Міністерство освіти і науки України

Державний університет «Житомирська політехніка»

Факультет інформаційно-комп'ютерних технологій

Кафедра інженерії програмного забезпечення

**КУРСОВА РОБОТА**

з дисципліни: «Сучасні бази даних та аналіз даних»

на тему **«Аналіз даних вступників ВНЗ України»**

Студентки 1-го курсу, групи ІПЗм-19-1  
напряму підготовки 121 «Інженерія програмного забезпечення»

Бойко Таїси Олегівни

(прізвище, ім’я та по-батькові)

Керівник доцент кафедри ІПЗ ,

,к.т.н., доцент Сугоняк І.І, ,

Національна шкала\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кількість балів:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Оцінка: ECTS \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Члени комісії \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Сугоняк І.І.

(підпис) (прізвище та ініціали)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Марчук Г.В. .

(підпис) (прізвище та ініціали)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Ковальчук А.М.

(підпис) (прізвище та ініціали)

Житомир – 2019

ЗМІСТ

[ВСТУП 3](#_Toc58527684)

[1. ТЕРЕОТИЧНИЙ АНАЛІЗ МОДЕЛЕЙ ТА МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 5](#_Toc58527685)

[1.1 Основні поняття великих даних 5](#_Toc58527686)

[1.2 Постановка задачі 10](#_Toc58527687)

[1.3 Обґрунтування вибору методів реалізації 10](#_Toc58527688)

[2. СТРУКТУРА ІНФОРМАЦІЙНОГО СХОВИЩА ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ 12](#_Toc58527689)

[2.1 Виміри, показники, метрики, ER-моделі сховища даних 12](#_Toc58527690)

[2.2 Характеристика джерела даних для інформаційного сховища 13](#_Toc58527691)

[2.3 Опис структури БД 16](#_Toc58527692)

[2.4 Заповнення сховища даних 21](#_Toc58527693)

[3. ФІЗИЧНА МОДЕЛЬ ТА ПРОТОТИП ПРОГРАМНОГО КОМПЛЕКСУ 22](#_Toc58527694)

[3.1 Створення проекту аналітичної системи 22](#_Toc58527695)

[3.2 Розгортання OLAP-куба 23](#_Toc58527696)

[3.3 Кластеризація, лінійна регресія та прогнозування даних 26](#_Toc58527697)

[3.4 Організація звітності 29](#_Toc58527698)

[ВИСНОВКИ 35](#_Toc58527699)

[ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ 36](#_Toc58527700)

# ВСТУП

З стрімким розвитком інформаційних технологій кількість користувачів та їх даних стрімко збільшується щодня. В умовах цієї тенденції, виникла гостра необхідність змінювати підходи до зберігання та обробки даних, для того щоб впоратись з їх кількістю. В побутовому житті кожного з нас, на підприємствах та в державних установах невпинно зростає потік інформації, важливої для роботи. В результаті розвитку соціальних мереж та різноманітних сервісів для мультимедійних даних безперервно ростуть потреби в засобах, що надають можливість ефективно працювати з цими даними. Окрім збереження та обробки інформації, гостро стало питання аналізу даних користувачів. В результаті для органів державної влади й управління, телекомунікаційних та Інтернет-компаній, банків, підприємств роздрібної торгівлі, енергетики, житлово-комунальному господарстві накопичена інформація стає особливо важливим ресурсом, від ефективності управління яким суттєво залежить діяльності підприємств та компаній.

Для роботи з такою кількістю інформації постійно створюються нові апаратні та програмні засоби, які здатних швидко обробляти великі обсяги даних і дозволяють робити процес обробки більш економічно вигідним. Саме тому, такі умови сучасного світу інформаційних технологій створили напрямок в інженерії програмного забезпечення, що носить назву великі дані.

Значення терміну великі дані має інтуїтивно зрозумілий характер. Великі дані (англ. Big Data) це набори структурованої або неструктурованої інформації які за розмірами настільки великі, що традиційні способи та підходи їх управління не можуть бути застосовані до них[]. Інколи, великими даними називають феноменальне прискорення нагромадження даних та їх ускладнення.

Галузь освіти не є виключенням серед інших галузей людської діяльності, що продукують велику кількість даних. Оцінки студентів, навчальні програми, методи оцінювання досягнень учнів та дані створені іншими процеси діяльності різних навчальних закладів. Вся ця інформація накопичується щоденно. Саме, важливо мати інструменти аналізу цих даних.

Враховуючи цю інформацію, можна встановити завдання до курсової роботи.

*Мета* курсової роботи. Проектування та аналіз даних про успішність вступників до вищих навчальних закладів на території України.

З мети можна сформувати наступні *завдання* до курсової роботи:

* Теоретичний аналіз проектування та реалізації OLAP-систем
* Вибір фактів та вимірів для збереження
* Проектування сховища даних
* Заповнення сховища даних
* Вибір методів інтелектуального аналізу даних
* Реалізація звітності для сховища даних

*Предметом* дослідження є засоби аналізу та розробки сховища даних, проектування сховища даних.

*Об’єктом* дослідження є данні вступників до ВНЗ України, враховуючи успішність заяв на вступ.

# ТЕРЕОТИЧНИЙ АНАЛІЗ МОДЕЛЕЙ ТА МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

## Основні поняття великих даних

Терміном Big Data ("великі дані") окреслюють групу технологій та методів, за допомогою яких аналізують та обробляють величезну кількість даних, як структурованих так і неструктурованих, для отримання якісно нових знань. Якщо підсумувати, то це інформація, що не піддається обробці класичними способами через її величезний об'єм.

Точний розмір «великих» даних важко сформувати, оскільки розуміння розміру значно змінилось, порівняно з початком тисячоліття. За даними сайту statista.com, в 2025 році об’єм даних що сворюється досягне 175 Зетабайтів (за прогрозованим результатом). 1 зетабайт дорівнює 1e+12 гігабайтів. На рисунку 1.1 зображений графік росту кількості даних, що було чи буде створено (за прогнозом) з 2010 року.

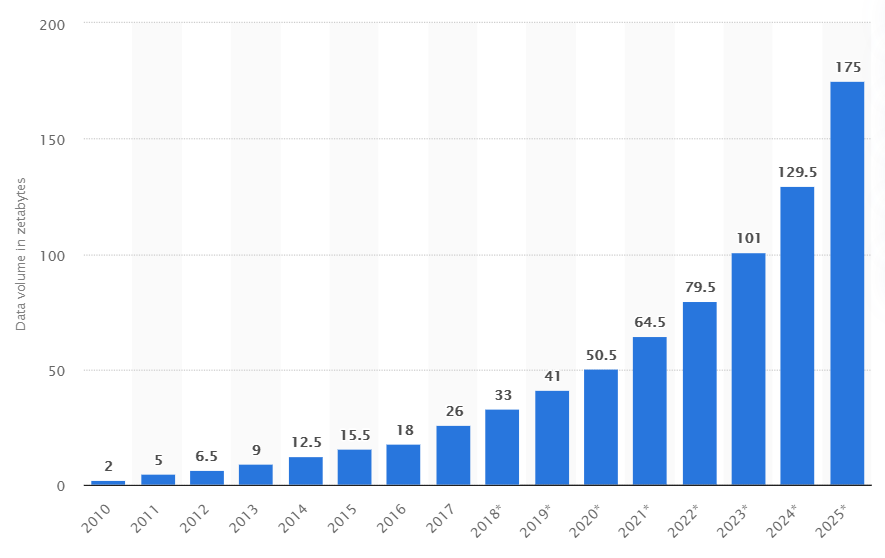


Рисунок . – Об’єм даних створених у світі починаючи з 2010 року

Технологія великих даних створена з метою вирішення наступних завдань:

* Обробляти більші, у порівнянні зі «стандартними» сценаріями, об’єми даних.
* Уміти працювати з даними, що швидко надходять у дуже великих об’ємах. Тобто даних не просто багато, а їх постійно стає все більше й більше.
* Вміти працювати зі структурованими і мало стуктурованими даними паралельно і у різних аспектах.

*Три ключові умови роботи з великими даним*

Визначальними характеристиками для великих даних є, окрім їх фізичного об’єму, й інші, які підкреслюють складність задачі обробки і аналізу цих даних.

Velocity (швидкість) – ця характеристика вказує на швидкість накопичення даних, яка постійно збільшується. Наприклад, 90 відсотків всієї інформації, якою оперує людство, зібрано за останні два роки. Також ця характеристика має на увазі швидкість обробки даних. Останнім часом збільшується попит на технології, що дозволяють використовувати обробку даних в режимі реального часу. Для швидкої роботи необхідно мати принцип відмовостійкісті системи. Оскільки обчислювальних вузлів у кластері може бути багато (іноді десятки тисяч) та їх кількість, не виключено, буде збільшуватись, зростає ймовірність виходу машин з ладу. Методи роботи з великими даними мають враховувати ймовірність таких ситуацій і передбачати превентивні заходи.

Volume (об'єм) – накопичена база даних охоплює настільки великий обсяг інформації, що його практично нереально обробляти та зберігати традиційними способами. Для них потрібен зовсім новий підхід та вдосконалені інструменти. Об’єм забезпечує горизонтальна масштабованість. Це базовий принцип обробки великих даних. Як вже було зазначено, великих даних з кожним днем стає все більше. Відповідно, необхідно збільшувати кількість обчислювальних вузлів, за якими розподіляються ці дані, при чому обробка має відбуватись без погіршення продуктивності.

