МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ПВНЗ
«МІЖНАРОДНИЙ НАУКОВО-ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ АКАДЕМІКА Ю. БУГАЯ»

Кафедра інформаційних та комунікаційних технологій

Освітня компонента: «Інтелектуальний аналіз даних»

Дослідницька РГР – Візуальні компоненти

Виконав:
студент 4 курсу денної форми навчання
група І26
Бондаренко О.В.

Перевірила:
доц. Гриб'юк О.О.

Київ – 2025

Тема проєкту: Аналіз поведінкових даних та розробка аналітичних моделей для прийняття рішень.

Прізвище та ім’я виконавців проєкту: Бондаренко Олег.

Концепція проєкту: посилання на github.

Мета проєкту: Формування системи інтелектуального аналізу поведінкових та предметно-орієнтованих даних з метою моделювання сценаріїв розвитку бізнес-рішень, виявлення трендів та оптимізації взаємодії.

Завдання проєкту:

1. Отримати вибіркові дані поведінкових метрик користувачів веб-проєкту з різних джерел (первинна предметна область) та провести їх первинну обробку.
2. Здійснити регресійний та survival-аналіз для виявлення впливу поведінкових ознак на ймовірність покупки або конверсії.
3. Провести розкладання часових рядів поведінкових метрик на складові (тренд, сезонність).
4. Провести апробацію розробленого комплексу методів на новій предметній області – ідентифікація діамантів.
5. Сформувати висновки та рекомендації щодо оптимізації бізнес-рішень на основі виявлених закономірностей, трендів та результатів моделювання.

ВІЗУАЛЬНІ КОМПОНЕНТИ

Вхідні данні. Вибірка містить такі змінні:

1000 записів про сертифікацію та продаж діамантів за 3-річний період (2023-2025). Слугує єдиним джерелом даних для всіх аналітичних методів РГР

report_id (Унікальний ID) – Маска ID: DR-[NNNNN]

Блок 3 (Часовий вимір):
report_date (Дата сертифікації)

Блок 1 (Фізичні характеристики):
carat_weight (Вага, ct)
color_grade (Колір)
clarity_grade (Чистота)
polish_grade (Полірування)
proportions_grade (Пропорції)
symmetry_grade (Симетрія)
cut_grade (Огранювання) – Розрахункове поле
fluorescence_grade (Флуоресценція)
stone_origin (Походження)

Блок 4 (Мета-дані процесу оцінки):
expert_id (ID експерта)
evaluation_time_min (Час оцінки, хвилини)
report_notes_length (Довжина коментаря, кількість слов)
report_sentiment (Тональність коментаря)

Блок 2 (Цільові змінні / Ринкові показники):
price (Ціна, $)
is_investment_grade (Інвест. клас: 1/0)
is_report_rejected (Звіт відхилено: 1/0)
is_sold (Продано: 1/0)

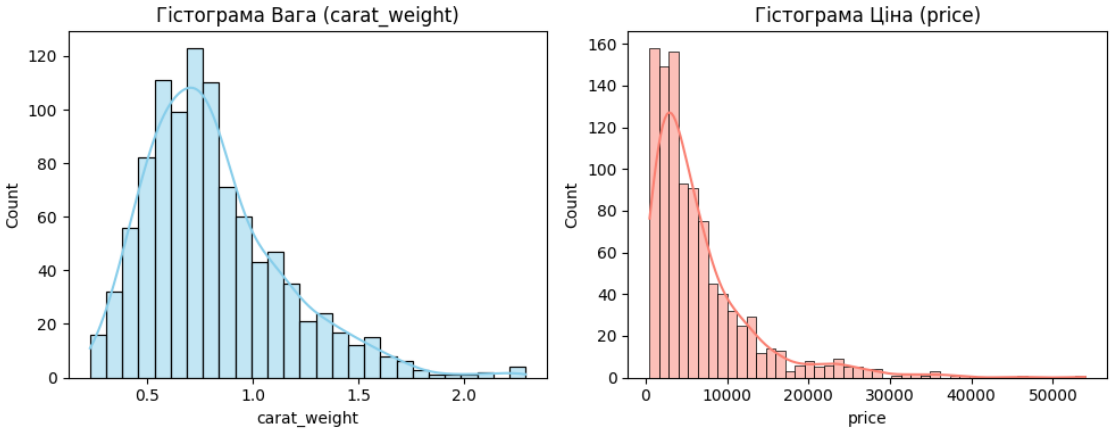
days_on_market (Днів на ринку)
sale_date (Дата продажу) – це поле може бути NaT (Not a Time), якщо is_sold = 0

