vTargetMail2_NN	17	Number Cars O...	4	878	0.06785636683...	0	4

Рис. 10.7 – Результат запиту до вузла граничної статистики

Самостійна робота №11. Робота з моделями
інтелектуального аналізу даних з SQL Server
Management Studio

SQL Server Management Studio, в першу чергу, розглядається як засіб створення і адміністрування реляційних баз даних. У той же час, у випадку обізнаності з мовою DMX, деякі завдання, що пов'язані з проведенням інтелектуального аналізу даних, можна швидко вирішити з цього середовища.

Для початку роботи, після запуску Management Studio, повинні бути підключені до аналітичних служб SQL Server (рис. 10.7).



Рис. 10.8 – Діалогове вікно підключення до аналітичних служб
СКБД SQL Server 2012

Для цього в списку, що випадає **Server Type** треба вибрати **Analysis Server**, а в поле **Server Name** ввести ім'я сервера ("local", якщо ви підключаєтеся до неіменованого примірника, SQL Server встановленому на тому ж комп'ютері, де запущена Management Studio). Після підключення можна переглянути наявні бази аналітичних служб і створювати нові, а також створювати запити, виконувати резервне копіювання, відновлення та ін. Розглянемо створення нової моделі за допомогою запиту DMX.

У контекстному меню для обраної бази даних знайдіть **NewQuery ->DMX** (рис. 10.9).

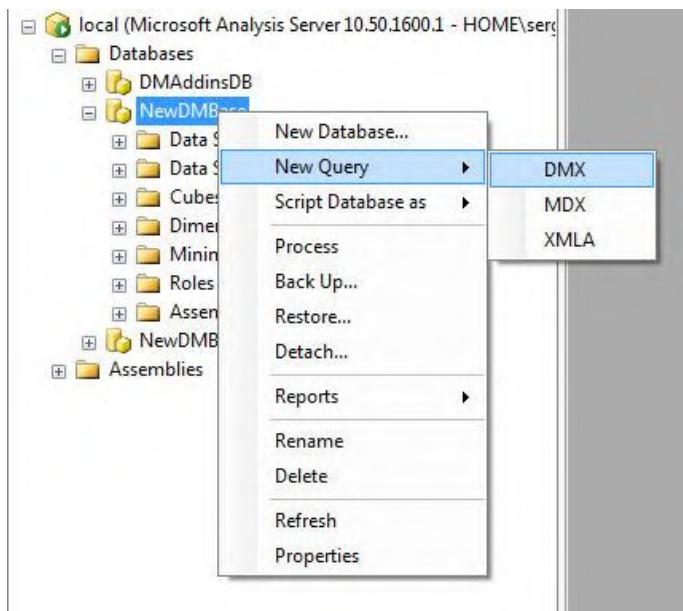


Рис. 10.9 – Вікно створення запиту мовою DMX

Буде відкрито нове вікно, куди треба ввести тест запиту (рис. 10.10). Зверніть увагу на те, щоб на панелі інструментів у спадному списку з назвою бази даних була вказана саме та база, в якій ви хочете створити модель. Виконаємо наступний запит

```
CREATEMININGMODEL vTM1_NB (
  [Customer Key] LONGKEY,
  [Age] LONGDISCRETIZED,
  [Number Cars Owned] LONGDISCRETE,
  [Bike Buyer] LONGDISCRETEPREDICT_ONLY)
USINGMICROSOFT_Naive_Bayes
```

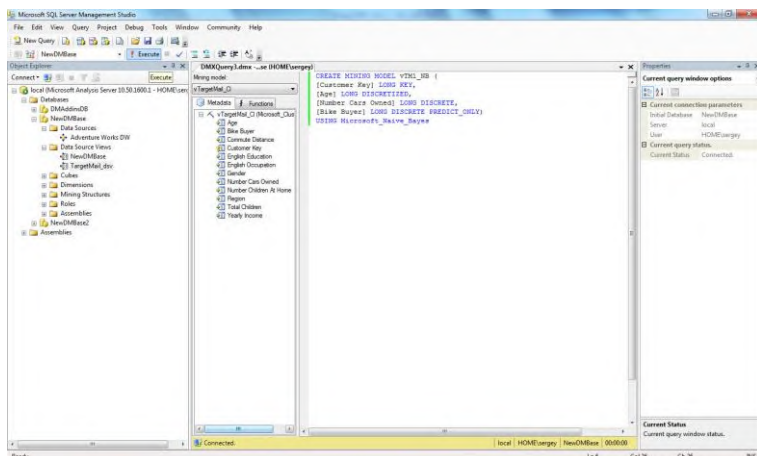


Рис. 10.10 – Вікно створення запиту мовою DMX (продовження)

Наведений вище фрагмент коду створює структуру і модель інтелектуального аналізу для прогнозування атрибута [BikeBuyer] на основі атрибутів [Age] і [NumberCarsOwned]. Подібну модель можна навчити на вже відомих нам даних з бази AdventureWorksDW. З огляду на те, що ми не створювали структуру інтелектуального аналізу, а скористалися оператором CREATEMININGMODEL ім'я створеної структури буде формуватися як <імямоделі>_Structure (рис. 10.11).

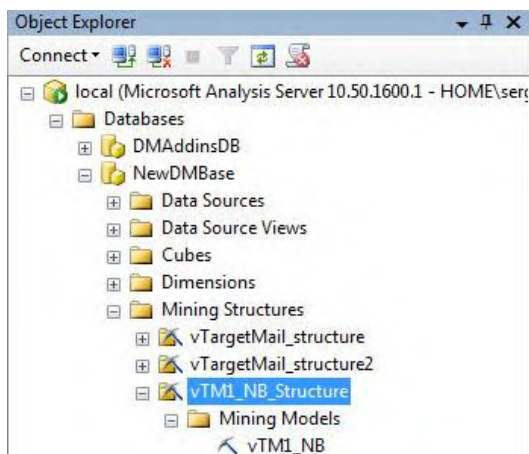


Рис. 10.11 – Вікно з створеною структурою і моделлю

До речі. Якщо нам знадобиться скрипт, який створює такий же об'єкт, як вже є в базі даних, його можна отримати як показано на рисунку 10.12.

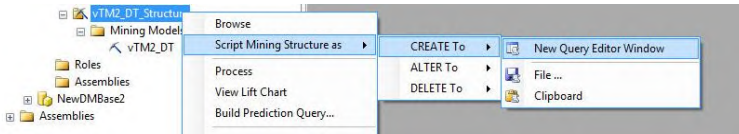


Рис. 10.12 – Генерація скрипта на основі існуючого об'єкта (структури або моделі)

Правда, отримаємо ми не код мовою DMX, а опис в форматі XMLA. Створений скрипт можна змінити, виконати для іншої бази, зберегти та ін.

Але повернемося до роботи з моделлю інтелектуального аналізу даних. Після створення її треба обробити, завантаживши в неї дані. Робиться це за допомогою оператора INSERT INTO. Серед Management Studio полегшує написання запиту, тому що назва об'єктів і атрибутів можна просто перетягувати мишею з вікна Object Explorer.

```
INSERTINTOMININGMODEL [vTM1_NB] ([Customer
Key], [Age], [Number Cars Owned], [Bike Buyer])
OPENQUERY ([Adventure Works DW], 'Select
CustomerKey, Age, [NumberCarsOwned], [BikeBuyer]
FROM [vTargetMail]')
```

Зверніть увагу, що в реченні OPENQUERY ми використовуємо створений під час “Початок роботи в BI Dev Studio” джерело даних [AdventureWorks DW].

Після обробки, клацнувши правою клавішею миші на моделі у вікні Object Explorer, можна з контекстного меню вибрати опцію Browse і переглянути модель, аналогічна тій, як це робилося в BI Dev Studio (рис. 10.13).

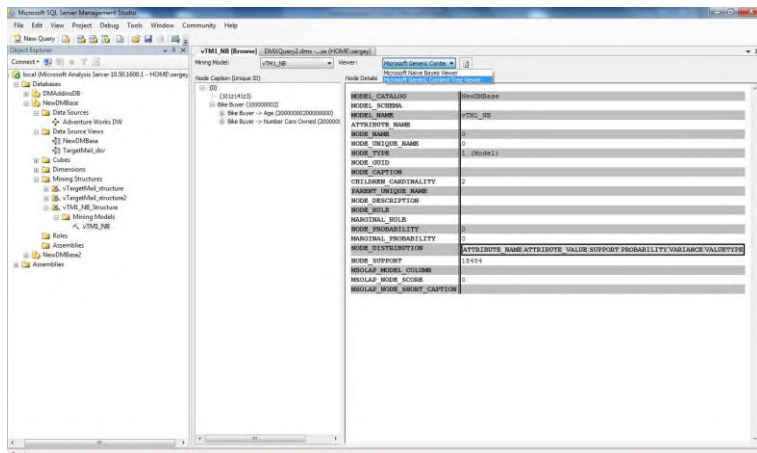


Рис. 10.13 – Перегляд моделі з Management Studio

Тепер, коли модель оброблена, можна написати який прогнозує запит. Спочатку дізнаємося, чи стане покупцем велосипеда клієнт у віці 36 років, який має 1 автомобіль. Нас цікавить прогноз, тобто значення атрибута Bike Buyer і оцінка ймовірності для цього прогнозу. Код запиту виглядає наступним чином

```
SELECT Predict ([Bike Buyer]) as Result,
PredictProbability ([Bike Buyer]) as
ResultProbability
FROM [vTM1_NB]
NATURALPREDICTIONJOIN
(SELECT 36 as [Age], 1 as [Number Cars
Owned]) as t
```

Його виконання призведе до висновку результату для Bike Buyer і оцінки ймовірності (зверніть увагу, щоб запит би виконаний щодо потрібної бази).