Variety (різноманітність) – можливість одночасно обробляти структуровану та неструктуровану інформацію. Структурована інформація – це така, яку можна класифікувати. Наприклад, це може бути інформація з банківської бази даних, де чітко вказаний перелік клієнтів та їхні фінансові транзакції.

*Технології і тенденції роботи з Big Data*

Початково у сукупність підходів і технологій включались засоби масово-паралельної обробки невизначено-структурованих даних, такі як СУБД NoSQL, алгоритми MapReduce і засоби проекту Hadoop. У подальшому до технологій великих даних почали відносити й інші рішення, що забезпечують схожі за характеристиками можливості обробки надвеликих масивів даних, а також деякі апаратні засоби.

MapReduce — модель розподілених обчислювань у комп’ютерних кластерах, представлена компанією Google. Згідно з цією моделлю, додаток розділяється на значну кількість однакових елементарних завдань, що виконуються на вузлах кластера і потім, природнім шляхом зводяться у кінцевий результат.

NoSQL (від англ. Not Only SQL, не лише SQL) — загальний термін для різних нереляційних баз даних і сховищ, не означає якусь конкретну технологію чи продукт. Звичайні реляційні бази даних добре підходять для досить швидких і однотипних запитів, а на складних і гнучко побудованих запитах, характерних для великих даних, навантаження перевищує розумні межі і використання СУБД стає неефективним.

Hadoop — набор утилітів, бібліотек і фреймворків, що вільно розповсюджується, для розробки і виконання розподілених програм, які працюють на кластерах із сотень і тисяч вузлів. Вважається однією з основоположних технологій більшості даних.

R — мова програмування для статистичної обробки даних і роботи з графікою. Широко використовується для аналізу даних і фактично став

стандартом для статистичних програм.

Апаратні рішення. Корпорації Teradata, EMC та ін. др. пропонують апаратно-програмні комплекси, призначені для обробки великих даних. Ці комплекси поставляються як готові до установки телекомунікаційні шафи, що містять кластер серверів і керівне програмне забезпечення для масово-паралельної обробки. Сюди іноді відносять апаратні рішення для аналітичної обробки в оперативній пам’яті, зокрема, апаратно-програмні комплекси Hana компанії SAP і комплекс Exalytics компанії Oracle, незважаючи на те, що така обробка початково не є масово-паралельною, а об’єми оперативної пам’яті одного вузлаобмежуються кількома терабайтами.

Сховище даних (англ. data warehouse) — предметно орієнтований, інтегрований, незмінний набір даних, що підтримує хронологію і здатний бути комплексним джерелом достовірної інформації для оперативного аналізу та прийняття рішень. В основі концепції сховища даних (СД) лежить розподіл інформації, що використовують в системах оперативної обробки даних (OLTP) і в системах підтримки прийняття рішень (СППР). Такий розподіл дозволяє оптимізувати як структури даних оперативного зберігання для виконання операцій введення, модифікації, знищення та пошуку, так і структури даних, що використовуються для аналізу. В СППР ці два типи даних називаються відповідно оперативними джерелами даних (ОДД) та сховищем даних.

*Методи і техніка аналізу великих даних*

Методи класу Data Mining (видобуток даних, інтелектуальний аналіз даних, глибинний аналіз даних) — сукупність методів виявлення у даних раніше невідомих, нетривіальних, практично корисних знань, необхідних для прийняття рішень. До таких методів, зокрема, належать: навчання асоціативним правилам (association rule learning), класифікація (розгалуження на категорії), кластерний аналіз, регресійний аналіз, виявлення і аналіз відхилень тощо.

Краудсорсинг — класифікація і збагачення даних силами широкого, неозначеного кола особистостей, що виконують цю роботу без вступу у трудові стосунки.

Змішання та інтеграція даних (data fusion and integration) — набір технік, що дозволяють інтегрувати різнорідні дані з розмаїття джерел з метою проведення глибинного аналізу (наприклад, цифрова обробка сигналів, обробка природньої мови, включно з тональним аналізом).

Машинне навчання, включаючи навчання з учителем і без учителя – використання моделей, побудованих на базі статистичного аналізучи машинного навчання для отримання комплексних прогнозів на основі базових моделей.

Штучні нейронні мережі, мережевий аналіз, оптимізація, у тому числі генетичні алгоритми (genetic algorithm — евристичні алгоритми пошуку, що

використовуються для розв’язання задач оптимізації і моделювання шляхом випадкового підбору, комбінування і варіації потрібних параметрів з використанням механізмів, аналогічних натуральному відбору у природі)

Розпізнавання образів

Прогнозна аналітика

Імітаційне моделювання (simulation) — метод, що дозволяє будувати моделі, що описують процеси так, як вони би проходили у дійсності. Імітаційне моделювання модна розглядати як різновид експериментальних випробувань.

Просторовий аналіз (spatial analysis) — клас методів, що використовують топологічну, геометричну і географічну інформацію, що вилучається із даних.

Статистичний аналіз — аналіз часових рядів, A/B-тестування (A/B testing, split testing — метод маркетингового дослідження; при його використанні контрольна група елементів порівнюється із набором тестових груп, у яких один чи кілька показників були змінені, щоб з’ясувати, які зі змін покращують цільовий показник.

Візуалізація аналітичних даних — подання інформації у вигляді малюнків, діаграм, з використанням інтерактивних можливостей і анімації, як для отримання результатів, так і для використання у якості вихідних даних для

подальшого аналізу. Дуже важливий етап аналізу великих даних, що дозволяє показати найважливіші результати аналізу у найбільш зручному для сприйняття вигляді.

## Постановка задачі

За мету курсової роботи було визначено дослідження особливостей реалізації та проектування баз даних на основі технологій Big Data а також створення системи для аналізу даних вступників ВНЗ України.

Ця система дасть можливість оцінити успішність вступної кампанії для різних ВНЗ, на основі конкурсних балів вступників.

Для розробки курсової роботи потрібно:

* виконати порівняння різних засобів виконання поставленого завдання;
* розробити базу даних;
* заповнити базу даних;
* побудувати OLAP куб;
* проаналізувати отриманий куб;
* кластеризувати отримані дані.

## Обґрунтування вибору методів реалізації

На сьогоднішній день популярними засобами реалізації технологій великих даних є база даних Oracle та Microsoft SQL. Проведемо порівняння цих двох інструментів між собою.

Таблиця . – Порівняння MS SQL та Oracle

|  |  |  |
| --- | --- | --- |
| Параметр | MS SQL | Oracle |
| Синтаксис | Простіший та легший синтаксис | Складніший але ефективніший синтаксис |
| Підтримка платформи | Версії до 2017р. лише Windows, пізніше ще і Linux | Працює на різноманітних платформах |

Продоження таблиці 1.1

|  |  |  |
| --- | --- | --- |
| Параметр | MS SQL | Oracle |
| Мова | Transact SQL(T-SQL) | PL/SQL |
| Планування задач | SQL Server Agent | Планувальник задач Oracle або OEM |
| Опитимізація запитів | - | Зіркова (star) оптимізація |
| Roll back(відкат) | Не дозволено під час транзакції | Дозволено під час транзакції |
| Паралельний доступ | Недоступний під час процесу запису | Доступний під час запису, що зменшує час очікування |
| Паралельне виконання | Вирази INSERT, UPDATE, DELETE виконуються послідовно | Вирази INSERT, UPDATE, DELETE, and MERGE виконуються паралельно |
| Підтримка | Забезпечує технічні примітки, описи помилок, сценарії, виправлення та завантаження без додаткової плати | Виклик підтримки, за додаткову оплату за кожен випадок звернення |

ВИСНОВКИ ДО РОЗДІЛУ 1

У першому розділі курсової роботи було розглянуто теоритичні засади роботи з великими даними, проведено постановку задачі. Перед проектуванням та розробкою бази даних проведено порівняння та опис можливих методів реалізації та обґрунтування обраного методу.

# СТРУКТУРА ІНФОРМАЦІЙНОГО СХОВИЩА ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ

## Виміри, показники, метрики, ER-моделі сховища даних

Сховище даних — предметно орієнтований, інтегрований, незмінний набір даних, що підтримує хронологію і здатний бути комплексним джерелом достовірної інформації для оперативного аналізу та прийняття рішень. В основі концепції сховища даних лежить розподіл інформації, що використовують в системах оперативної обробки даних і в системах підтримки прийняття рішень.

Основними елементами даних, які зберігаються в сховищі даних, є:

* показники (змінні);
* виміри та їх ієрархія;
* факти.

Показник, або змінна, — це поле, значення якого однозначно визначаються фіксованим набором вимірів, що характеризують певний факт.

Показники, або змінні, складають, як правило, основний вміст сховища даних і можуть бути представленими:

* числовими характеристиками факту чи події, що відбувалися на об'єкті управління, для якого створюється сховище (наприклад, обсяги чи дохід від продажів);
* формулами, що, як правило, являють собою прості функції агрегування показників (змінних) для отримання узагальнених даних (наприклад, сума, яка консолідує значення змінної за кілька календарних періодів в одне підсумкове значення).

Вимір — це множина однотипних даних, що утворюють одну з граней куба і характеризують якусь ознаку показників, котрі знаходяться в комірці багатовимірного куба.

Наприклад, день, місяць, квартал, рік — це виміри часу; район, область, країна — це географічні виміри. За виміром виконується індексація даних у багатовимірній базі даних (ББД). Виміри бувають колективними (shared dimensional) та приватні (private dimensional).