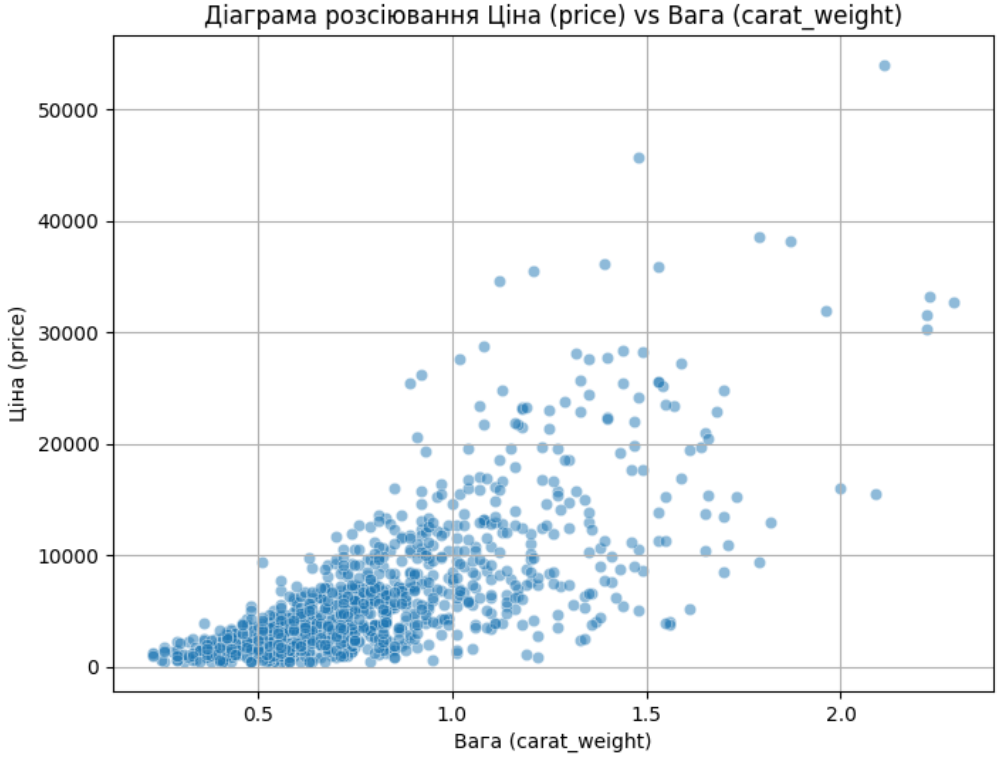
Датасет:
https://github.com/OlegBon/mntu-y4-sem7/blob/main/IntelligentDataAnalysis/data/diamond/diamonds_dataset.csv

1. Метод кореляційного аналізу даних

Побудовано дві гістограми для змінних Вага (carat_weight) та Ціна (price), а також діаграму розсіювання, яка демонструє наявність прямого зв'язку між ними. Графіки збережено у папці results/diamond:

- 01_correlation_analysis_histograms.png – розподіл значень
- 01_correlation_analysis_scatterplot.png – графік кореляції