Ну і на завершення розглянемо як виконати запит прогнозу для кількох рядків одночасно. Нехай нам потрібні прогнози для перших 10 клієнтів з таблиці Dim Customer бази Adventure Works DW. Водночас, нас цікавлять тільки ті клієнти, які є в поданні vDATA MININGPrep. Потрібний нам для передбачення параметр “вік клієнта” ми також беремо з цього подання. Запит на SQL буде виглядати наступним чином:

```

SELECTDISTINCTTOP 10
dbo.DimCustomer.CustomerKey,
dbo.DimCustomer.FirstName,
dbo.DimCustomer.LastName,
dbo.vDATA MININGPrep.Age,
dbo.DimCustomer. [NumberCarsOwned]
FROM dbo.DimCustomer INNER JOIN dbo.vDATA
MININGPrep
ON dbo.DimCustomer.CustomerKey = dbo.vDATA
MININGPrep.CustomerKey
ORDER BY dbo.DimCustomer.CustomerKey

```

Прогнозний запит мовою DMX наведено нижче, а результат його виконання представлений на рисунку 10.14.

```

SELECT          t.CustomerKey,          t.FirstName,
t.LastName, Predict ([Bike Buyer]) as Result,
      PredictProbability    ([Bike Buyer])    as
ResultProbability
FROM [vTM1_NB]
PREDICTIONJOIN
OPENQUERY ([Adventure Works DW],
'SELECT          DISTINCT          TOP          10
dbo.DimCustomer.CustomerKey,
dbo.DimCustomer.FirstName,
      dbo.DimCustomer.LastName,          dbo.vDATA
MININGPrep.Age, dbo.DimCustomer. [NumberCarsOwned]
FROM dbo.DimCustomer INNER JOIN dbo.vDATA
MININGPrep ON dbo.DimCustomer.CustomerKey =
      dbo.vDATA MININGPrep.CustomerKey
ORDER BY dbo.DimCustomer.CustomerKey ') as t
ON [vTM1_NB]. [Age] = t. [Age] AND [vTM1_NB].
[Number Cars Owned] = t. [NumberCarsOwned]

```

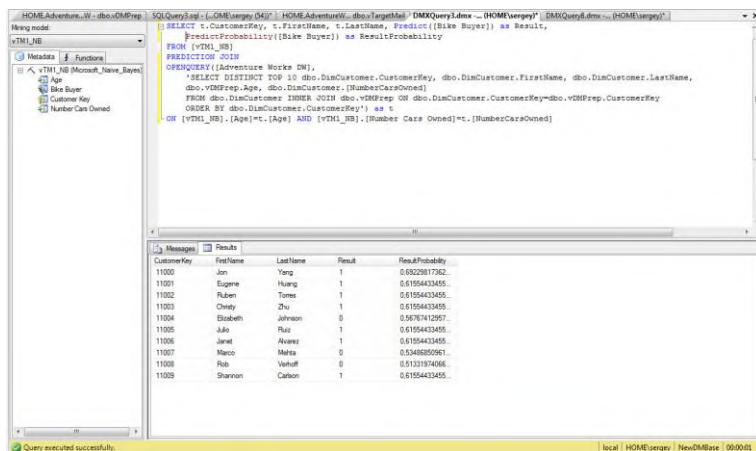


Рис. 10.14 – Результат виконання запиту

Завдання 1. Виконайте описані в самостійній роботі дії.

Завдання 2. На основі методу дерева рішень створити прогнозу модель. Вхідні атрибути – вік, стать, сімейний стан, кількість машин, рівень доходу. Прогнозований атрибут – “BikeBuyer”. Для тестування повинно бути зарезервовано 30% даних. Виконайте обробку моделі. Напишіть прогнозують запити, аналогічні розглянутим в самостійній роботі.

Самостійна робота №12. Використання алгоритму Microsoft Time Series для прогнозування значень часових рядів

На практиці ми часто стикаємося з тимчасовими рядами: біржові котирування цінних паперів, обсяги випуску товарів по місяцях, середньодобові значення температури повітря – все це приклади подібних послідовностей. У загальному випадку, часовий ряд – це набір числових значень, зібраних в послідовні моменти часу (в більшості випадків – через рівні проміжки часу). Під час роботи з тимчасовими рядами часто виникає бажання виявити залежності між поточним значенням і попередніми йому, і використовувати їх для прогнозування майбутніх значень. Таке завдання можна вирішити з використанням алгоритму Microsoft Time Series, що входить в набір стандартних алгоритмів аналітичних служб SQL Server 2012.

Для аналізу будемо використовувати дані з навчальної бази Adventure Works DW. Подання (view) dbo.vTimeSeries формувє часовий

ряд, що містить результати продажів різних моделей велосипедів по місяцях (рис. 10.15).

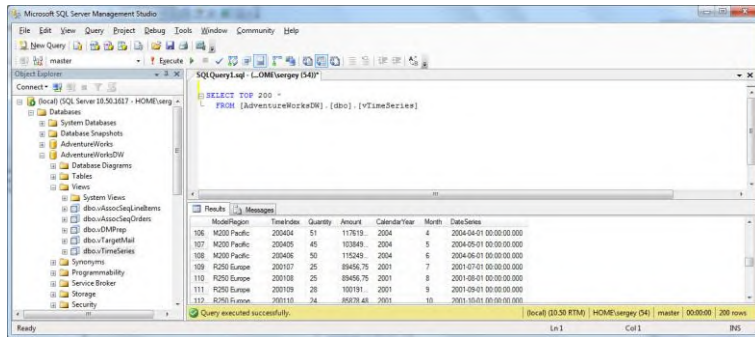


Рис. 10.15 – Дані з уявлення vTimeSeries бази AdventureWorks DW

Розглянутий набір даних представляє собою приклад чергується ряду: в ньому є кілька записів відповідних одній даті, що містять дані для різних моделей велосипеда.

Джерело даних, що вказує на базу AdventureWorks DW, у нас вже створено. Наступний крок – створити в середовищі BI Dev Studio уявлення джерела даних, що включає vTimeSeries. Параметри алгоритмів інтелектуального аналізу даних. Алгоритми нейронних мереж і логістичної регресії”. Назовем створене уявлення джерела даних TimeSeries_dsv.

Далі буде потрібно створити в середовищі BI Dev Studio модель для прогнозування рядів. Початок створення моделі: використовувати будемо дані з реляційної БД, алгоритм – Microsoft Time Series, уявлення джерела даних – TimeSeries_dsv, таблиця з даними – vTimeSeries. Тепер визначимося, які атрибути будуть потрібні для прогнозу. Ми маємо справу з чергується тимчасовим поруч, і щоб ідентифікувати один його елемент, треба вказати позначку часу і назва моделі. Значить і ключових атрибутів буде два – DateSeries (відмітка часу) і ModelRegion. Прогнозувати будемо значення обсягу продажів у грошовому еквіваленті (Amount) і кількість проданих велосипедів (Quantity). Ці атрибути також будуть розглядатися і в якості вхідних (рис. 10.16).

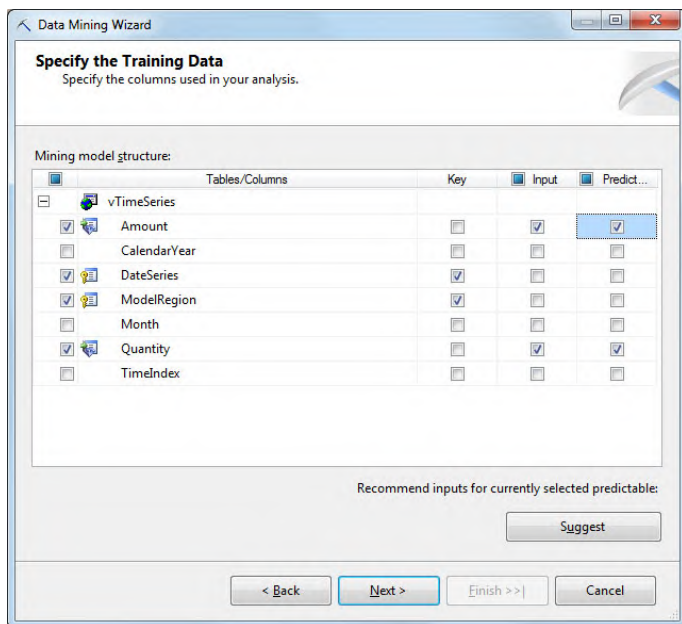


Рис. 10.16 – Діалогове вікно вибору прогнозованих атрибутів

Наступний крок – визначення типів даних і вмісту для обраних атрибутів. Тут можна погодитися з автоматично підібраними значеннями (рис. 10.17) Варто звернути увагу на тип вмісту KeyTime обраний для стовпця [DateSeries]. Стовпець з таким типом обов'язково повинен бути присутнім, якщо буде використовуватися алгоритм інтелектуального аналізу Microsoft Time Series.

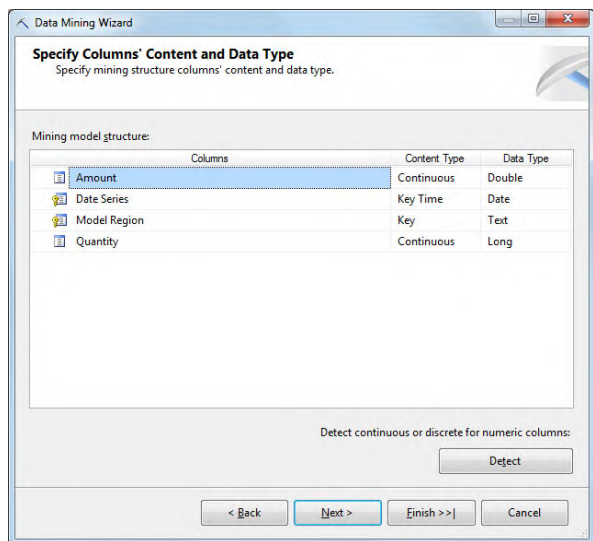


Рис. 10.17 – Діалогове вікно вибору типу вмісту і даних

Назвемо створювану структуру vTimeSeries_TS_Structure, а модель – vTimeSeries_TS (рис. 10.18).

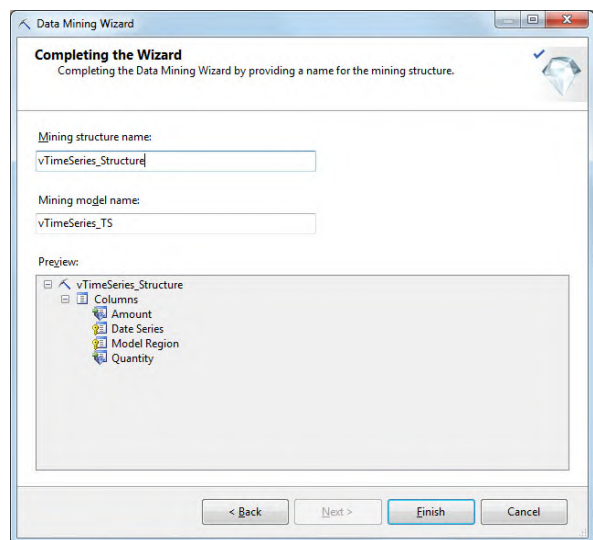


Рис. 10.18 – Діалогове вікно вибору назви моделі і структури

Після створення структури і моделі інтелектуального аналізу можна більш точно налаштувати параметри. Зокрема, у властивостях моделі можна явно вказати, що в тимчасовому ряді очікується періодичність 12 (число місяців в році, тому що у нас дані з продажами по місяцях). Така зміна налаштувань показана на рисунку 10.19.