Колективні виміри — це виміри, які можуть використовуватись одночасно в декількох кубах. Колективними вимірами можуть бути якісь ознаки, які використовуються при бізнес-аналізі різних предметних областей. Приватні виміри — це виміри, які належать конкретному кубу і створюються разом з ним. Інакше кажучи, це специфічні ознаки, що характеризують лише певну конкретну предметну область.

Сукупність вимірів визначають параметри простору, в якому можна буде виконувати бізнес-аналіз.

Відношення — це зв'язки між різними вимірами моделі та між окремими значеннями всередині певного виміру.

ER-модель (Entity-relationship model або Entity-relationship diagram) – це семантична модель даних, яка призначена для спрощення процесу проектування бази даних. З ER-моделі можуть бути породжені всі види баз даних: реляційні, ієрархічні, мережні, об’єктні. В основі ER-моделі лежать поняття “сутність”, “зв’язок” та “атрибут”.

Гранулювання - це рівень деталізації даних, що зберігаються в сховище даних. З погляду взаємозв'язку вимірювань і фактів останні можна розбити на наступні класи: адитивні факти, полуаддітивні факти, неаддитивні факти та числові заходи інтенсивності.

Таблиці фактів поділяють на три основні категорії, залежно від рівня деталізації фактів (гранульованості): транзакційна таблиця фактів, таблиця фактів періодичних моментальних знімків та таблиця фактів кумулятивних моментальних знімків

Схема «зірка» - має одну таблицю фактів, та кілька таблиць вимірів. Варто зазначити, що таблиці вимірів є денормалізованими.

Схема «сніжинка» - має одну таблицю фактів і кілька таблиць вимірів. Таблиці вимірів є нормалізованими.

## Характеристика джерела даних для інформаційного сховища

Для проведення аналізу результатів вступників до ВНЗ України, було отримано джерело даних з Єдиної Державної Електронної Бази з Питань Освіти за адресою https://vstup2019.edbo.gov.ua/uploads/KONKURS\_Universities.xlsx. Опис полів таблиці джерела даних наведено в таблиці 2.3. Оскільки дані знаходяться у вигляді Excel таблиці, було проведено нормалізацію даних.

Таблиця . – Опис полів джерела даних

|  |  |
| --- | --- |
| Поле таблиці джерела даних | Опис поля |
| Спеціальність | Спеціальність за якою подається заява |
| Спеціалізація | Спеціалізація визначеної спеціальності (опціональне значення) |
| Форма навчання | Обрана форма навчання: денна, заочна |
| ОКР | Освітньо-кваліфікаційний рівень |
| Орган управління | Орган, якому підпорядковується ВНЗ |
| Код закладу | Код ВНЗ в ЄДЕБО |
| Назва закладу | Повна назва ВНЗ |
| Усього подано заяв | Загальна кількість поданих заяв |
| Подано заяв на бюджет | Кількість поданих заяв на навчання за бюджетні кошти |
| Допущено до конкурсу | Кількість заяв, допущених до конкурсу |
| Середній пріоритет допущених | Середній пріоритет заяв допущених до конкурсу |
| Усього рекомендовано | Загальна кількість вступників, рекомендованих до вступу |
| Середній пріоритет рекомендованих | Середній пріоритет вступників, рекомендованих до вступу |
| Суперобсяг | Кількість місць на відкриті конкурсні пропозиції, які складають широку конкурсну пропозицію |
| Фіксований обсяг | Кількість місць на закриті конкурсні пропозиції, які складають фіксовану / регіональну конкурсну пропозицію |
| Рекомендовано за співбесідою | Загальна кількість вступників, рекомендованих до вступу за співбесідою |
| Рекомендовано за квотою-2 | Загальна кількість вступників, рекомендованих до вступу за квотою 2 |

В результаті обробки джерела, було отримано базу даних, що складається з шести таблиць. Перелік та опис таблиць бази даних наведено у таблиці 2.4. Схема структури бази даних зображена на рисунку 2.1. Розглянемо детально структуру бази даних за допомогою опису існуючих таблиць БД та схеми БД, що відображає зв’язки між таблицями.

Таблиця . – Опис таблиць бази даних

|  |  |
| --- | --- |
| Назва таблиці БД | Опис таблиці |
| Degrees | Освітньо-кваліфікаційні рівні навчання (бакалавр, магістр) |
| Departments | Органи управління, яким підпорядковується ВНЗ |
| Forms | Можливі форми навчання (денна, заочна) |
| Specialities | Спеціальності за якими подаються заяви |
| Submitions | Заяви на вступ |
| Universities | Вищі навчальні заклади |

Для побудови діаграми (схеми) структури бази даних було використано вбудовані інструменти Microsoft SQL Server у Microsoft SQL Server Management Studio 2017. В результаті на рисунку 2.1 зображена структура реляційної бази даних, з наявними зв’язками між таблицями по зовнішньому ключу. Таке розділення необхідне для побудови багатовимірного простору даних, що в свою чергу дозволить виконати та реалізувати певні аналітичні звіти.

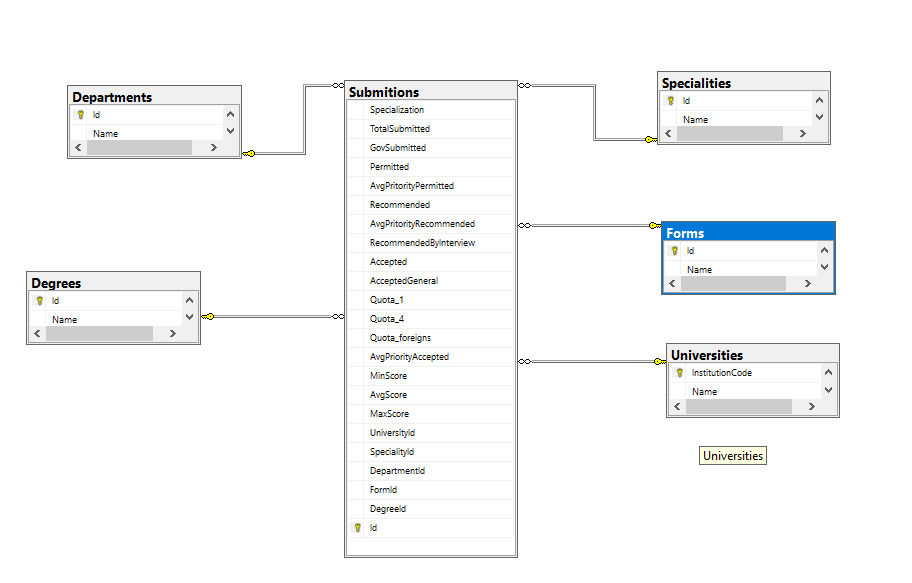


Рисунок . - Схема структури бази даних

## Опис структури БД

Для повного розуміння архітектури бази даних, розглянемо опис кожної таблиці БД (таблиці 2.5 – 2.10).

В таблиці Degrees (ступені) міститься перелік можливих освітньо-кваліфікаційних рівнів. В випадки системи моніторингу вступників до ВНЗ – це ступені бакалавр та магістр.

Таблиця . – Опис таблиці Degrees

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва поля | Тип даних | Primary key | Foreign key | Опис поля |
| Id | int | + | - | Унікальний індекс ОКР |
| Name | nvarchar (16) | - | - | Назва освітньо-кваліфікаційного рівня навчання |

Таблиця Departments зберігає можливі органи управління певних ВНЗ.

Таблиця . – Опис таблиці Departments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва поля | Тип даних | Primary key | Foreign key | Опис поля |
| Id | int | + | - | Унікальний індекс органу управління |
| Name | nvarchar (256) | - | - | Назва органу управління |

В таблиці Forms (форми навчання) міститься перелік можливих форм навчання у ВНЗ: денна та заочна.

Таблиця . – Опис таблиці Forms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва поля | Тип даних | Primary key | Foreign key | Опис поля |
| Id | int | + | - | Унікальний індекс форми навчання |
| Name | nvarchar (16) | - | - | Назва форми навчання |

В таблиці Specialities (спеціальність) міститься перелік всіх спеціальностей, за якими створюються заяви на вступ.

Таблиця . – Опис таблиці Specialities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва поля | Тип даних | Primary key | Foreign key | Опис поля |
| Id | int | + | - | Унікальний індекс спеціальністі |
| Name | nvarchar (100) | - | - | Назва спеціальність |

В таблиці Submitions (заяви) міститься перелік всіх поданих заяв на вступ за певною спеціальністю до певного ВНЗ.