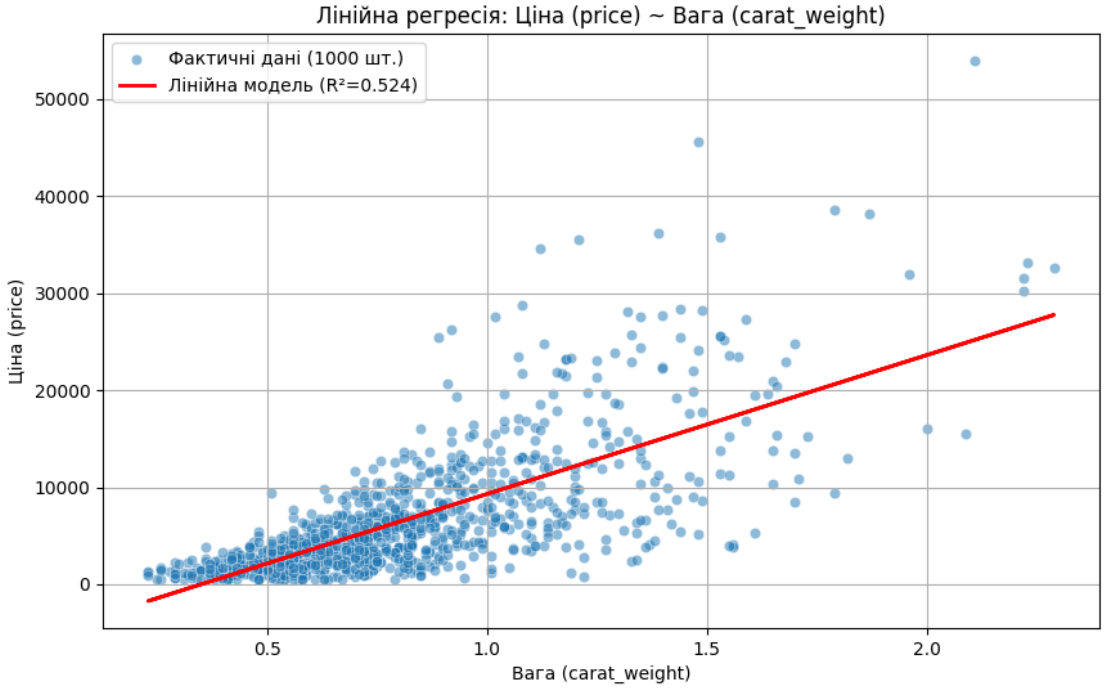




2. Лінійний регресійний аналіз + Інтелектуальна задача

Побудовано лінійну регресійну модель між залежною (Ціна (price)) та незалежною (Вага (carat_weight)) змінними за методом найменших квадратів (OLS).

Червона лінія намагається "вписатися", але очевидно програє.



Початковий стан:

- Є набір звітів (diamonds_dataset.csv) з метаданими процесу оцінки.
- Поточний відсоток відхилених звітів (Rejection Rate) є високим: 8.5% (Примітка: ця цифра буде розрахована зі 1000 записів, вона базується на закладеній нами логіці помилок Expert_5 та низького часу оцінки).

Ціль:

- Знизити відсоток відхилень до $\leq 2.0\%$.
- Мінімізувати кількість управлінських дій.

Алгоритм агента:

1. Зчитати дані з diamonds_dataset.csv.
2. Підготувати ознаки (Блок 4): expert_id (ID експерта), evaluation_time_min (Час оцінки, хвилини), report_notes_length (Довжина коментаря, кількість слов), report_sentiment (Тональність коментаря).
3. Побудувати класифікаційні моделі (LDA, QDA) для прогнозу is_report_rejected (Звіт відхилено: 1/0).
4. Оцінити точність моделей і вибрати стабільну.
5. Задати набір можливих управлінських дій:
    a. retrain_expert_5 (Провести тренінг для Експерта 5)
    b. enforce_time_policy (Заборонити оцінки < 20 хвилин)
    c. implement_checklists (Впровадити чек-листи для покращення якості коментарів)
6. Змоделювати вплив кожної дії на метрики (simulate_action).
7. Прогнозувати відсоток відхилень після кожної дії.
8. Обрати найкращу дію (або комбінацію), яка дає відсоток відхилень $\leq 2.0\%$.
9. Зберегти результати у звіт.

Роль аналітичного агента:

- Отримує метадані про процес оцінки.
- Навчається класифікувати ймовірність відхилення звіту.
- Має набір дій, які можуть змінити ці мета-дані.
- Моделює вплив кожної дії на фінальний відсоток відхилень.
- Приймає рішення, які дії виконати для досягнення цільової мети.

Агент діє в умовному середовищі, де кожна дія змінює стан системи. Його мета – знайти оптимальний шлях до бажаного стану (низький % відхилень).

Показник ефективності:

- Прогнозований відсоток відхилень після дії.
- Матриця помилок класифікації.
- Візуалізація метрик.