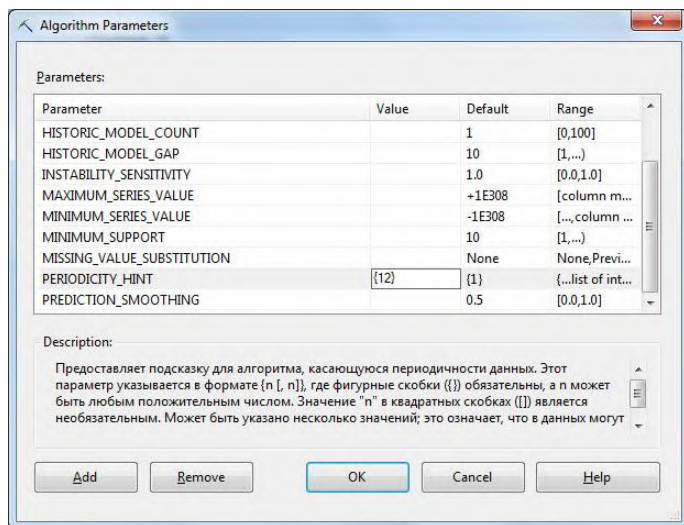


Рис. 10.19 – Діалогове вікно для вказівки періодичності – 12 (число місяців в році)

Після цього треба провести повну обробку моделі, і на вкладці MiningModelViewer з'явиться можливість ознайомитись з результатами (рис. 10.20). Безперервної лінією зображені фактичні дані, пунктиром – прогнозовані результати. Поле Predictionsteps дозволяє вказати, на скільки кроків вперед передбачається значення ряду, а випадає під ним – вибрати значення, для яких будуються графіки.

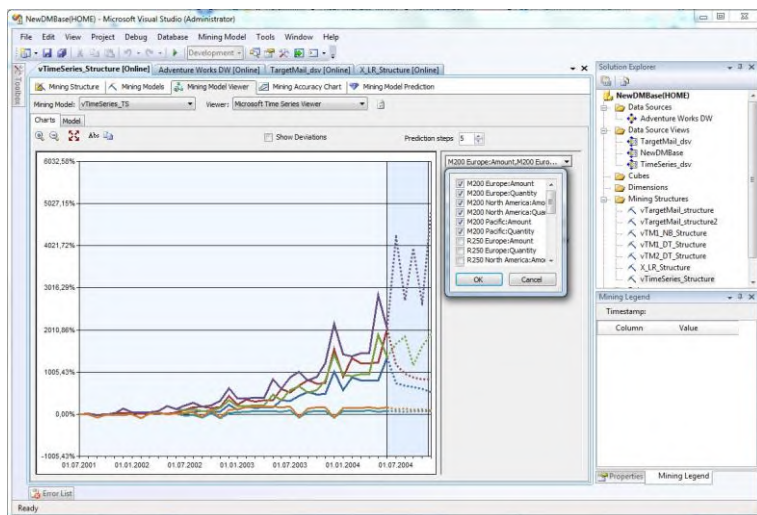


Рис. 10.20 – Перегляд графіків; пунктирна лінія – прогнозовані значення

Завдання 1. За аналогією з розглянутим прикладом створіть структуру і модель інтелектуального аналізу даних для прогнозування продажів AdventureWorks. Не вказуйте значення для періодичності. Проведіть обробку моделі, побудуйте графіки продажів велосипедів моделі M200 в різних регіонах з прогнозом на рік (12 значень). Проаналізуйте результати.

Завдання 2. Вкажіть в параметрах алгоритму періодичність. Проведіть повторну обробку моделі і знову проаналізуйте графіки прогнозу на рік. Чи є зміни? Чим можна пояснити отриманий результат?

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Наведіть приклад створення моделі з використанням спрощеного алгоритма Байеса.
2. Наведіть приклад створення моделі з використанням дерева рішень.
3. Наведіть приклад створення моделі з використанням лінійної регресії.
4. Наведіть приклад створення моделі з використанням алгоритму часових рядів.

5. Наведіть приклад створення моделі з використанням алгоритму кластеризації.
6. Наведіть приклад створення моделі з використанням алгоритму взаємозв'язків.
7. Наведіть приклад створення моделі з використанням алгоритму кластеризації послідовностей.
8. Наведіть приклад створення моделі з використанням алгоритму нейронних мереж.
9. Наведіть приклад створення моделі з використанням алгоритму логістичної регресії.
10. Опишіть параметри алгоритмів ARIMA і ARTxp, за якими обраховуються прогнозування.

ЛІТЕРАТУРА

1. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.
2. Службы SQL Server Analysis Services – интеллектуальный анализ данных. – [Електронний ресурс]. – Режим доступу: <https://docs.microsoft.com/ru-ru/sql/analysis-services/analysis-services-tutorials-ssas?view=sql-server-2017>
3. Барсегян А.А. Анализ данных и процессов : учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д.Тесс. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
4. Data Mining Extensions (DMX) Reference [Електронний ресурс] // SQL Docs. – 2018. – Режим доступу: <https://docs.microsoft.com/en-us/sql/dmx/data-mining-extensions-dmx-reference?view=sql-server-2017>.

РОЗДІЛ 11.

СТВОРЕННЯ СТРУКТУРИ І МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

В уявленні **vTargetMaily** у нас є інформація щодо клієнтів (ім'я, дата народження, сімейний стан, кількість машин тощо) і про те, придбав конкретний клієнт велосипед чи ні (стовпець **Bike Buyer** = 1, якщо велосипед купувався, інакше = 0). Шлях необхідно розділити всіх клієнтів на кілька груп, подібних за значенням параметрів. Подібна задача називається задачею кластеризації [1]. Засобами надбудов інтелектуального аналізу для MS Office 2007 ми вирішували це завдання в ході виконання **Analyze Key In fluencers** і **Detect Categories**. Тепер розберемо, як її вирішити в середовищі BI Dev Studio.

Спочатку потрібно буде зробити структуру і модель інтелектуального аналізу. Відкриємо створену раніше базу аналітичних служб в середовищі BI Dev Studio, у вікні **Solution Explorer** клацнемо правою клавішею миші на вузлі вузол **Mining Structures** і в контекстному меню виберемо **New Mining Structure** (рис. 11.1).

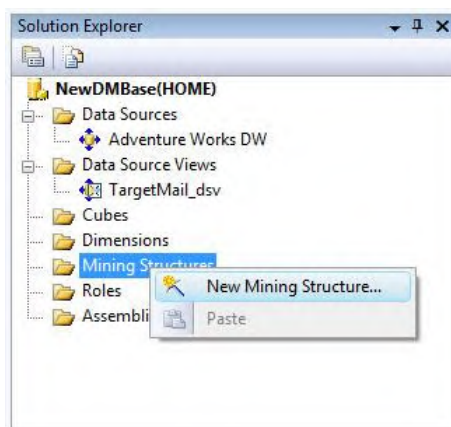


Рис. 11.1 – Діалогове вікно створення нової структури інтелектуального аналізу даних

Після виведення вікна вітання, майстер **Data Mining Wizard** здійснить запит, чи буде створювана структура заснована на

реляционном джерелі даних або на кубі OLAP (рис. 11.2). Нам потрібен перший варіант, який використовує реляційну базу даних.

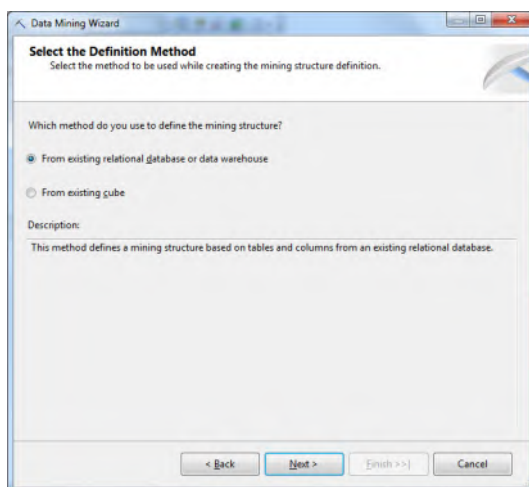


Рис. 11.2 – Діалогове вікно створення нової структури інтелектуального аналізу даних (продовження)

Наступний крок – вибір між одночасним створенням структури і моделі інтелектуального аналізу і тільки створенням структури. Для вирішення поставленого завдання нам знадобиться модель, яка використовує алгоритм кластеризації (рис. 11.3).

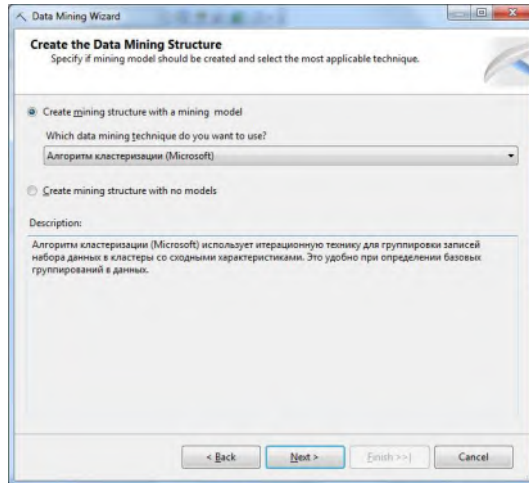


Рис. 11.3 – Діалогове вікно створення нової структури і моделі інтелектуального аналізу: вибір алгоритму

Після чого буде запропоновано вибрати вигляд джерела даних (рис. 11.4).

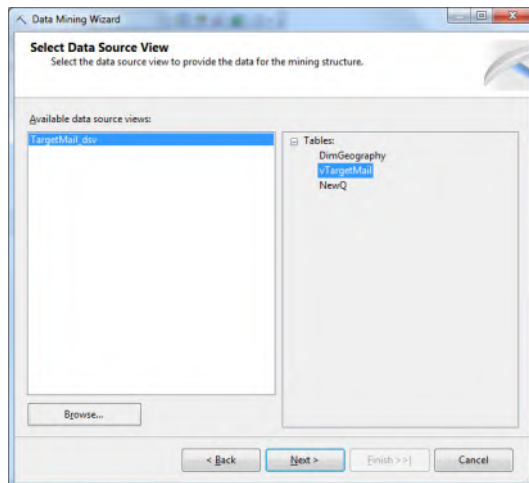


Рис. 11.4 – Діалогове вікно створення нової структури і моделі інтелектуального аналізу: вибір уявлення джерела даних

Наступний крок – вибір таблиці варіантів і вкладених таблиць. У нашому випадку, таблиця варіантів – **vTargetMail**, вкладених таблиць немає (рис. 11.5).

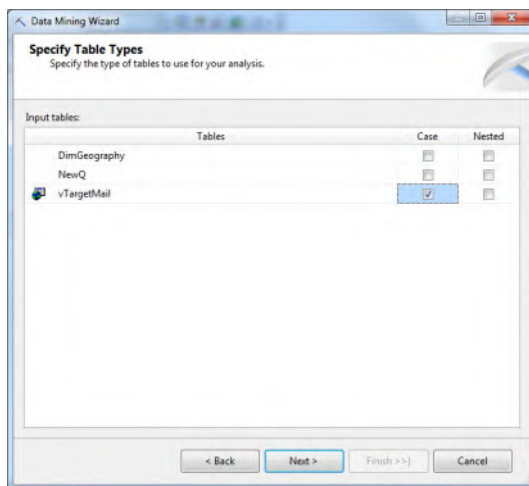


Рис. 11.5 – Діалогове вікно вибору таблиці варіантів

Далі потрібно вказати ключовий стовпець, вхідні стовпці і стовпець, значення якого буде проорокувати. Ключовий стовпець VI Dev Studio визначила автоматично, це первинний ключ таблиці **Customer Key** (рис. 11.6).