Таблиця . – Опис таблиці Submitions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва поля | Тип даних | Primary key | Foreign key | Опис поля |
| Id | int | + | - | Унікальний індекс запиту |
| SpecialityId | int | - | + | Ідентифікатор спеціальності в таблиці Specialities |
| Specialization | nvarchar(255) | - | - | Назва спеціалізації |
| FormId | int | - | + | Ідентифікатор форми навчання в таблиці Forms |
| DegreeId | int | - | + | Ідентифікатор ОКР в таблиці Degrees |
| UniversityId | int | - | + | Ідентифікатор ВНЗ в таблиці Universities |
| DepartmentId | int | - | + | Ідентифікатор органу управління в таблиці Departments |
| TotalSubmitted | Float | - | - | Кількість поданих заяв |
| GovSubmitted | Float | - | - | Кількість поданих заяв на бюждетну форму навчання |
| Permitted | float | - | - | Кількість поданих заяв допущених до конкурсу |
| AvgPritorityPermitted | float | - | - | Середній пріоритет заяв допущених до вступу |
| Recommended | float | - | - | Рекомендовані до вступу |
| AvgPritorityRecommended | float | - | - | Середній пріоритет заяв рекомендованих до вступу |
| RecommendedByInterview | float | - | - | Середній пріоритет заяв рекомендованих до вступу за співбесідою |
| Accepted | float | - | - | Кількість заяв переданих до офіційного зарахування |
| AcceptedGeneral | float | - | - | Кількість обсягу зарахування на загальних умовах |
| Quota\_1 | float | - | - | Кількість обсягу зарахування за першою квотою |
| Quota\_4 | float | - | - | Кількість обсягу зарахування за четветрою квотою |
| Quota\_foreigns | float | - | - | Кількість обсягу зарахування за квотою для іноземців |
| AvgPriorityAccepted | float | - | - | Середній пріоритет заяв прийнятих до вступу |
| MinScore | float | - | - | Мінімальний середній бал заяв на вступ |
| AvgScore | float | - | - | Середнє значення середнього балу заяв на вступ |
| MaxScore | float | - | - | Максимальний середній бал заяви на вступ |

## Заповнення сховища даних

База даних була заповнена шляхом імпорту Excel таблиці за допомогою вбудованих інструментів Microsoft SQL Server Management Studio 2017. Після імпорту даних була проведена нормалізація загальної таблиці для розділення на менші таблиці.

На рисунку 2.2 зображено джерело даних в вигляді Excel таблиці.

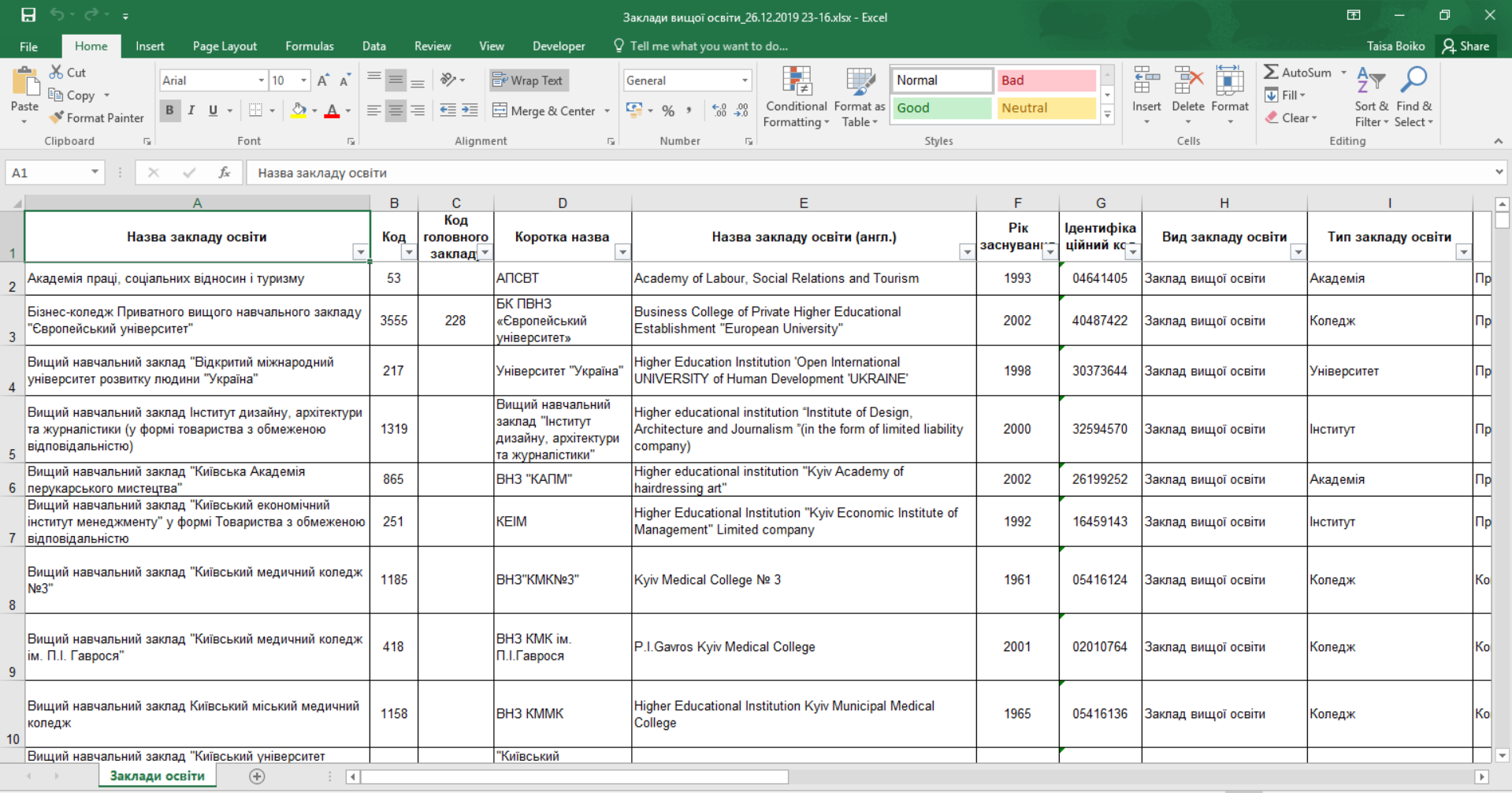


Рисунок . – Джерело даних в вигляді Excel таблиці

ВИСНОВКИ ДО РОЗДІЛУ 2

У другому розділі курсової роботи були описані метрики та моделі сховища даних, проведена робота по обробці джерела даних для інформаційного сховища, розроблена та описана детальна структура бази даних та заповнене сховище.

# ФІЗИЧНА МОДЕЛЬ ТА ПРОТОТИП ПРОГРАМНОГО КОМПЛЕКСУ

## Створення проекту аналітичної системи

У середовищі Microsoft Visual Studio 2019 було створено Analysis Services Multidimensional and Data Mining проект.

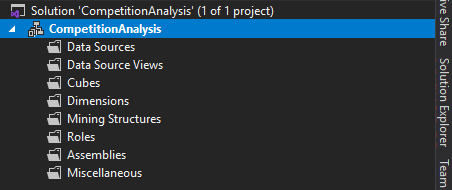


Рисунок 3. – Створений проект для проведення інтелектуального аналізу

Наступним кроком було додано джерело даних, у ролі якого виступає вже заздалегідь створена база даних на основі CSV документу з статистики вступної кампанії.

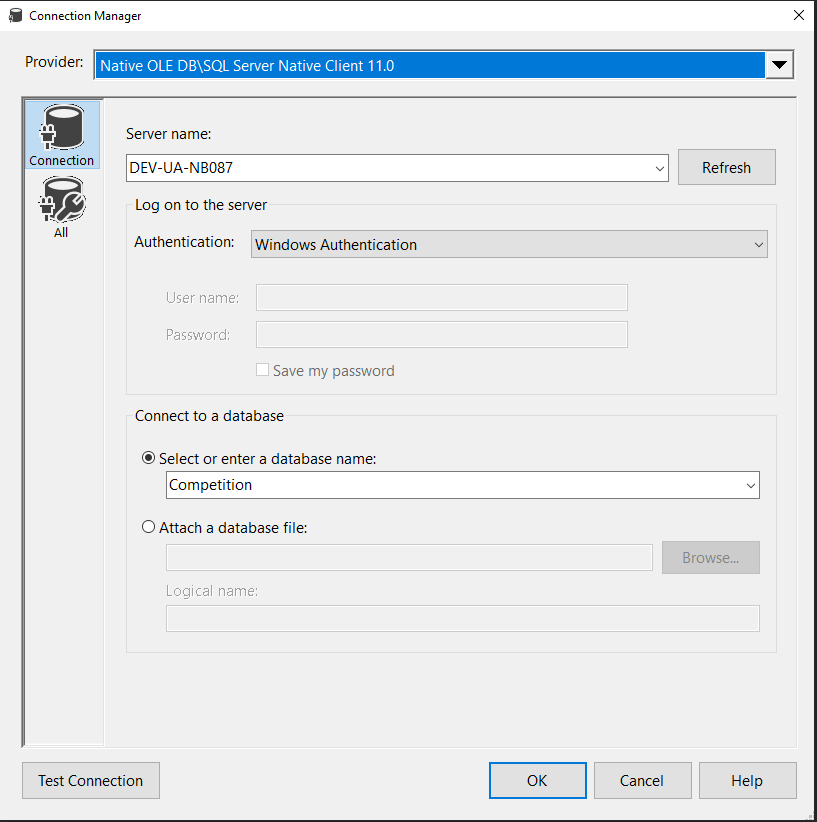


Рисунок 3.2 – Створений проект для проведення інтелектуального аналізу

Ім’я для нового джерела даних (Data Source) було обране Competition, що відповідає вже раніше створеній базі даних за допомогою пакетних інструментів застосунку Miscrosoft SQL Management Studio. Авторизація підключення була виконана за допомогою Windows Authentication для поточного користувача.

## Розгортання OLAP-куба

Засобами Microsoft Visual Studio було виконано розгортання OLAP-куба. Першим кроком було створено джерело даних на основі вже існуючого підключення до бази даних, щ обуло сконфігуровано у попередньому підрозділі. Після чого було виконане створення відображення джерелах даних.

