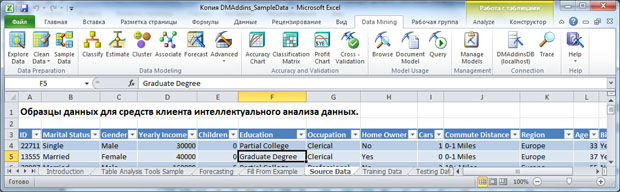
**Практична робота 6**

**Використання інструментів Data Mining Client для Excel 2007 для підготовки даних**

Розглянуті в попередніх лабораторних роботах «Засоби аналізу таблиць для *Excel* » ( **TableAnalysisTools** ) дозволяють швидко провести «стандартний» *аналіз* наявних даних. У той же час цей набір інструментів не надає особливих можливостей для підготовки даних до аналізу, оцінки результатів і т.д. В *Excel* це можна зробити, використовуючи інструмент **DataMiningClient** , який також входить до набору надбудови інтелектуального аналізу. У ході [«Надбудови інтелектуального аналізу даних для](https://www.intuit.ru/studies/courses/2312/612/lecture/13263) MicrosoftOffice[»](https://www.intuit.ru/studies/courses/2312/612/lecture/13263) відзначалося, що бажано зробити повну установку надбудови, в яку входить і **DataMiningClient** .

Відкриємо вже використаний нами набір даних, що входить в поставку надбудови ( *меню* **"Пуск"** , найдіть **Надстройки інтелектуального аналізу даних- >Образцы данных Excel** ). Щоб можна було спокійно вносити зміни, краще зберегти його під новим іменем .П ерейдіть на *лист* **«Ісходні дані» ( SourceData )** і натисніть на закладку **DataMining** . Стрічка з запропонованими інструментами представлена на [рис. 6.1](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.1) .

**Рис. 6.1.** Інструменти Data Майнінг Клієнт



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

Наприклад, у попередніх роботах ми неодноразово зіткнулися з тим, що ряд алгоритмів (MicrosoftNaiveBayes та ін.) вимагають попередньої дискретизації непереривних значень числових параметрів. Але в ряді випадків користувачеві бажано переглянути можливі діапазони уточнити їх число і т.д. Окремий інтерес може представляти і розподіляти рядок за значенням вибраного параметра.

### Explore Data

Інструмент Explore Data дозволяють проаналізувати значення стовпця (або діапазону яулітинок) і відобразити їх на діаграмі. Розглянемо його роботу на прикладі значення річного доходу клієнта ( Income ). Додатковий інтерес полягає в тому, що це значення може розглядатися і як безперервне, і як дискретне. І так, запускаємо інструмент ( [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) ).

|  |  |
| --- | --- |
| 1 | 2 |
|  |  |
| 3 | 4 |
| **Рис. 6.2.** Інструмент Explore Дані |  |
|  |  |

У процесі роботи потрібно вказати, для якої таблиці (або діапазону ячеек) і таблиці буде проведено аналіз підсумку *(* рис [. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -1 і [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -2). Після чого вказані значення будуть проаналізовані і результат представлений у вигляді гістограмми.

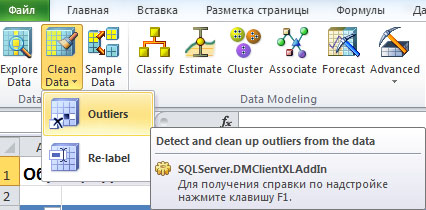
Як уже відзначалося вище, *значення* річного доходу можна розглядати і як постійне, і як дискретне (за рахунок того, що в нашому наборі даних присутні тільки значення, короткі 10 тисячам). Для безперервного значення буде запропоновано варіант розбиття на діапазони ( [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -3). Число діапазонів можна поміняти, а *діаграма* з розподілом значень буде побудована заново. Натискнувши кнопку **" Додати новий Стовпець "** можна додати в вихідну таблицю новий стопчик з інтервалами річного доходу. Наприклад, якщо для стрічки *значення*  **щорічно Дохід = 30000** , то *значення* нового параметра **Річний Дохід 2** при використанні представленого на рисунку розбіжності складе **"'30000 - 50000"** (звичайно так, з апострофом на початку, щоб розглядалося як стрічкове). В ході інтелектуального аналізу, отриманий стовпчик може бути використаний замість вихідного (включення обох стовпців одночасно небажано).