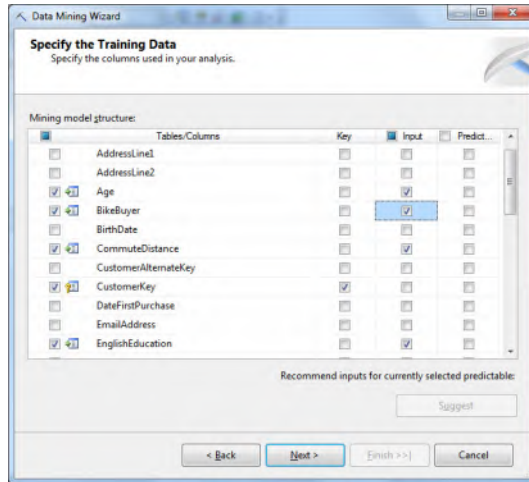


Рис. 11.6 – Діалогове вікно вибору входних і прогнозних атрибутів

У разі використання вкладених таблиць, вказівкою ключів треба приділити особливу увагу. Для завдання кластеризації вказівку пророкує (**Predictable**) атрибута не потрібно. А в якості входних атрибутів будемо використовувати:

- Age – вік клієнта;
- BikeBuyer – ознака того, чи клієнт придбав велосипед;
- CommuteDistance – відстань до роботи або інших “регулярних” поїздки;
- EnglishEducation – освіта;
- EnglishOccupation – посада або рід заняття;
- Gender – стать;
- NumberCarsOwned – число автомобілей у власності;
- NumberChildrenAtHome – число дітей, що потребують догляду;
- Region – регіон проживання;
- TotalChildren – загальне число дітей;
- YearlyIncome – річний дохід.

Наступний крок – уточнення типів даних. На рис. 11.7 представлені вихідні значення.

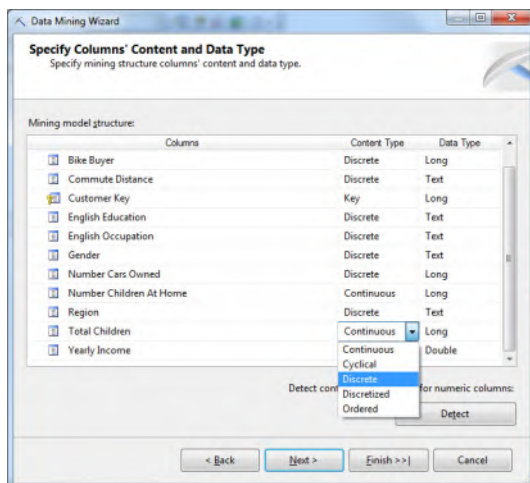


Рис. 11.7 – Діалогове вікно уточнення типів даних

Не всі вони можуть бути визнані вдалими. Наприклад, тип вмісту (**Content Type**) атрибута **Total Children** встановлений як “безперервний” (**Continuous**), тоді як більш коректно було б вибрати тип “дискретний”, тому що кількість дітей в сім’ї буде належати обмеженому безлічі значень {0,1,2, ..}. Можна змінити тип даних і тип вмісту вручну, вибираючи потрібне значення зі списку. А можна скористатися кнопкою **Detect** для автоматичного визначення. Правда в нашому випадку тип вмісту для атрибутів **Total Children** і **Number Children At Home** доведеться міняти все одно вручну.

Після уточнення типів даних буде запропоновано зарезервувати частину даних для цілей тестування. В принципі, для вирішення завдання кластеризації тестове безліч створювати не потрібно. Тому краще тут тестовий набір не створювати і в поле **Percentage of data for testing** вказати значення 0% (рис. 11.8).

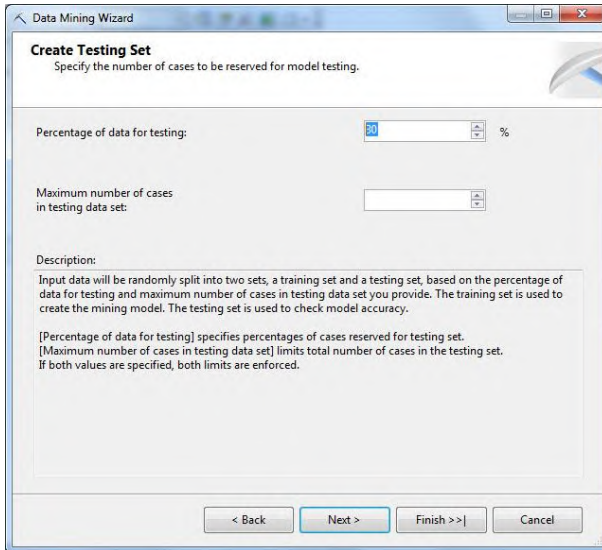


Рис. 11.8 – Діалогове вікно резервування даних для цілей тестування

Останнє вікно майстра дозволяє вказати назви для структури і моделі інтелектуального аналізу. Пропоновані за умовчанням назви краще відредагувати, так щоб вони явно вказували, що це за об'єкт. Наприклад, створену структуру назовемо **vTargetMail_structure**, а засновану на алгоритмі кластеризації модель – **vTargetMail_Cl**. Встановимо також прапорець **Allow drill through**, що дозволить проводити деталізацію, наприклад, переглядати окремі записи, що відносяться до того чи іншого кластеру (рис. 11.9).

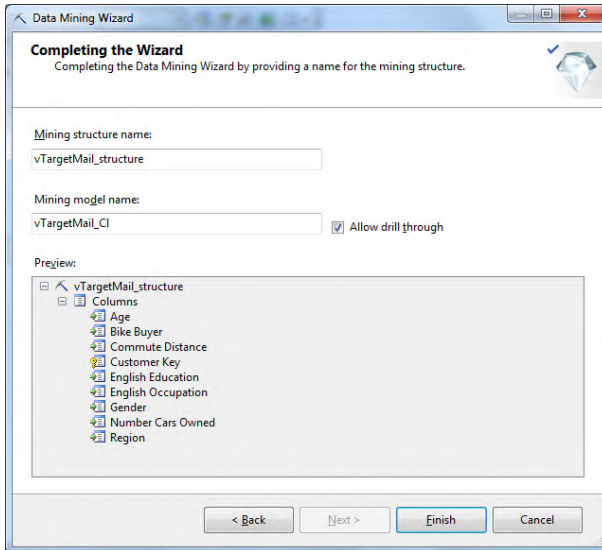


Рис. 11.9 – Діалогове вікно вказівкою імені структури і моделі інтелектуального аналізу

Після чого, тому що ми працює в режимі *immediate*, на сервері інтелектуального аналізу будуть створені структура і модель. А в середовищі BI Dev Studio буде відкрито вікно редактора, що дозволяє проводити слушні роботу з створеними об'єктами. Зокрема, з контекстного меню на вкладці **Mining Structure** можна запустити обробку структури і всіх моделей (рис. 11.10).

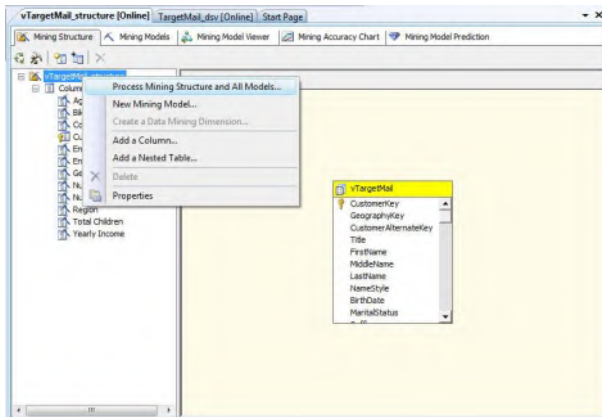


Рис. 11.10 – Діалогове вікно редактора, запуск обробки структури даних

В процесі обробки дані буде завантажено в структуру і станеться навчання моделей. Можна вибрати тип обробки об'єкта (рис. 11.11).

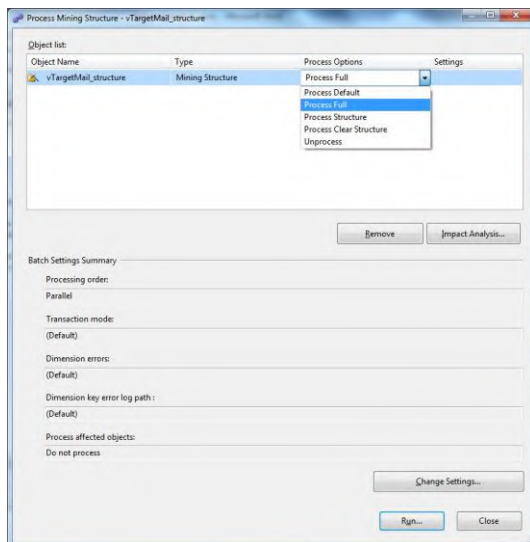


Рис. 11.11 – Діалогове вікно налаштування під час обробки структури

Нижче наведено опис типів обробки об'єкта:

- Process Full (повна обробка) – об'єкт повністю обробляється, для випадку структури відбувається обробка структури і всіх її моделей;

- Process Default (обробка за замовчуванням) – сервер виконує дії, необхідні для приведення даного об'єкта в оброблене стан. Наприклад, якщо ProcessDefault виконується для окремої моделі, буде оброблена саме ця модель (обробка інших моделей, що відносяться до структури проводитися не буде);

- Process Structure (обробка структури) – може виконуватись тільки для структури; водночас, читаються і кешуються дані, обробка моделей не проводиться;

- Process Clear Structure (очищення структури) – використання цієї операції щодо структури призведе до того, що кеш структури буде

очищений від вихідних даних, але містяться в ній моделі збережуться в обробленому стані;

– Unprocess (скасування обробки) – переводить об’єкт в необроблене стан; в случаеструктури, з кешей будуть видалені дані і що містяться в структурі моделі будуть переведені в необроблене стан.

Запускаємо повну обробку для створеної структури з настройками за замовчуванням. І отримуємо повідомлення про помилку, що в джерелі даних вказано значення **Impersonation Mode**, яка не підтримується для операцій обробки (рис. 11.12).

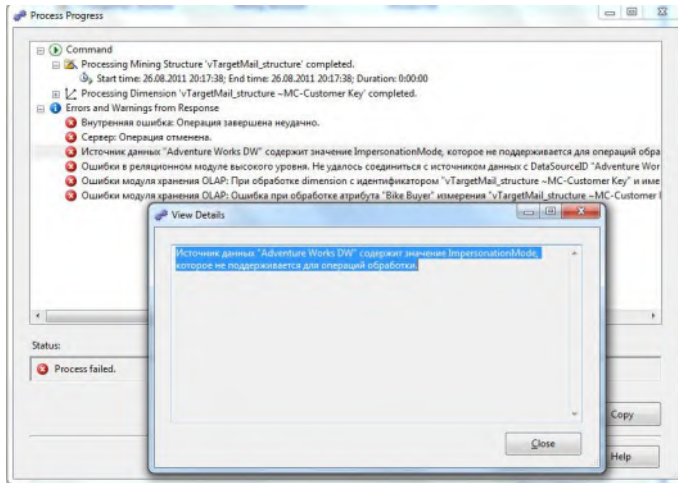


Рис. 11.12 – Повідомлення про помилку під час спроби обробити структуру

В “Початок роботи в BI Dev Studio”, присвяченій створенню джерела даних, зазначалося, що до налаштування **Impersonation Mode** ми ще повернемося. Зробимо це зараз. Закриємо вікна з повідомленнями про помилку та вікна майстра обробки. У вікні **Solution Explorer** знайдемо використовуваний джерело даних і відкриємо його в редакторі (рис. 11.13).