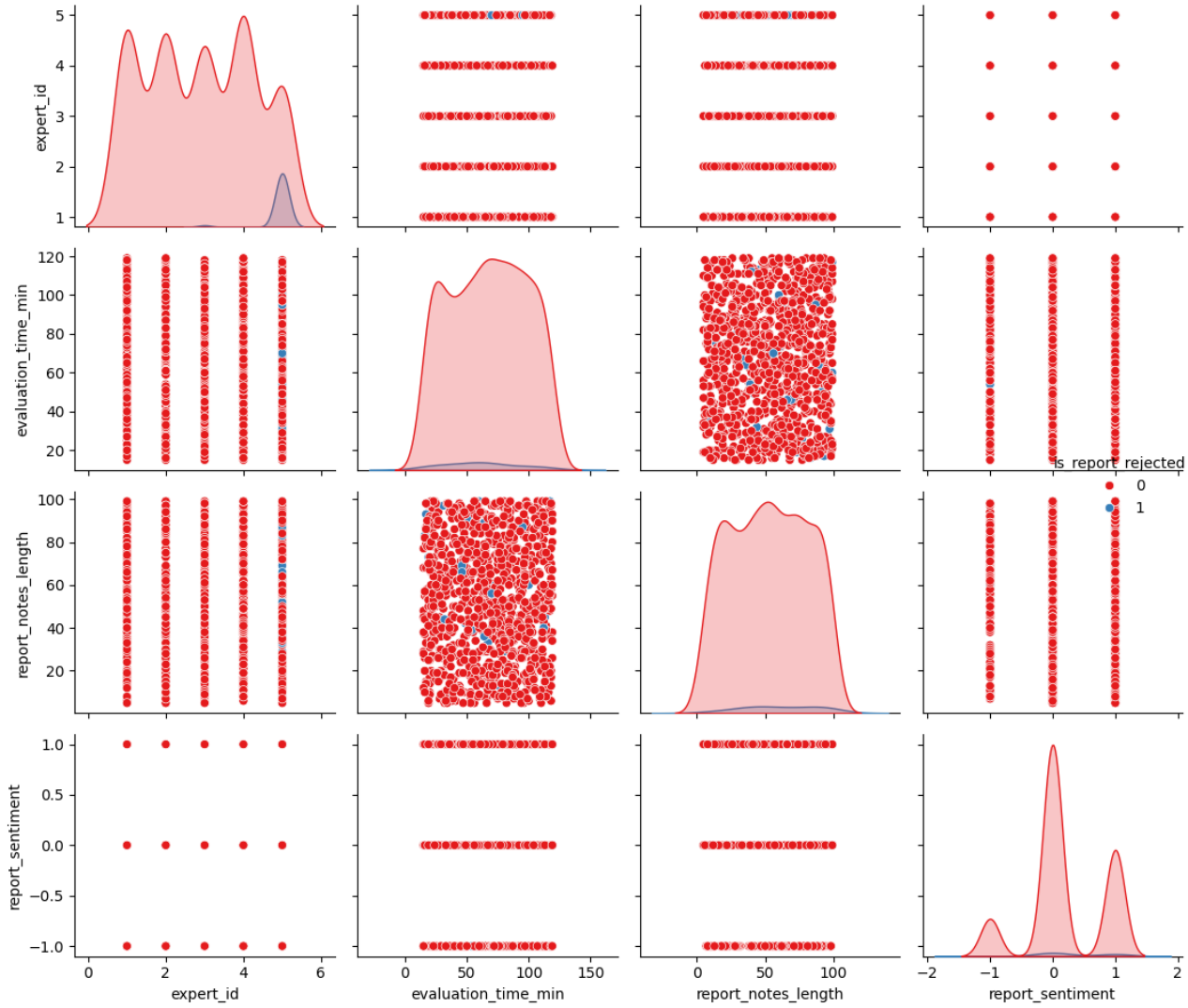
Очікуваний результат:

- Побудова класифікаційної моделі (LDA/QDA).
- Сценарне моделювання впливу управлінських дій на метрики.
- Вибір оптимальної комбінації дій для досягнення цілі.