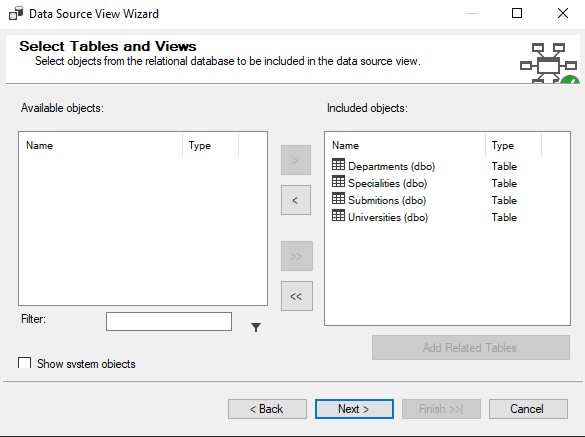


Рисунок 3.3 – Створення відображення джерела даних

Створене відображення джерела даних дозволяє робити певну задану вибірку із існуючого джерела даних. Після створення джерела було вирішено додати ще дві таблиці «Forms» та «Degrees», для формування ширшої та інформативнішої збірки системи. Саме відображення джерела даних дозволяє, як і джерело даних редагувати створений об’єкт у проекті Analysis Services Multidimensional and Data Mining, що впроваджує гнучкість у процесі роботи над таким проектом. У результати отримали таке відображення джерела даних:

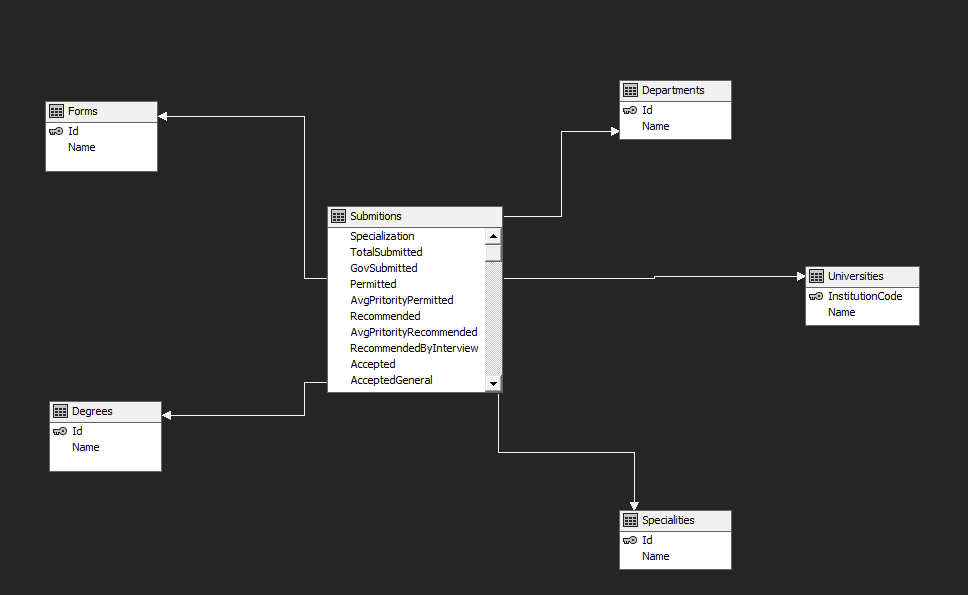


Рисунок 3.4 – Діаграма відображення джерела даних

Наступним кроком було власне створення OLAP-куба за допомогою інструменту Cube Wizard.

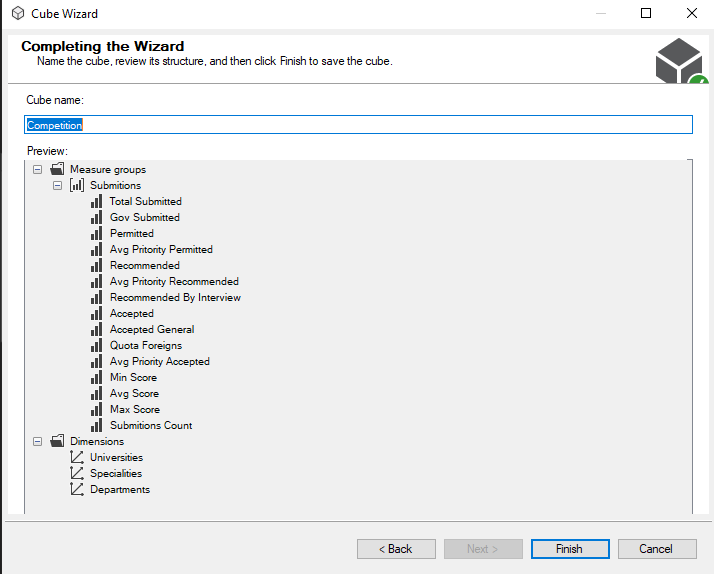


Рисунок 3.5 – Помічник створення OLAP-куба

Завершуючи процес налаштування OLAP-куба, було виконано процес

розгортання, за обраними вимірами. Як і у випадку з відображенням джерела даних внаслідок було додано ще виміри «Forms» та «Degrees».

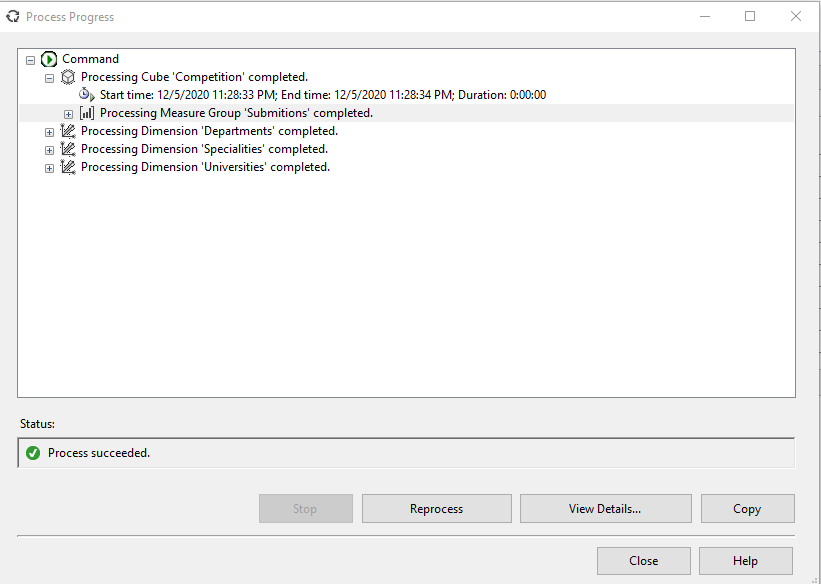


Рисунок 3.6 – Процес розгортання OLAP-куба

Виконавши, процес розгортання куба нам стане доступно багато аналітичних інструментів. Сюди входить перегляд структури даних, які використовувались для побудови куба, звітності по використаним вимірам, різні обчислення над зібраними даними, KPIs, а також різноманітні маніпуляції та агрегації даних. До того ж вбудований інструментарій Microsoft Visual Studio дозволяє виконати експорт побудованого та розгорнутого OLAP-куба у застосунки Microsoft Office, а саме Microsoft Office Excel. Це дозволяє використовувати отримані мультивимірні дані у вигляді таблиць, графіків діаграм, вбудованими засобами пакету Microsoft Office Excel. Виконавши такий імпорт, для прикладу побудуємо кругову діаграму зарахованих абітурієнтів за кожним освітнім департаментом, вибравши відповідні поля таблиць із секції побудови даних вимірів.

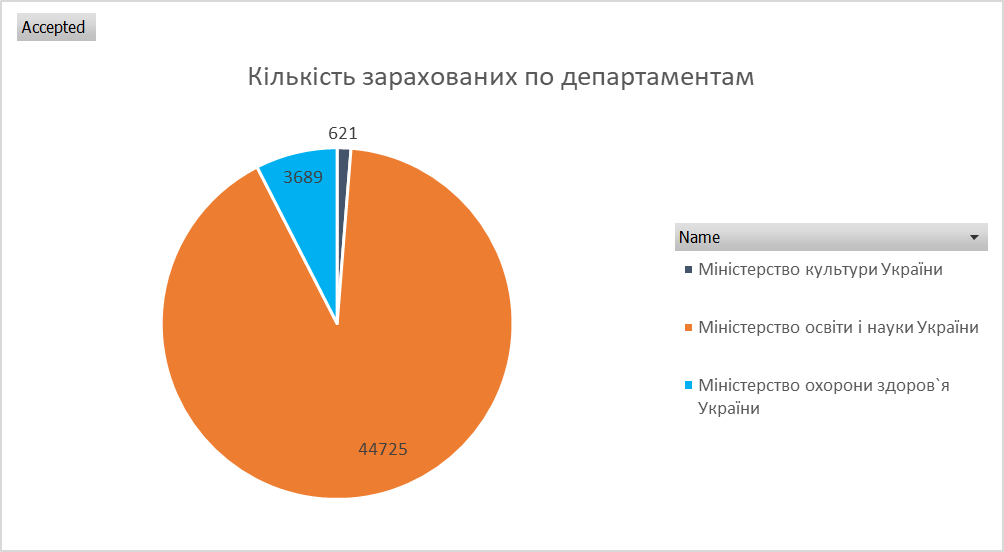


Рисунок 3.7 – Використання імпортованих даних з OLAP-куба у середовищі Microsoft Office Excel

## Кластеризація, лінійна регресія та прогнозування даних

Для того, щоб провести автоматичний аналіз даних, використовується Data Mining (здобич (витягання) знань). Це нова технологія інтелектуального аналізу даних з метою виявлення прихованих закономірностей у вигляді значущих особливостей, кореляцій, тенденцій і шаблонів. Сучасні системи здобичі даних використовують засновані на методах штучного інтелекту засоби уявлення і інтерпретації, що і дозволяє знаходити розчинену в терабайтних сховищах не очевидну, але вельми цінну інформацію. Фактично, ми говоримо про те, що в процесі Data mining система не відштовхується від наперед висунутих гіпотез, а пропонує їх сама на основі аналізу. Існує безліч визначень Data Mining, але в цілому вони співпадають у виділенні чотирьох основних ознак. Згідно визначенню, Г. Пиатецкого-Шаниро (G. Pia-tetsky Shapiro,GTE Labs), одного з ведучих світових експертів в даній області, Data Mining — дослідження і виявлення алгоритмами, засобами штучного інтелекту в "сирих даних"  
прихованих структур, шаблонів або залежності, яка: раніше не були відомі;

* нетривіальні;
* практично корисні;
* доступні для інтерпретації людиною і необхідні для ухвалення рішень в різних сферах діяльності.