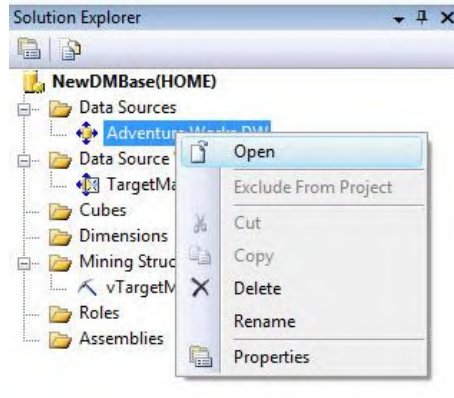


Рис. 11.13 – Діалогове вікно відкриття джерел даних в редакторі

На вкладці **Impersonation Information** замість встановленого раніше **Use the credentials of the current user** вибираємо варіант **Inherit**. Як зазначається в довідці. У цьому варіанті буде використовуватися різні облікові записи (користувача, служби) в залежності від виконуваної операції (рис. 11.14).

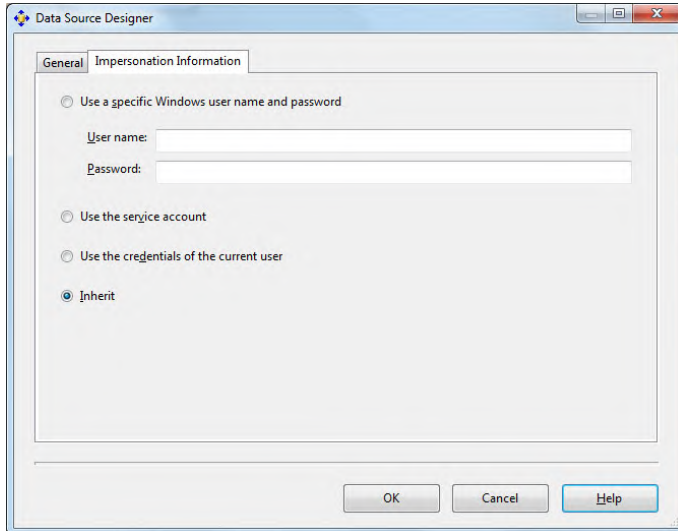


Рис. 11.14 – Діалогове вікно зміни налаштувань Impersonation Information

Після зміни настройки, повторно запускаємо повну обробку структури, яка зараз повинна завершитися успішно (рис. 11.15).

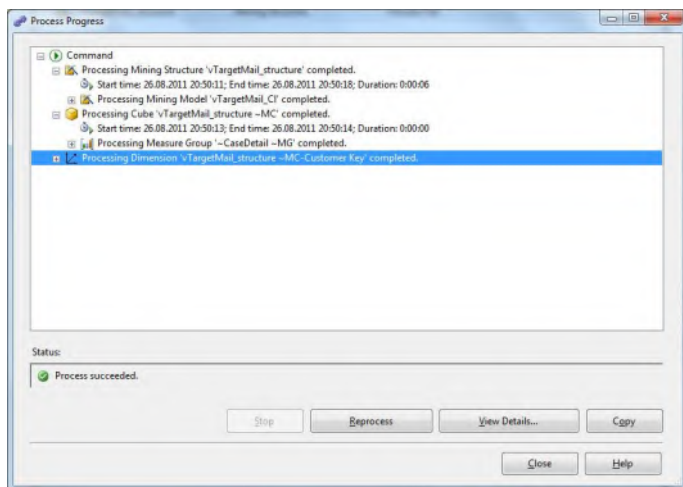


Рис. 11.15 – Результат успішної обробки структури

Після обробки можна відкрити в редакторі структуру і на вкладці **Mining Model Viewer** ознайомитися з побудованою моделлю (рис. 11.16).

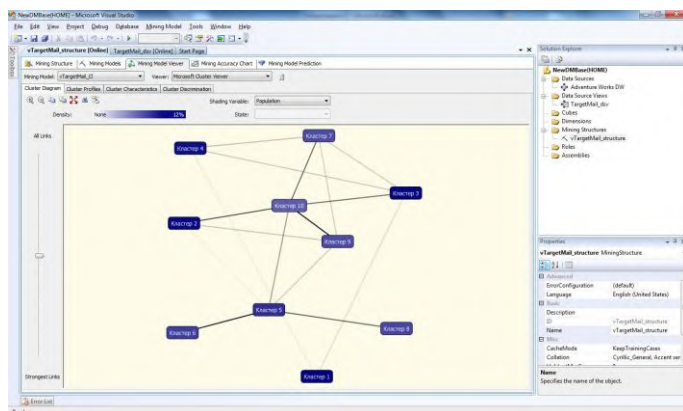


Рис. 11.16 – Перегляд моделі – діаграма кластерів

Інструмент **Cluster Profiles** дозволяє побачити характеристики виявлених кластерів (рис. 11.17).

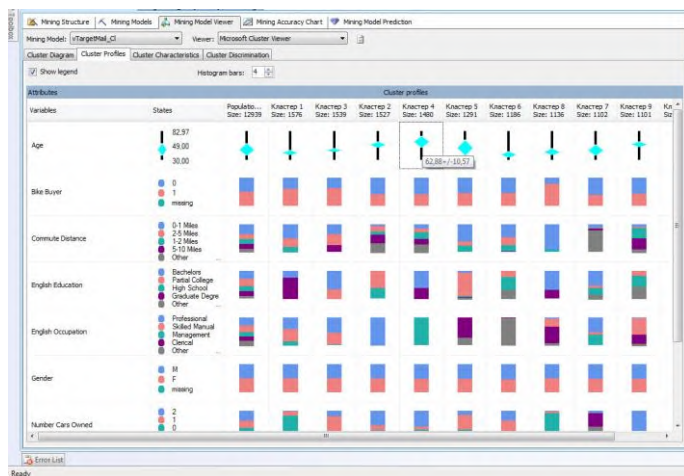


Рис. 11.17 – Перегляд характеристик виявлених кластерів

Наприклад, кластер 4 об'єднує клієнтів старшого віку (середній вік близько 63 років), які працюють в сфері управління (English Occupation = 'Management').

Тут же можна перейменувати кластери, провести деталізацію (опція **Drill Through** в контекстному меню), щоб побачити записи, що відносяться до кожного кластеру.

Отримання номера кластера

Тепер розглянемо, як можна отримати список клієнтів з ідентифікаторами присвоєних їм кластерів. Для цього можна використовувати конструкцію **Prediction Join** і функцію **Cluster()**.

У середовищі BI Dev Studio в вікні перегляду моделі перейдемо на вкладку **Mining Model Prediction**. У вікні **Select Input Tables** натиснемо кнопку **Select Case Table** і вкажемо звідки брати варіанти: уявлення джерела даних **TargetMail_dsv** і в ньому **vTargetMail**. Після цього треба вибрати список відображуваних в результатах запиту атрибутів. Нехай це буде ключ клієнта (CustomerKey), його ім'я, по батькові або друге ім'я і прізвище (FirstName, MiddleName, LastName) і регіон проживання (Region). Також нам потрібно отримати значення функції **Cluster()** (рис. 11.18).

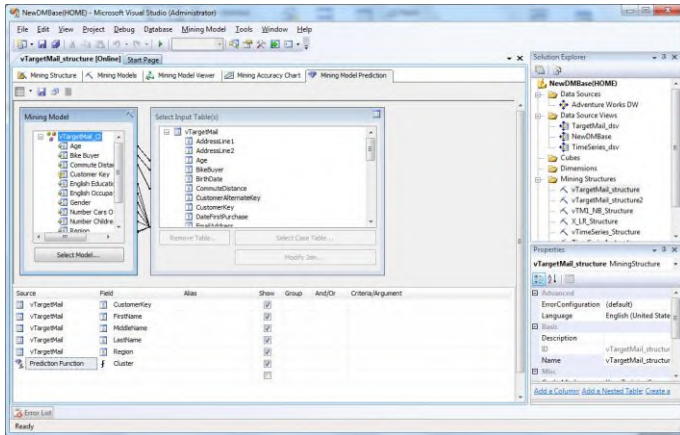


Рис. 11.18 – Конструктор прогнозує запитів

Коли запит сформований в конструкторі можна переключитися до подання результатів (кнопка з зображенням таблиці на панелі інструментів вікна **Mining Model Prediction**). Результат виконання прогнозує запити наведено на рисунку 11.19, а його код представлений нижче.

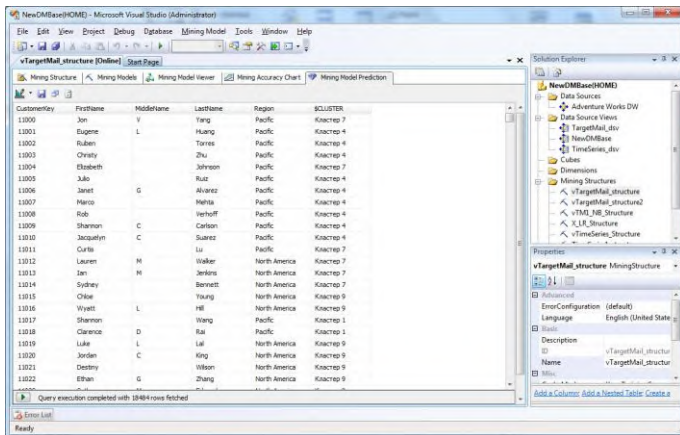


Рис. 11.19 – Результат виконання запити

```
SELECT
t.
t. [FirstName],
```

```

t. [MiddleName],
t.
t. [Region],
Кластер ()
Від
[vTargetMail_Cl]
ПЕРЕДБАЧЕННЯ
OPENQUERY ([Adventure Works DW],
'SELECT
    [CustomerKey],
    [Ім'я],
    [MiddleName],
    [Прізвище],
    [Region],
    [Стать],
    [Річний прибуток],
    [TotalChildren],
    [NumberChildrenAtHome],
    [EnglishEducation],
    [EnglishOccupation],
    [NumberCarsOwned],
    [CommuteDistance],
    [Bik],
    [BikeBuyer]
ВІД
[dbo]. [vTargetMail]
') AS t
ON
[vTargetMail_Cl]. [Гендер] = t
[vTargetMail_Cl]. [Річний прибуток] = t.
[YearlyIncome] AND
[vTargetMail_Cl]. [Всього дітей] = t.
[TotalChildren] I
[vTargetMail_Cl]. [Кількість дітей на дому]
= t.
[vTargetMail_Cl]. [Англійська освіта] = t
[vTargetMail_Cl]. [Англійська окупація] = t
[vTargetMail_Cl]. [Кількість власних
автомобілів] = t
[vTargetMail_Cl]. [Відстань до комунації] =
t
[vTargetMail_Cl]. [Region] = t. [Region] AND

```

[vTargetMail_Cl]. [Bik] = t
 [vTargetMail_Cl]. [Bike Buyer] = t.