Кнопки з зображеннями графіки та гістограми (на [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -3, [рис. 6.2 -4 вони](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) *підчеркнуті* ), можна вказати тип аналізованого значення - постійне або дискретне. Якщо *значення* річного доходу розглядаємо як дискретне, то для нього буде побудована *діаграма ,* яка показує розподіл числа рядків за значенням річного доходу ( [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -4). При цьому *сортування* проводиться за зменшенням числа рядків зі значенням даних, із-за чого перший стовпчик гістограми відповідає значенню **"60000"** , другий - **"40000"** і т.д. Сформовану гістограму можна скопіювати в *буфер* (кнопка права**" Додати новий Стовпець "** , [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -3, [рис. 6.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -4) і використовувати для подальшої роботи.

### чистий Дані

Інструмент **Clean Дані** ( [рис. 13.3](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.3) ) дозволяють підготувати дані для аналізу, відкинувши нетипові або помилкові дані (викиди), а також навести заміну окремих значень. Як зазначається в документах, під викидом розуміється *значення* даних, що є проблематичним за однією з наступних причин:

* значення знаходиться за межами очікуваного діапазону;
* дані були введені неправильно;
* значення відсутнє;
* дані представляють собою пропуск або пусту строку;
* значення може значно відклонитися від розподілу, якому підчиняються дані в моделі.



**Рис. 13.3.** Інструмент CleanData

Використання даного інструменту проілюстровано на прикладі всіх той же таблиці з даними про клієнтів ( *лист* **Джерело Дані** ). Обратимся до столбцу з віком. Пусть нам потрібно очистити набір даних з інформації про нехарактерних за віком покупців. Запускаем інструмент **Clean Дані -> Викиди** , у вікні аналогічно представленому на [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -1 вибираємо таблицю для аналізу, потім у вікні **Виберіть Колонка** ( [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -2)- столбец **Вік** .

В розглядуваному наборі даних є строки зі значеннями столбця **Вік** від 25 до 96 років. Якщо цей *параметр* вважаємо непереривним, то він буде представлений графіком, де по осі X показано вік, по осі Y-число клієнтів з таким віком. В наборе данных доля клиентов преклонного возраста очень мала. на рис [. 13.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.4) -1 показано, що встановивши порогове *значення* в 75 років, ми відбираємо заштрихований "хвост", що включає нехарактерні значення (покупці велосипедів у віці від 76 до 96 років, які подають меншинство).

У багатьох аналогічно виглядає робота з параметром, що приймає дискретні значення. Для него будується *гістограмма* , а для визначення породи потрібно вказати мінімальне число прикладів, «підтримуючих» *значення* . Наприклад, на [рис. 13.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.4) -2, встановлено порогове *значення* в 15. На жаль, при великому числі столбцов гістограмми , з наведення параметрів на них не відображаються. Тому не поняти, що саме потрапляє в "хвост" розподілу.

|  |  |
| --- | --- |
| 1 | 2 |
|  |  |
| 3 | 4 |
| **Рис. 13.4.** Використання інструменту CleanData для виключення викидів |  |

Так, ми виділили нехарактерні дані. Тепер потрібно визначити, що з ними робити. Пропоновані майстерні рішення відрізняються для постійного і дискретного параметра. Відповідну строку можна видалити ( **Видалити рядки містить outliners** ) або змінити *значення* параметра на пусте ( **Змінити значення до null** ). Крім того, для непереривних даних ( [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -3) можна змінити нехарактерне *значення* середнім або граничним (зверху або внизу, в залежності від того, який діапазон *відбирається* ). Для дискретного параметра ( [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) -4) можна вказати значення (із чисел, які вже є в наборі), на які будуть змінюватися « *виброси* ».

Останнє пропонує вікно майстра (його на рисунку не представлено) вибрати, куда занести зміни - у вихідні дані ( **Змінити даних inplace** ), в їх копію на новому листі *Excel* ( **Копіювати лист даних з зміни до a новий працювати аркуш** ) або в новий стовпець у вихідній таблиці ( **Дод як a новий колонка до в поточний працювати sheet** ) .Остання опція *для* випадку видалення рядка недоступна.

### CleanData.Re-label

У деяких випадках у вихідних даних можуть бути значення, які затрудняють автоматизований *аналіз* . Наприклад, є *параметр* "город" і серед його значень - Санкт-Петербург, С-Петербург, СПб. Для того , щоб у процесі інтелектуального аналізу ці значення враховувалися правильно, надо їх замінити на одне. Для цього можна використовувати інструмент **Re-label** . Його можна змінити, якщо потрібно знизити *рівень деталізації значення* параметра. Слід зазначити, що інструмент працює тільки з дискретними значеннями (ну або розглядаються як дискретні ).