LDA (Linear Discriminant Analysis) — це лінійна модель класифікації, яка шукає проєкцію ознак, що найкраще розділяє класи. Вона припускає однакову коваріаційну матрицю для всіх класів, що робить її стабільною при невеликих вибірках.

QDA (Quadratic Discriminant Analysis) — це нелінійна модель, яка дозволяє кожному класу мати власну коваріаційну матрицю. Вона краще працює при складних розподілах, але чутливіша до колінеарності та розміру вибірки.

Для аналізу було обрано 4 ознаки: expert_id (ID експерта), evaluation_time_min (Час оцінки, хвилини), report_notes_length (Довжина коментаря, кількість слов), report_sentiment (Тональність коментаря).



Поточний % відхилень (Базовий рівень): 3.10%.

Агент протестував 3 можливі управлінські дії для досягнення цілі (< 2.0% відхилень):

Дія	Прогнозована конверсія
retrain_expert_5	10.5%
enforce_time_policy	28.2%
implement_checklists	29.4%

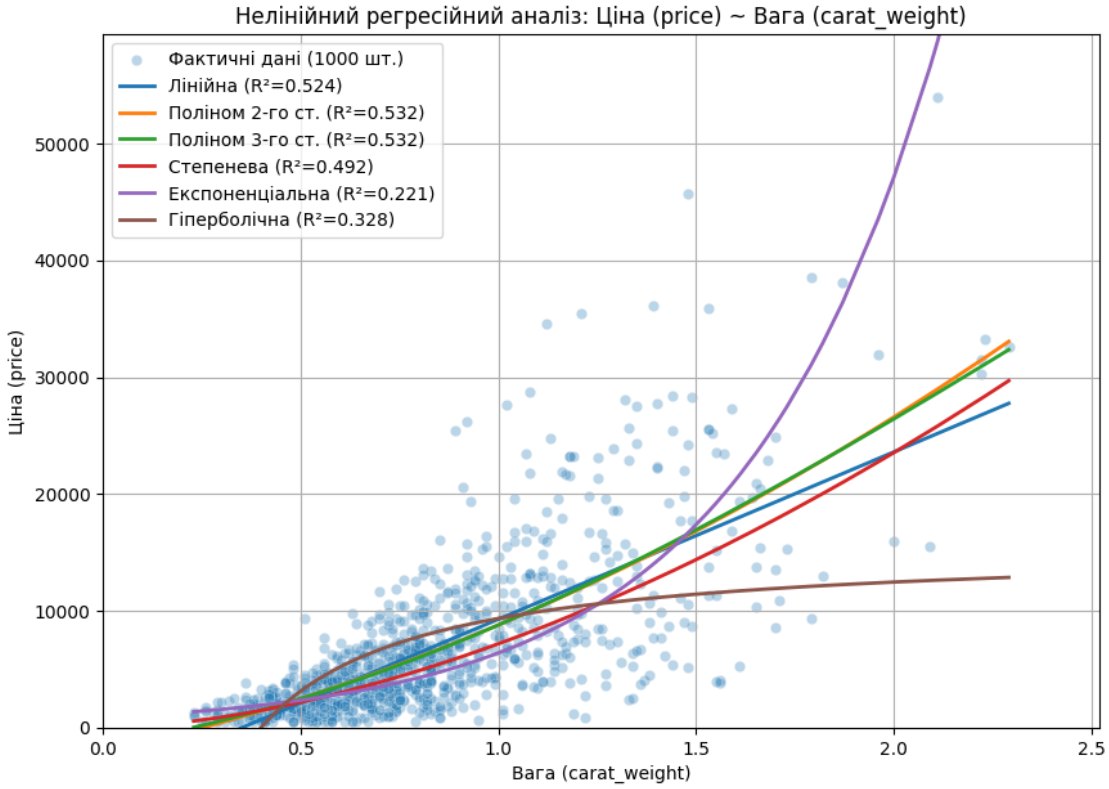
Робота агента успішна (він виконав симуляцію), але модель агента (LDA) виявився неадекватним для цієї задачі через сильний дисбаланс класів. Для реального застосування необхідно або зібрати більше даних про відхилення (клас 1), або використовувати альтернативні моделі, більш стійкі до дисбалансу (наприклад, RandomForest з class_weight='balanced').