Самостійна робота №13. Завдання класифікації.
 Створення структури і моделей інтелектуального
 аналізу. Порівняння точності моделей

Нехай, використовуючи наявні дані компанії **Adventure Works**, необхідно визначити, чи купить новий клієнт велосипед чи ні. Це приклад завдання класифікації, яку можна вирішити за допомогою різних алгоритмів: спрощеного алгоритму Баєса, нейронних мереж, дерев рішень. Розглянемо, як для однієї структури можна створити кілька моделей і оцінити якість формованого ними прогноз.

За аналогією із завданнями попередньої лабораторної створимо структуру інтелектуального аналізу і модель, яка використовує спрощеного алгоритму Баєса. Назвемо структуру – **vTargetMail_structure2**, модель на основі алгоритму **Naive Bayes** (спрощений алгоритм Баєса) – **vTargetMail_NB**. Дані будемо брати, як і раніше, з уявлення **vTargetMail**, використовуючи створені в попередніх лабораторних джерело даних (DataSource) і уявлення джерела даних (Data Source View).

При визначенні структури і моделі в переліку стовпців відзначимо ключовий атрибут – Customer Key і пророкує (Predictable) значення – Bike Buyer (1 – ознака того, що клієнт купив велосипед; 0 – не купив). Щоб визначити, які атрибути надають на нього вплив, скористаємося кнопкою **Suggest** і відзначимо пропоновані стовпці, в найбільшій мірою впливають на цільовий (рис. 11.20).

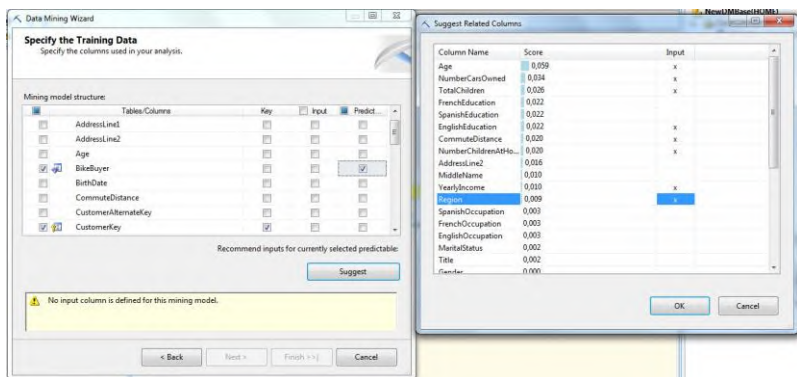


Рис. 11.20 – Вибір стовпців відповідно до рекомендації Data Mining Wizard

Наші знання про предметну область підказують, що адреса і ім'я варто виключити, а включивши в список **English Education**, можна виключити **French Education** і **Spanish Education**, тому що це те ж саме, тільки на іншій мові. Таким чином, в якості вхідних (Input) атрибутів будемо використовувати:

- Age – вік клієнта;
- NumberCarsOwned – кількість автомобілей у власності;
- TotalChildren – загальна кількість дітей;
- EnglishEducation – освіта;
- CommuteDistance – відстань до роботи “регулярних” поїздов);
- NumberChildrenAtHome – кількість дітей, що потребують догляду;
- YearlyIncome – річний дохід;
- Region – регіон проживання.

На рисунку 11.21 представлені типи даних, автоматично встановлені для стовпців.

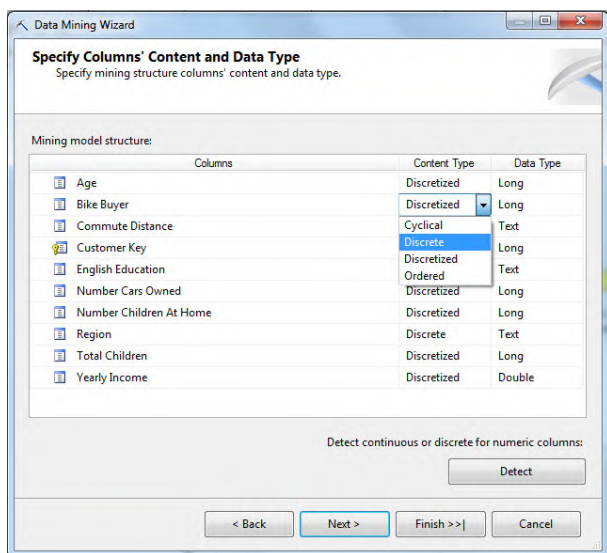


Рис. 11.21 – Діалогове вікно уточнення типів вмісту

Зверніть увагу, що для деяких стовпців встановлений тип вмісту **Discretized**. Пов'язано це з тим, що обраний для моделі спрощений алгоритм Баєса не працює з числовими атрибутами з типом вмісту **Continuous** і для цих стовпців буде проведена дискретизація (розбиття на інтервали) значень. Такий тип вмісту коректний в разі атрибутів Age і YearlyIncome. Але, NumberCarsOwned, NumberChildrenAtHome, TotalChildren треба змінити тип на Discrete – ми знаємо, що безліч можливих значень цих атрибутів невелика.

Для цілей тестування відповідно до установкою за замовчуванням резервуємо 30% записів.

Таким чином, ми створили структуру інтелектуального аналізу даних. Якщо відкрити структуру в редакторі і перейти на вкладку **Mining Models** можна впевнитися, що створена і модель (рис. 11.22).

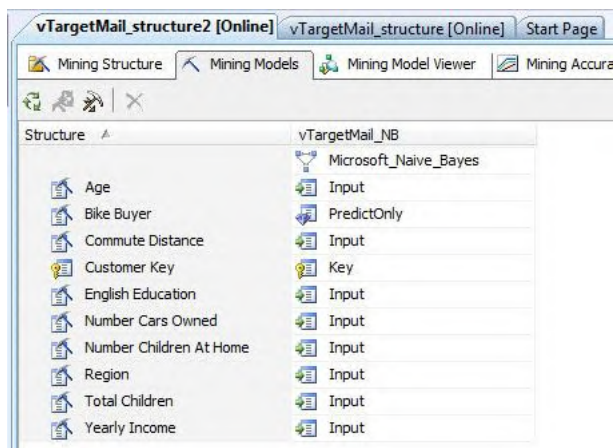


Рис. 11.22 – Створена структура, що включає одну модель

Зверніть увагу, що атрибут BikeBuyer відзначений як що передбачається (PredictOnly). Список, що розкривається навпроти назви стовпчика дозволяє змінити це налаштування. Якщо якийсь атрибут треба виключити з розгляду під час навчання моделі, змініть значення Input на Ignore (ігнорувати).

Скориставшись контекстним меню (рис. 11.23) можна отримати доступ до параметрів використовуваного моделлю алгоритму.

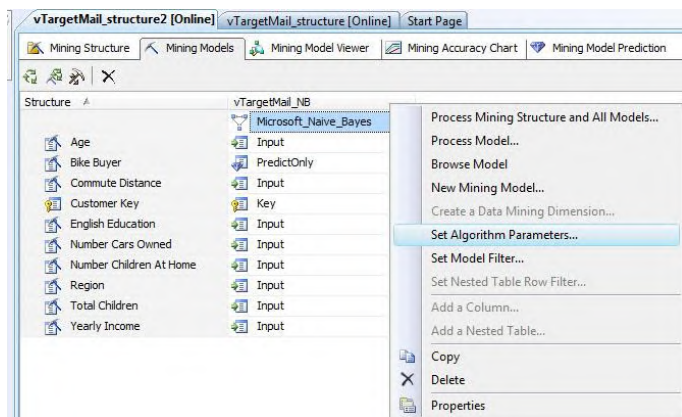


Рис. 11.23 – Контекстне меню дозволяє переглянути і змінити параметри алгоритму

Також можна налаштувати фільтр (пункт **Set Model Filter ...**), тоді для аналізу будуть використовуватися тільки варіанти, що відповідають умовам фільтрації (рис. 11.24).

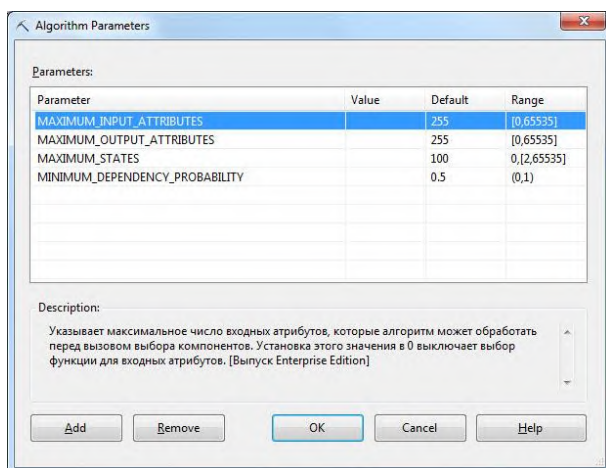


Рис. 11.24 – Параметри спрощеного алгоритму Баєса для створеної моделі

Наприклад, це дозволяє виключити з розгляду клієнтів з якогось регіону, який, як ми вважаємо, істотно відрізняється від інших.

Скориставшись пунктом **NewMiningModel ...** контекстного меню (рис. 11.23) можна створювати аналогічні по набору атрибутів моделі, засновані на інших алгоритмах інтелектуального аналізу.

Завдання 1. Створіть структуру і модель інтелектуального аналізу, аналогічно тому, як це було описано вище.

Завдання 2. У тій же структурі створіть дві додаткові моделі, одна з яких (назвемо її **vTargetMail_DT**) буде використовувати алгоритм **Microsoft Decision Trees**, інша (**vTargetMail_NN**) – **Microsoft Neural Network**. У підсумку, повинен вийти результат, представлений на рисунку 11.25.

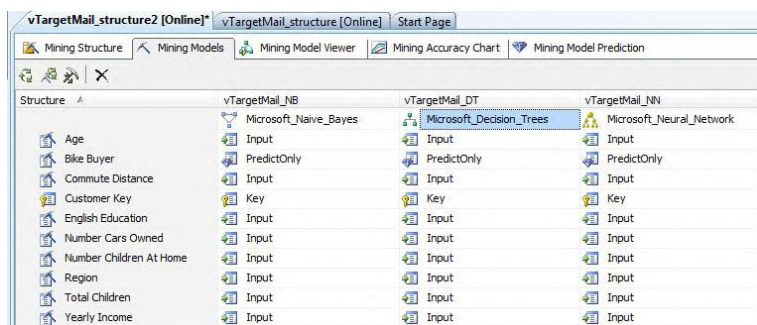


Рис. 11.25 – Три моделі в одній структурі

Завдання 3. Виконайте обробку структури і всіх її моделей (повну обробку структури).