Для прикладу, у таблиці з інформацією про клієнтів нам потрібно зменшити число значень параметра **CommuteDistance** ( *відстань щоденних поїздок* ). Вихідні значення "0-1 *миль* ", "1-2 *милі* ", "2-5 *миль* ", "5-10 *миль* ", "10+ *миль* ". Пусть все, що менше 2 миль, буде "близько", інше - "далеко". Додайте в таблицю дві порожні строки і укажіть для **однієї CommuteDistance** " близько" , а для іншої - "далеко". Делается це тому, що значення, на які замінюємо, теж повинні бути присутніми в столбці.

|  |  |
| --- | --- |
| 1 | 2 |
| **Рис. 13.5.** Заміна символів |  |

Запустимо інструмент: **CleanData -> Re-label** . Перші два екрани, як і раніше, дозволено вказати таблицю і стовпець. Далі вказуємо порядок заміни ( [рис. 13.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.5) -1) і вибираємо створення нового стовпця ( [рис. 13.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.5) -2), щоб не втратити вихідні дані. Заміна буде зроблена, після чого не забудемо видалити додані пусті строки з "близько " - "далеко".

### SampleData

Останній інструмент в групі **Data Приготування** називається **Зразок Дані** (Образцы данных). Він дозволяє вирішити задачу формування навчального і тестового безлічі даних, а також виконати «балансування» даних.

У тих випадках, коли використовуваний метод інтелектуального аналізу вимагає попередньої моделі навчання (наприклад, для вирішення завдань класифікації), необхідно сформувати кілька наборів даних - для моделей навчання, перевірки її роботи, власного аналізу. Зразок **інструменту Дані** дозволяють підготувати потрібні набори.

Нехай необхідним випадковим способом розділити наявний набір даних на навчальний і тестовий вибір. Для цього потрібно запустити інструмент **Зразок Data** , указати звідки беремо дані для обробки ( [рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6) -1) і тип формованої вибору. Спочатку зробимо випадкову виборку, т.е. тип - **Випадковий Відбір проб** ( [рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6) -2). *Далі показано* відсоток записів із вихідного набору (або *точне число* записів), поміщених у вибір ( [рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6) -3 ) і *місце* для збереження отриманих результатів. на рис [. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6) -4 видно, що можна окремо зберегти сформовану вибірку і дані, в яких не попали. В підсумку можна отримати навчальний і тестовий набори. Хочеться звернути увагу на можливість використання зовнішнього джерела даних при формуванні вибору ( [рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6) -1). Це дозволяє використовувати дані, що зберігаються на MS SQLServer для формування наборів значень. Але як зазначено в описі інструменту, при використанні зовнішнього джерела даних у вікні, представленому на [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2) , буде доступний тільки *параметр* випадкової вибору.

При використанні засобів інтелектуального аналізу для виявлення редких подій, в навчальному наборі рекомендується збільшити частоту виявлення потрібного події в порівнянні з вихідними даними. Формування подібної вибору часто називають балансуванням даних, а інструмент **SampleData дозволяє виконати її** .

С помощью інструмента **Explore Дані** проаналізуємо розподіл клієнтів в наборі даних по регіонам. на рис [. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7) -1 видно, що приблизно п'ята частина клієнтів у нас з **Тихоокеанського регіону** (будемо читати це Азіатсько-Тіхоокеанським регіоном). Формуємо набір даних, де таких клієнтів буде 50 %.

|  |  |
| --- | --- |
| 1 | 2 |
|  |  |
| 3 | 4 |
| **Рис. 13.6.** Зразок інструменту Дані |  |

Запустим інструмент **Зразок Data** , наведемо в якості джерела даних використовувану таблицю *Excel* і виберемо варіант формування ізбутової вибору з балансуванням даних ( **Oversample до баланс даних розподіли** , [рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7) -2). Далі приведемо таблицю, для якої виконується балансування, і частоту виявлення потрібного значення і розмір вибору ( [рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7) -3) . Буде створена нова *таблиця* з указаним користувачем назвою. Нова застосовна **Explore Дані** і переконаємося в тому, що вибірка сформована *відповідно до зазначених вище* вимог ( [рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7) -4).

|  |  |
| --- | --- |
| 1 | 2 |
|  |  |
| 3 | 4 |
| **Рис. 13.7.** Формування вибору із заданим розподілом клієнтів по регіонам |  |

**Завдання.** Проведіть описану в роботі обробку вибраного набору даних.