Специфіка сучасних вимог до продуктивної переробки інформації наступна:

* дані мають необмежений обсяг;
* дані є різнорідними (кількісними, якісними, текстовими);
* результати повинні бути конкретний і зрозумілий;
* інструменти для обробки "сирих даних" повинні бути прості у використовуванні [11].

Отже, підсумовуючи, Data Mining – це процес виявлення у необроблених

даних раніше невідомих нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах діяльності.

Виконаємо процес кластеризації даних, для цього створимо нову Mining Structure у комбінації з моделлю Microsoft Clustering.

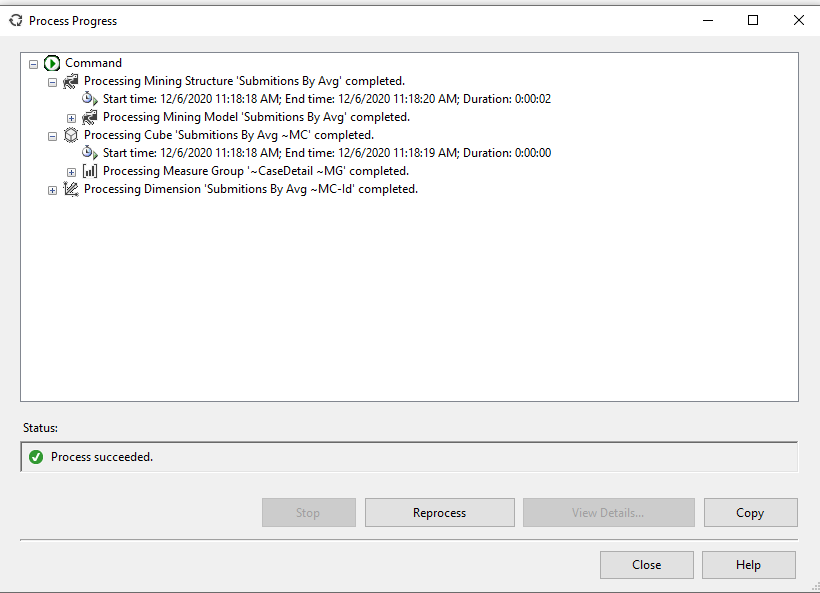


Рисунок 3.8 – Процес створення Mining Structure

Аналогічно виберемо колонки, на основі яких будемо проводити аналіз лінійною регресією. Для цього до створеної структури добування даних додамо нову модель на основі шаблону Microsoft Linear Regression. Отримаємо наступну структуру:

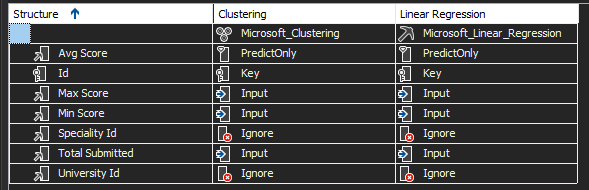


Рисунок 3.9 – Моделі кластеризації та лінійної регресії у комплексі Data Mining Strcuture

Після цього матимемо змогу переглянути візуальне представлення утворених кластерів даних за допомогою інструменту Mining Model Viewer.

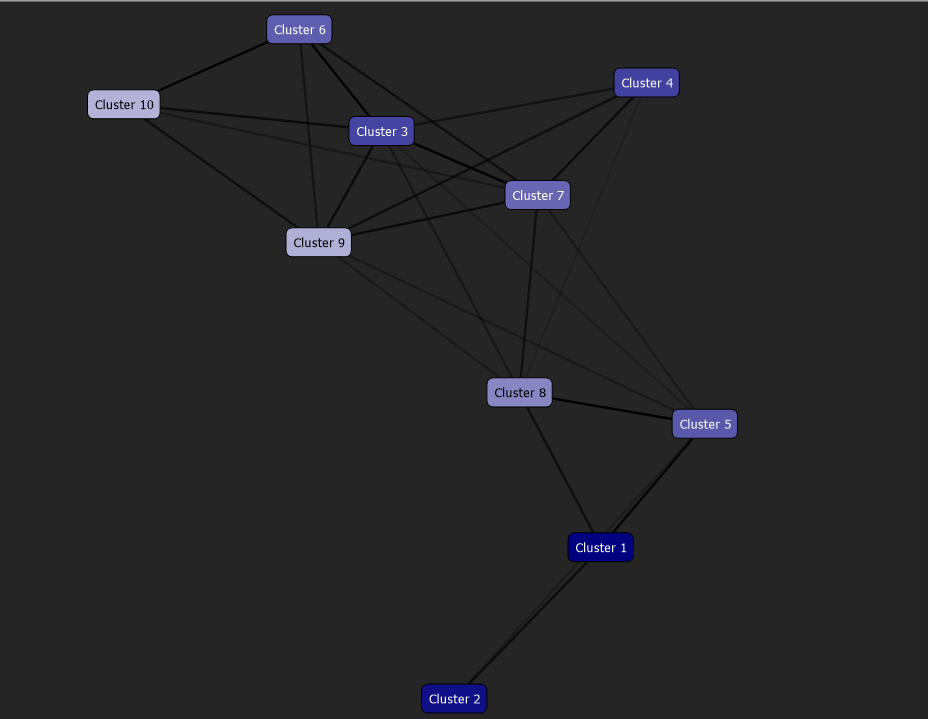


Рисунок 3.9 – Діаграма кластерів у Mining Model Viewer

Далі проведемо порівняння двох моделей за допомогою графіку підйому (англ. lift chart). Графік підйому графічно представляє поліпшення, яке забезпечує модель видобутку в порівнянні із випадковим припущенням, і вимірює це поліпшення за допомогою величини lift score. Порівнюючи показники підйому для різних моделей, можна визначити, яка модель найкраща. Також можна визначити момент, коли прогнози моделі стають менш корисними. Наприклад, переглянувши графік підйому, можна зрозуміти, що рекламна кампанія, ймовірно, буде ефективною лише для 30% клієнтів, і використати цю цифру для обмеження обсягу рекламної кампанії.

У SQL Server Data Mining графік підйому може порівнювати точність декількох моделей, що мають однаковий передбачуваний атрибут. [12]

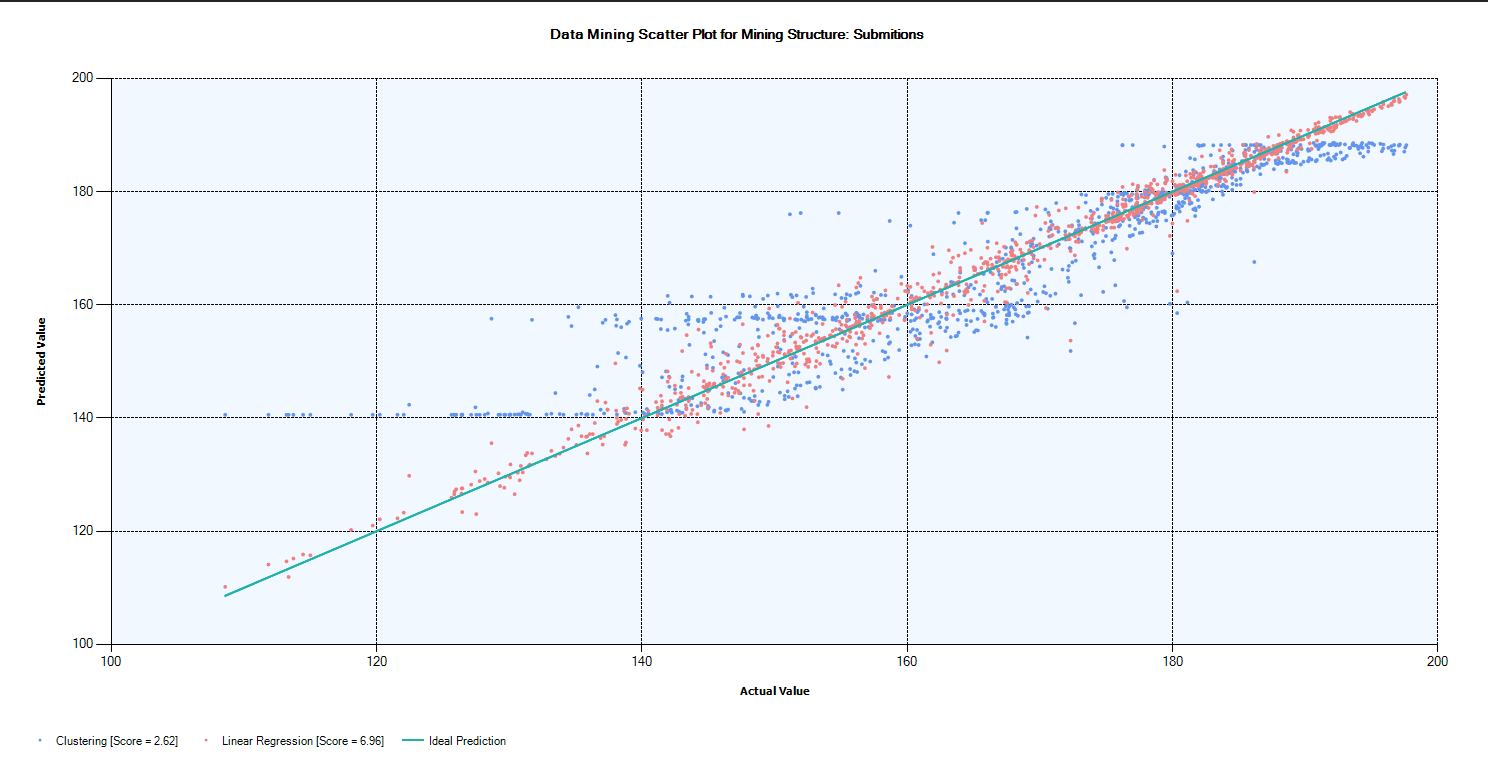


Рисунок 3.10 – Графік підйому, для аналізу точності обраних моделей.