Після того, як структура і моделі створені і оброблені, хотілося б з'ясувати, яка з трьох моделей дає більш точний прогноз. Для цього можна використовувати діаграми точності (**Mining Accuracy Chart**) і зарезервоване в ході створення структури безліч варіантів для тестування (ці дані не використовувалися під час навчання моделі).

Перейдемо на вкладку **Mining Accuracy Chart** (рис. 11.26).

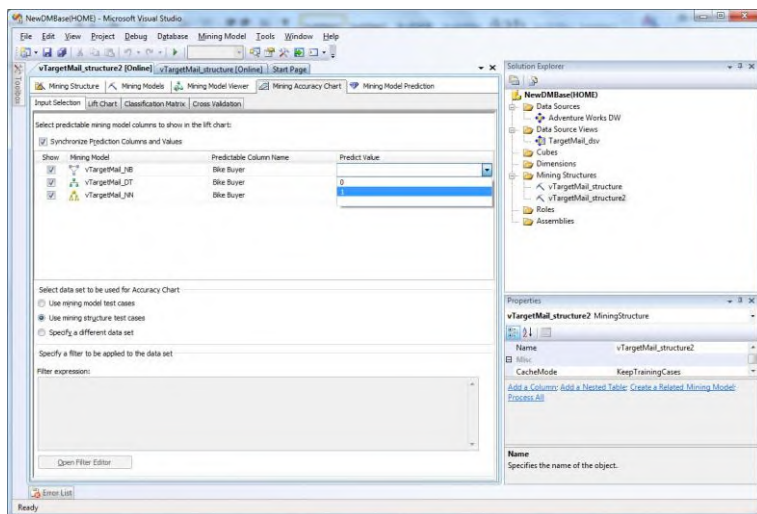


Рис. 11.26 – Встановлені для оцінки точності моделі

Там можна відзначити, для яких моделей будуть будуватися діаграми, і які дані будуть використовуватися в процесі тестування. Можна використовувати набір даних, зарезервований в моделі або в структурі, а також зовнішній набір даних. Перші два варіанти будуть відрізнятися, якщо під час створення моделі задавався фільтр варіантів. Перевірки набір в моделі буде включати тільки варіанти, відповідні фільтру, а перевірки набір в структурі фільтр не враховує. У нашому випадку фільтрація не використовувалася, так що ці варіанти рівнозначні.

Крім того, для дискретного цільового атрибута можна вибрати пророкує значення (Predict Value, рис. 11.26). У розглянутому прикладі нам цікавіша 1, тобто клієнт, який зробить покупку.

Вибираємо тестовий набір для структури, PredictValue = 1, і переходимо на вкладку **Lift Chart**. Стандартна діаграма точності, звана діаграмою зростання буде виглядати, як представлено на рисунку 11.27.

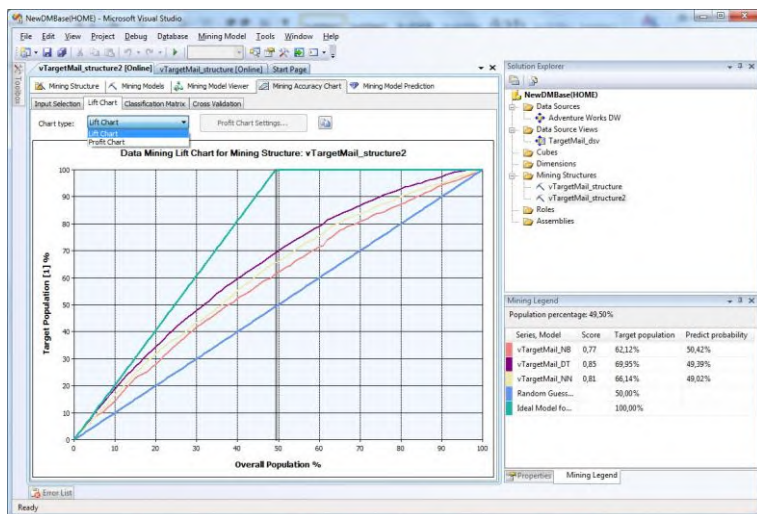


Рис. 11.27 – Діаграма зростання

Зелена лінія зображує ідеальну модель, синя – випадковий вибір. У нас в тестовому наборі приблизно 50% варіантів мають значення BikeBuycr рівне 1. Можна уявити, що кожна модель відповідно до свого прогнозу формує упорядкований за ступенем близькості до згаданого значення список варіантів. Так ось, у ідеальній моделі шукані варіанти будуть в перших 50% позицій списку. У випадкової моделі – в першій половині списку буде тільки 50% клієнтів, які зробили покупку. Чим ближче результат до ідеальної моделі, тим точніше прогноз. У нашому прикладі найкращий результат дає модель, яка використовує алгоритм дерев прийняття рішень.

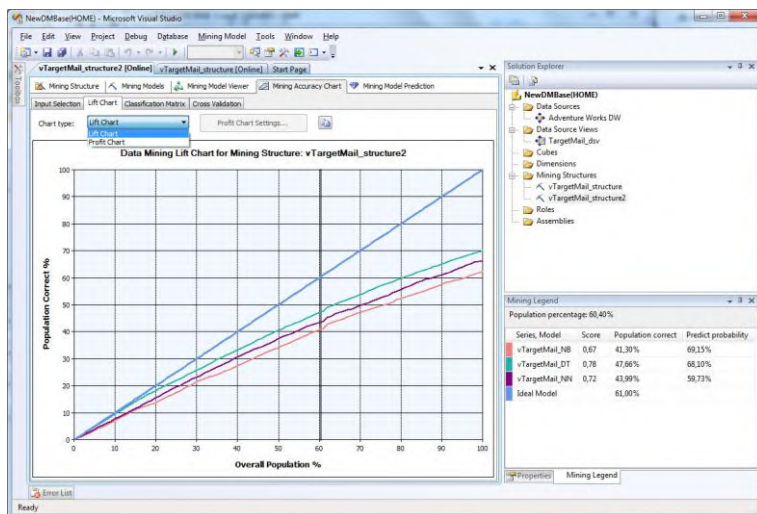


Рис. 11.28 – Діаграма зростання, коли цільове значення не вибрано

Якщо не вказувати цільове значення (тобто на рисунку 11.26 не ставити PredictValue = 1), то діаграма точності буде виглядати, як представлено на рисунку 11.28. На ній теж видно, що модель **vTargetMail_DT** дає більш точний прогноз.

Щоб отримати вартісну оцінку якості моделі, можна використовувати діаграму зростання прибутку (на рисунку 11.27 – **Profit Chart**). BI Dev Studio запросить дані про загальну кількість варіантів – Population (наприклад, це число клієнтів, яким збираємося провести розсилку), обмеження на сумарну вартість – FixedCost (наприклад бюджет рекламної розсилки), витратах на одиницю – IndividualCost (наприклад вартість відправки одного листа з рекламною пропозицією) , виручці від одного покупця – RevenueperIndividual (рис. 11.29).

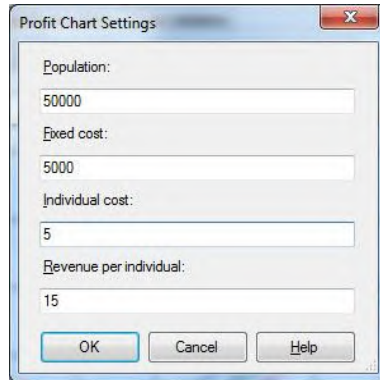


Рис. 11.29 – Діалогове вікно налаштування для побудови діаграми росту прибутку

Отримана діаграма дозволяє зрозуміти яке число пропозицій треба розіслати для отримання максирисньо прибутку. Будемо вважати, що на основі прогнозів кожної моделі сформований список клієнтів, упорядкований по спадаючій прогнозованій ймовірності покупки клієнтом велосипеда. На рисунку 11.30 максимум прибутку буде досягнуто, коли під час використання моделі **vTargetMail_DT**, пропозиції будуть відправлені приблизно 64% клієнтів, починаючи з верхньої частини списку.

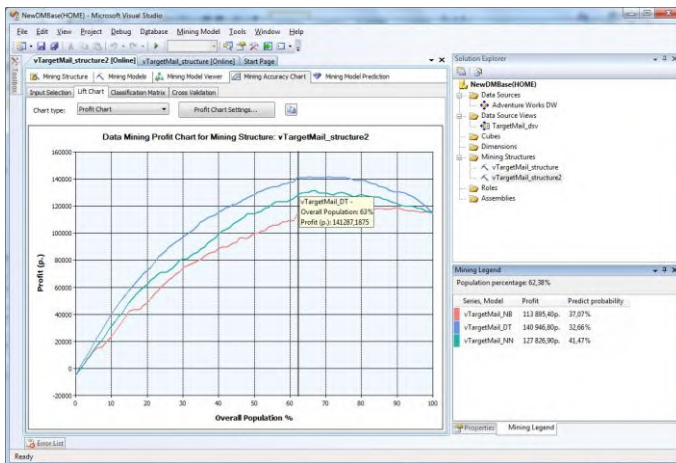


Рис. 11.30 – Діаграма зростання прибутку

Завдання 4. За аналогією з описом, проведіть аналіз точності отриманих моделей. Прокоментуйте результати.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Які задачі можна вирішити при інтелектуальному аналізі даних за допомогою оператора SELECT?
2. Для чого використовується інструкція FLATTENED?
3. Що дозволяють зробити інструкції ORDER BY і TOP <n>?
4. Що дозволяє зробити інструкція WHERE?
5. Для чого використовується інструкція FROM?
6. Приведіть синтаксис оператора, що дозволяє отримати значення зазначених стовпців моделі?
7. Наведіть синтаксис оператора для виводу вмісту моделі
8. Наведіть синтаксис оператора SELECT, що використовується для прогнозного з'єднання
9. Що таке Одноелементний запит?
10. В якому випадку застосовується запит з використанням функції OPENQUERY?

ЛІТЕРАТУРА

1. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.