## Організація звітності

Для побудови та організації звітності оброблюваних даних було використано рішення компанії Microsoft – Power BI Desktop. Power BI Desktop – програмний застосунок для встановлення на персональні комп'ютери, що входить до платформи Microsoft Power Platform, який дозволяє проводити різні аналітичні операції над даними за допомогою зручного інтерфейсу та великої кількості вбудованих інструментів.

Для того щоб працювати з даними необхідно виконати їх імпорт. Додаток дозволяє використовувати різні типи джерел даних для імпорту, проте у контексті даної роботи, найбільш доцільним є використання імпорту напряму з бази даних Competition. Для цього було виконане підключення до серверу, за допомогою аутентифікації Windows. Після цього було обрано усі таблиці бази даних, для імпорту у систему Power BI.

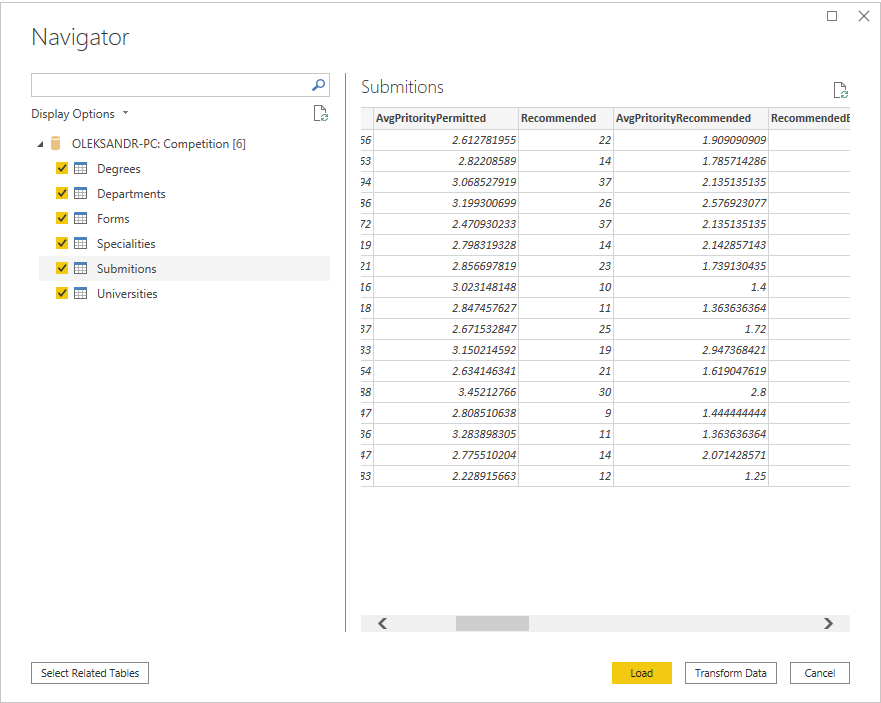


Рисунок 3.11 – Імпорт з бази даних Competition у Power BI

Виконавши імпорт даних, ми можемо обрати певні поля, для побудови звітів. Широкий функціонал додатку Power BI Desktop дозволяє проводити різного роду операції та фільтрації над імпортованими даними перед відображенням власне звітів системи. Для прикладу, було сформовано наступні звіти:

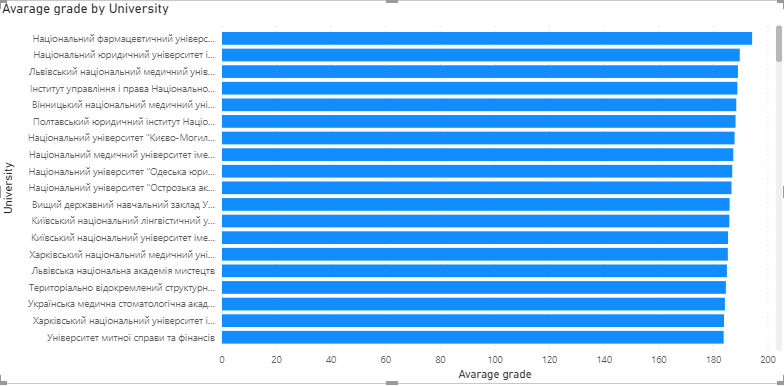


Рисунок 3.12 – Середній бал серед поданих заяв по університетам

Далі виконаємо звіт, який включатиме кількісне відношення зарахованих абітурієнтів до загального числа заяв в залежності від ступеня.

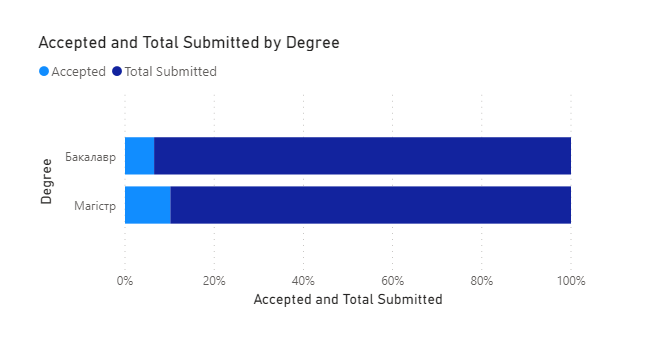


Рисунок 3.13 – Відношення зарахованих абітурієнтів до загальної кількості заяв залежно від ступеню

З даного звіту можна зробити висновок, що абітурієнти, які подають заяви на магістра є більш впевненими у спеціальності, на якій вони збираються навчатися, що дійсно знаходить об’єктивне підкріплення у життєвому досвіді.

Також створимо звіти за кількістю зарахованих абітурієнтів та загальною кількістю поданих заяв за спеціальностями.

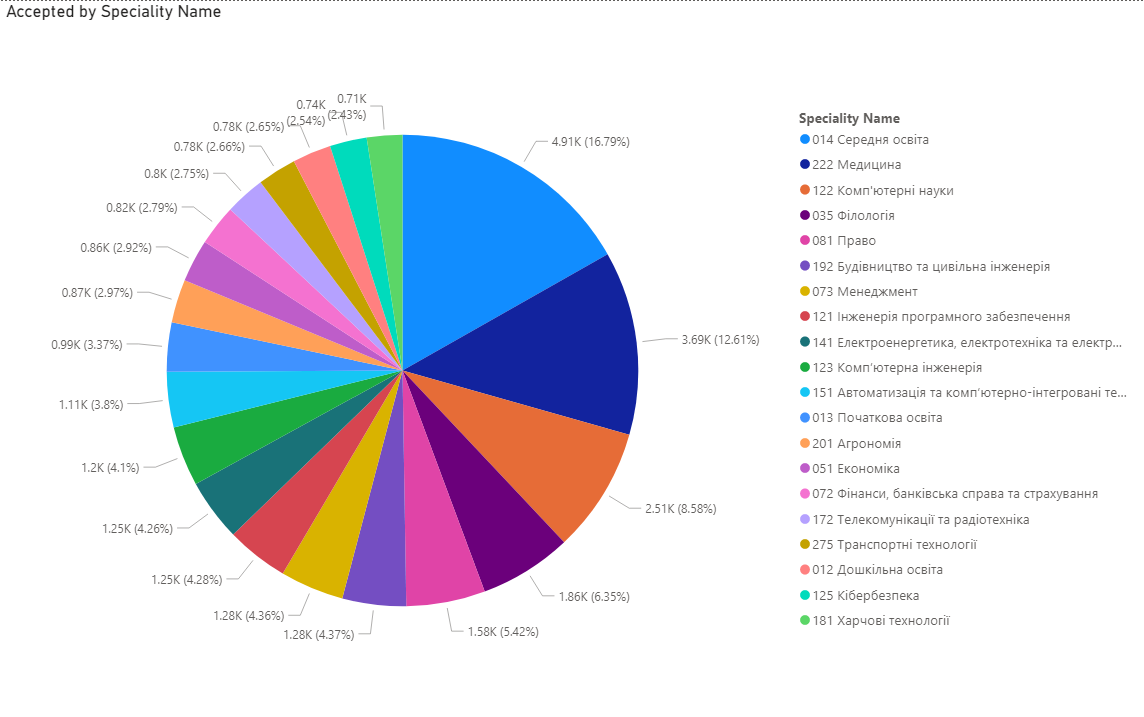


Рисунок 3.14 – Кількість зарахованих заяв по спеціальностям з фільтрацією

Як бачимо найбільше доступних місць, за якими зарахували абітурієнтів, мають спеціальності:

1. Середня освіта
2. Медицина
3. Комп’ютерні науки
4. Філологія
5. Право.

Проте кількість зарахованих абітурієнтів зовсім на свідчить про популярність спеціальності, скоріше це говорить про попит Міністерства щодо певних спеціальностей. Для того, щоб виконати порівняльний аналіз створимо звіт за кількістю поданих заяв за спеціальностями. Таким чином можна побачити динаміку, щодо конкурсу на певні фахові напрями навчання. На рисунку 3.15 зображено такий звіт у вигляді круговою діаграми. Також тут застосована фільтрація для того щоб відсіяти «малопопулярні» спеціальності.

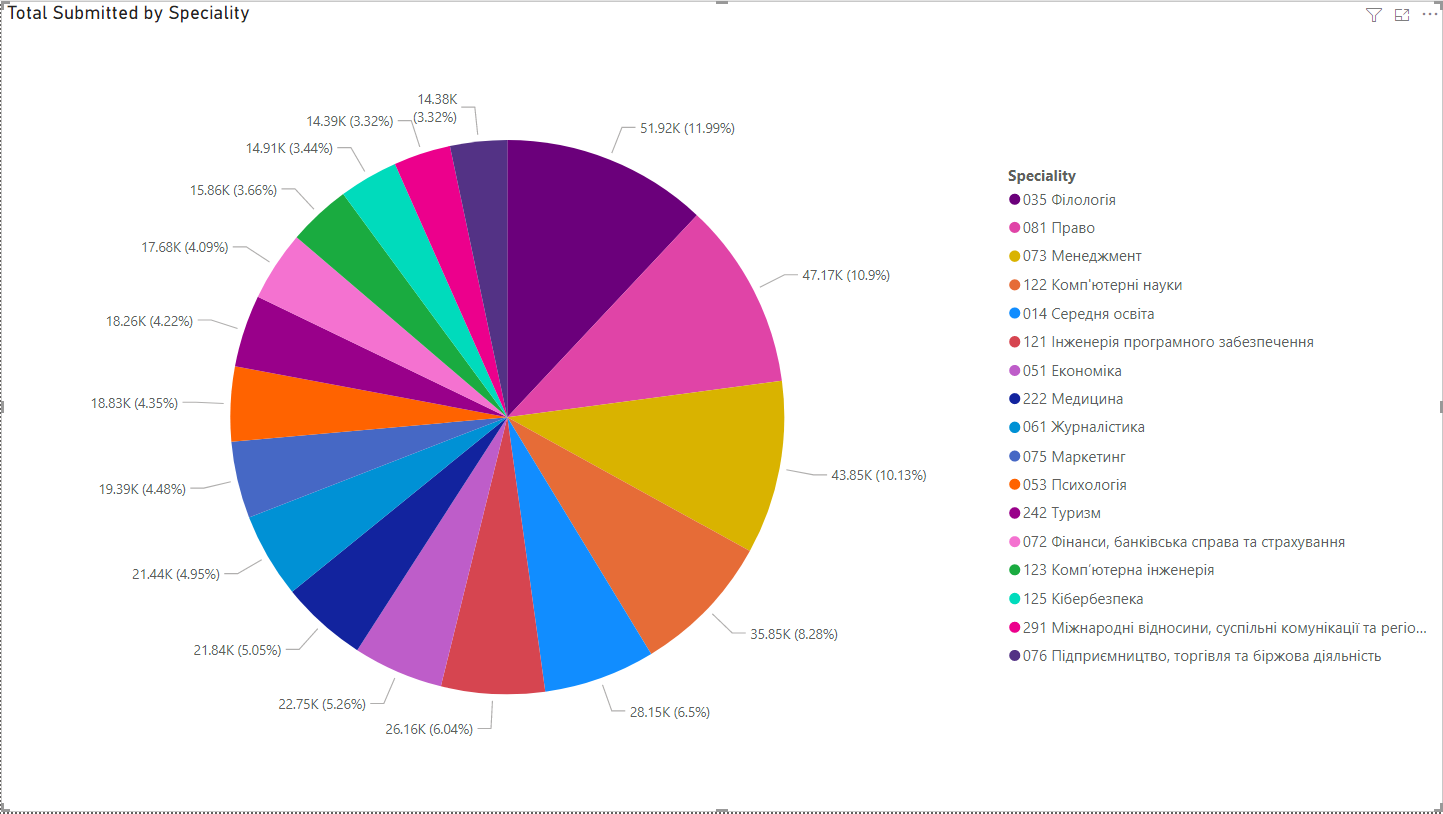


Рисунок 3.14 – Кількість зарахованих заяв по спеціальностям з фільтрацією

Як бачимо найбільше поданих заяв абітурієнтами, мають спеціальності:

1. Філологія
2. Право
3. Менеджмент
4. Комп’ютерні науки
5. Середня освіта.

Оберемо ці спеіальності для наочного співвідношення та побудуємо наступний звіт:

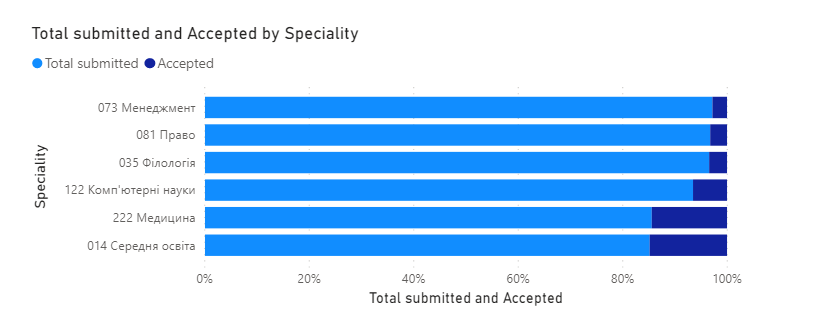


Рисунок 3.14 – Кількість зарахованих заяв по спеціальностям з фільтрацією

Отже, можна зробити висновок, що найбільший конкурс припадає на напрям підготовки Менеджменту, за ним дуже близько йдуть Право та Філологія, після чого вже Комп’ютерні науки, Медицина та Середня освіта.

ВИСНОВКИ ДО РОЗДІЛУ 3

У третьому розділі курсової роботи було описано процес створення та розгортання OLAP-куба з джерела даних у вигляді бази даних MS SQL Server. Також було описано процес аналізу та передбачення атрибутів даних за допомогою комплексу Data Mining. В останньому підрозділі було продемонстровано можливості Power BI Desktop у побудові звітності аналізованої системи.

# ВИСНОВКИ

В ході виконання курсової роботи було проведено аналіз моделей та методів інтелектуального аналізу даних вступників до ВНЗ України. Тому в процесі роботи було виконано збір та підготовку даних, які в подальшому використовувались для проведення інтелектуального аналізу та побудови звітності системи.

Предметом дослідження були засоби аналізу та розробки сховища даних, проектування сховища даних. Об’єктом було дослідження кількісних даних вступників до ВНЗ України, враховуючи їх успішність.

У процесі виконання було закріплено навички роботи з інструментами інтелектуального аналізу даних в середовищі Microsoft Visual Studio, використовуючи можливості платформи Microsoft SQL Analysis Services. Для отримання результатів було проведено розгортання OLAP-куба з визначеним набором вимірів. Також у проекті було створено Data Mining структуру для підбору найкращої моделі передбачення атрибутів середнього значення балу вступників. За допомогою додатку Power BI Desktop було виконано звітність системи та проведено аналіз даних звітності на основі отриманих візуальних матеріалів.

Отже, в результаті виконання курсової роботи було виконано усі поставлені задачі по проведенню інтелектуального аналізу даних та на практиці закріплено навички роботи з інструментами по реалізації такого аналізу.

# ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Jeffrey Richter. CLR via C# (4th Edition). Redmond, Washington – 2012.
2. Data created worldwide 2010-2024 [Електронний ресурс] – Режим доступу: https://www.statista.com/statistics/871513/worldwide-data-created/.
3. Data Warehousing – Concepts [Електронний ресурс] – Режим доступу: https://www.tutorialspoint.com/dwh/dwh\_data\_warehousing.htm.
4. Microsoft Power BI [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Microsoft\_Power\_BI.
5. Power BI Desktop [Електронний ресурс] – Режим доступу до ресурсу: https://powerbi.microsoft.com/ru-ru/desktop/.
6. Бази даних. ER-модель. Сутність. Атрибут. Види атрибутів [Електронний ресурс] – Режим доступу: https://www.bestprog.net/uk/2019/01/24/the-concept-of-er-model-the-concept-of-essence-and-communication-attributes-attribute-types-ua/.
7. Великі дані [Електронний ресурс] – Режим доступу до ресурсу: https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye.
8. Єдина державна електронна база з питань освіти [Електронний ресурс] – Режим доступу: https://vstup.edbo.gov.ua/.
9. Технологія сховищ даних атрибутів [Електронний ресурс] – Режим доступу: https://skaz.com.ua/informatika/6950/index.html.
10. Що таке Big Data? [Електронний ресурс] – Режим доступу: https <http://thefuture.news/bigdata>.
11. Grid і інтелектуальна обробка даних Data Mining [Електронний ресурс] – Режим доступу: <http://allted.kpi.ua/downloads/DataMining.pdf>
12. Lift Chart (Analysis Services - Data Mining) [Електронний ресурс] – Режим доступу: https://docs.microsoft.com/uk-ua/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=asallproducts-allversions