БІБЛІОГРАФІЧНИЙ СПИСОК

1. Witten I.H. Data Mining: Practical Machine Learning Tools and Techniques / I.H. Witten, E.Frank. – Morgan Kaufmann, 2005. – 558 p.
2. Чубукова И.А. Data Mining / И.А. Чубукова. – М.: Интернет-университет информационных технологий, Бином. Лаборатория знаний, 2008 – 382 с.
3. Lohr S. Data-IsM: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else / S. Lohr. – HarperBusiness, 2015. – 489 p.
4. Data Mining History and Current Advances [Електронний ресурс]. – 2017. – Режим доступа до ресурсу: https://www.sas.com/en_us/insights/analytics/data-mining.html.
5. Бармак О.В. Интеллектуальный анализ данных. Методичні вказівки до самостійного вивчення курсу «Интеллектуальный анализ данных» для студентов напрямую подготовки 0804 «комп'ютерні науки» спеціальності 7.080402 – «Інформаційні технології проектування» / О.В.Бармак. – Хмельницький: ХНУ, 2012. – 93 ст.
6. Макленнен Дж. Microsoft SQL Server 2012 : Data Mining – интеллектуальный анализ данных: [пер. с англ.] / Дж. Макленнен, Чжаохуэй Танг, Б. Криват. – СПб.: БХВ-Петербург, 2009. – 720 с.
7. Службы SQL Server Analysis Services – интеллектуальный анализ данных. – [Електронний ресурс]. – Режим доступа: <https://docs.Microsoft.com/ru-ru/sql/analysis-services/analysis-services-tutorials-ssas?view=sql-server-2017>
8. Барсегян А.А. Анализ данных и процессов : учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д.Тесс. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
9. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: «БХВ-Петербург», 2004. – 336 с.
10. Baker L.D. Distributional Clustering of Words for Text Classification In ACM SIGIR 98 / L.D. Baker, A.K. McCallum. – 1998.
11. Воронцов К.В. Машинное обучение: курс лекций [Електронний ресурс] / К.В. Воронцов. – Режим доступа: [http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_\(курс_лекций,_К.В.Воронцов](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов)
12. Башмаков А.И. Интеллектуальные информационные технологии: учеб. пособ. / А.И. Башмаков, И.А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

13. Data Mining and Image Processing Toolkits. – [Електронний ресурс]. – Режим доступу: <http://Data Mining.itsc.uah.edu/adam/>
14. Дюк В. Data Mining: учебный курс / В. Дюк, А. Самойленко. – СПб: Изд. Питер, 2001. – 368 с.
15. Palmer M. Data is the New Oil / M. Palmer // CMO News, Michael Palmer, Reinventing Marketing, Technology's Impact, The Customer. – 2006. – Електронний ресурс. – Режим доступу: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html
16. Акіменко В.В. Проектування СППР на основі нечіткої логіки : навчально-методичний посібник / В.В. Акіменко, Ю.В. Загородній. – К.: Вид-во КНУ, 2007. – 94 с.
17. Дюк В.А. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях / В.А. Дюк, А.В. Флегонтов, И.К. Фомина // Известия Российского государственного педагогического университета им. А.И. Герцена. Область наук: Информатика. – 2011. – С. 77–84.
18. Нікольський Ю. Дерева прийняття рішень та їхнє застосування для прогнозування діагнозу у медицині / Ю. Нікольський, Ю. Щербина, Р. Якимечко // Вісник Львівського університету. Серія прикладна математика та інформатика. – № 6, 2003. – С. 191–211.
19. Кречетов Н. Продукты для интеллектуального анализа данных / Н. Кречетов // Рынок программных средств, N14-15_97. – 1997. – С. 32–39.
20. Службы Analysis Services – видеоролики по интеллектуальному анализу данных (SQL Server 2012). URL: [http://Microsoftdn.Microsof.com/ru-ru/library/dd776389\(v=SQL.100\).aspx](http://Microsoftdn.Microsof.com/ru-ru/library/dd776389(v=SQL.100).aspx)
21. Киселев М. Средства добычи знаний в бизнесе и финансах / М. Киселев, Е. Соломатин // Открытые системы. – 1997. – № 4. – С. 41–44.
22. Хайкин С. Нейронные сети : полный курс, 2-е издание : пер. с англ / С.Хайкин. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
23. Любунь З.М. Основи теорії нейромереж / З.М. Любунь // Текст лекцій. – Львів: Видавничий центр ЛНУ імені Івана Франка, 2007. – 142 с.
24. DMX Tutorials (Analysis Services – Data Mining) [Електронний ресурс] // SQL Docs. – 2018. – Режим доступу до ресурсу: <https://docs.microsoft.com/en-us/sql/dmx/dmx-tutorials-analysis-services-data-mining?view=sql-server-2017>.

25.Data Mining Extensions (DMX) Reference [Электронный ресурс] // SQL Docs. – 2018. – Режим доступа: <https://docs.microsoft.com/en-us/sql/dmx/data-mining-extensions-dmx-reference?view=sql-server-2017>.

У зв'язку з удосконаленням технологій запису та зберігання даних людство отримало колосальні об'єми даних у різних сферах. Діяльність будь-якого підприємства (комерційного, виробничого, медичного, наукового тощо) тепер супроводжується реєстрацією та записом усіх подробиць його діяльності. Став очевидним той факт, що без попередньої обробки потоки сирих даних є малоцінними. Специфіка сучасних вимог до такої обробки є такою: дані мають необмежений обсяг; дані є різномірними (кількісними, якісними, текстовими); результати мусять бути конкретними та зрозумілими; інструменти для обробки сирих даних мають бути прості у використанні. Тому для ефектичної роботи з даними застосовують методи інтелектуального аналізу даних.

В основу інтелектуального аналізу даних концепція шаблонів, що відбивають фрагменти багатоаспектних взаємовідносин у даних. Ці шаблони є закономірностями, яким властиві підвибірки даних. Пошук шаблонів проводиться методами, не обмеженими апріорними припущеннями про структуру вибірки, та типами розподілів значень аналізованих показників. Важливе положення інтелектуального аналізу даних – нетривіальність розшукуваних шаблонів.

Завдяки використанню методичних вказівок до самостійної роботи навчальної дисципліни «Методи інтелектуального аналізу даних в інформаційних технологіях» студенти опанували базові методи сучасної обробки даних, навчилися використовувати програмні продукти й різні інструментальні засоби. У рамках навчальної дисципліни студенти виконували самостійну роботу, що полягала в поглибленні, узагальненні й закріпленні теоретичних знань і практичних умінь. У результаті проходження навчальної дисципліни «Методи інтелектуального аналізу даних в інформаційних технологіях» студенти готові до самостійної роботи з вирішення завдань засобами інтелектуального аналізу даних і розробки інтелектуальних систем.

Отриманні знання та навички будуть використовуватися й доповнюватися студентами в процесі вивчення дисциплін «Теорія прийняття рішень» та «Методи та системи штучного інтелекту».

ЗМІСТ

ВСТУП	4
РОЗДІЛ 1. ОСНОВНІ ПОНЯТТЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ	6
1.1 Поняття інтелектуального аналізу даних	6
1.2 Поняття даних	8
Самостійна робота №1. Надбудова інтелектуального аналізу даних для Microsoft Office	11
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	21
ЛІТЕРАТУРА.....	22
РОЗДІЛ 2. ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В СКБД MICROSOFT SQL SERVER 2012.....	23
2.1 Задачі інтелектуального аналізу даних	23
2.2 Інтелектуальний аналіз даних засобами СКБД Microsoft SQL Server 2012	27
Самостійна робота №2. Використання інструментів «Analyze Key Influencers» та «Detect Categories»	29
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	37
ЛІТЕРАТУРА.....	38
РОЗДІЛ 3. ОСНОВНІ ЕТАПИ ПРОВЕДЕННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ	40
Самостійна робота № 3. Використання інструментів «Fill From Example» та «Forecast»	44
Самостійна робота № 4. Використання інструментів аналізу «Highlight Exceptions» та «Scenario Analysis»	51
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	62
ЛІТЕРАТУРА.....	62
РОЗДІЛ 4. МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: СПРОЩЕНИЙ АЛГОРИТМ БАЕСА ТА ДЕРЕВА РІШЕНЬ.....	64
4.1 Спрощений алгоритм Баєса	64
4.2 Алгоритм дерева рішень	66
Самостійна робота №5. Використання інструментів Data Mining Client для Excel 2007 для підготовки даних	70
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	80
ЛІТЕРАТУРА.....	80

РОЗДІЛ 5. МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: ЛІНІЙНА РЕГРЕСІЯ ТА АНАЛІЗ ЧАСОВИХ РЯДІВ	81
5.1 Метод лінійної регресії	81
5.2 Методи аналізу часових рядів	83
Самостійна робота №6. Використання інструментів DMC Excel 2007 для створення моделі інтелектуального аналізу даних	88
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	95
ЛІТЕРАТУРА.....	96
РОЗДІЛ 6. МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: МЕТОДИ КЛАСТЕРИЗАЦІЇ, АЛГОРИТМ ВЗАЄМОЗВ'ЯЗКІВ ТА КЛАСТЕРИЗАЦІЯ ПОСЛІДОВНОСТЕЙ.....	97
6.1 Методи кластеризації	97
6.2 Алгоритм взаємозв'язків.....	100
6.3 Алгоритм кластеризації послідовностей	101
Самостійна робота №7. Побудова моделі кластеризації, трасування і перехресна перевірка.....	103
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	112
ЛІТЕРАТУРА.....	113
РОЗДІЛ 7. МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: ШТУЧНІ НЕЙРОННІ МЕРЕЖІ ТА ЛОГІСТИЧНА РЕГРЕСІЯ	114
7.1 Штучні нейронні мережі	114
7.2 Метод логістичної регресії.....	117
Самостійна робота №8. Аналіз точності прогнозу і використання моделі інтелектуального аналізу.....	117
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	124
ЛІТЕРАТУРА.....	125
РОЗДІЛ 8. КОНЦЕПЦІЇ МОВИ DATA MINING EXTENSIONS	126
8.1 Базові поняття мови Data Mining Extensions	126
8.2 Мова DMX: створення структури	130
8.3 Мова DMX: створення моделі	133
Самостійна робота №9. Початок роботи в BI Dev Studio	136
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	144
ЛІТЕРАТУРА.....	145
РОЗДІЛ 9. ЗАПИТИ МОВИ DATA MINING EXTENSIONS	146
9.1 Мова DMX: обробка, очищення, видалення та відновлення структур і моделей.....	146

9.2 Мова DMX: запити	149
9.3 Мова DMX: запити для прогнозу	154
Самостійна робота №10. Створення уявлення джерела даних	158
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	163
ЛІТЕРАТУРА.....	164
РОЗДІЛ 10. ОСОБЛИВОСТІ ВИКОРИСТАННЯ ВБУДОВАНИХ АЛГОРИТМІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В СКБД MICROSOFT SQL SERVER 2012.....	165
10.1 DMX. Параметри алгоритмів інтелектуального аналізу даних: спрощений метод Баєса, алгоритм дерева рішень, метод лінійної регресії	165
10.2 DMX. Параметри алгоритмів інтелектуального аналізу даних: часові ряди, алгоритм кластеризації	168
10.3 DMX. Параметри алгоритмів інтелектуального аналізу даних: алгоритм взаємозв'язків, кластеризація послідовностей	172
10.4 DMX. Параметри алгоритмів інтелектуального аналізу даних: алгоритми нейронних мереж, логістична регресія	174
Самостійна робота №11. Робота з моделями інтелектуального аналізу даних з SQL Server Management Studio.....	178
Самостійна робота №12. Використання алгоритму Microsoft Time Series для прогнозування значень часових рядів	184
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	189
ЛІТЕРАТУРА.....	190
РОЗДІЛ 11. СТВОРЕННЯ СТРУКТУРИ І МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ.....	191
Самостійна робота №13. Завдання класифікації. Створення структури і моделей інтелектуального аналізу. Порівняння точності моделей	206
ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ	215
ЛІТЕРАТУРА.....	215
БІБЛІОГРАФІЧНИЙ СПИСОК.....	216
ПІСЛЯМОВА.....	219