3. Нелінійний регресійний аналіз + Дерево рішень + Нейрона мережа (3-1)

На графіку нижче зображено фактичні дані (чорні точки) та криві шости побудованих моделей:

- Лінійна (СР-2) – синя лінія
- Поліном 2-го ступеня – оранжева лінія
- Поліном 3-го ступеня – зелена лінія
- Степенева модель – червона лінія
- Експоненціальна модель – фіолетова лінія
- Гіперболічна модель – коричнева лінія

Графік дозволяє візуально порівняти, як кожна модель апроксимує залежність між Ціна (price) та Вага (carat_weight). Найкраще узгодження з фактичними даними демонструє поліном 3-го ступеня.



Вхідні данні.
Цільова змінна (Y): is_investment_grade (Бінарна: 1 - Так, 0 - Ні).
Ознаки (X) (4C):

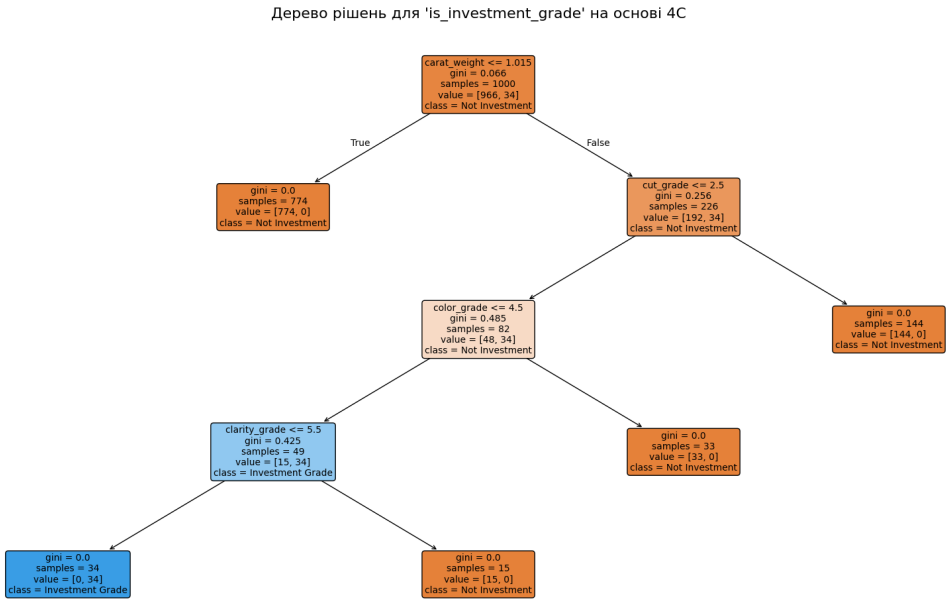
- carat_weight (Вага, ct)
- color_grade (Колір)
- clarity_grade (Чистота)
- cut_grade (Огранювання)

Побудоване дерево рішень (див. рисунок) дозволяє нам формалізувати правила, за якими діамант відноситься до інвестиційного класу. Термінологія вузлів:

- gini (Індекс Джині). Ступінь "неоднорідності" вузла (від 0.0 – "чистий" вузол, всі об'єкти одного класу, до 0.5 – максимальна неоднорідність).
- samples. Кількість зразків (діамантів) у цьому вузлі.
- value. Розподіл зразків по класах. Список: [кількість класу 0 (Not Investment), кількість класу 1 (Investment Grade)].
- class. Прогнозований клас для цього вузла (той, що має більшість у value).

Загальна логіка дерева (логіка прийняття рішень на основі дерева):

Умова	Прогноз
carat_weight <= 1.015	Not Investment
carat_weight > 1.015 AND color_grade > 4.5	Not Investment
carat_weight > 1.015 AND color_grade <= 4.5 AND clarity_grade > 5.5	Not Investment
carat_weight > 1.015 AND color_grade <= 4.5 AND clarity_grade <= 5.5 AND cut_grade > 2.5	Not Investment
carat_weight > 1.015 AND color_grade <= 4.5 AND clarity_grade <= 5.5 AND cut_grade <= 2.5	Investment Grade



Вхідні данні.
Цільова змінна (Y): is_investment_grade (Бінарна: 1 - Так, 0 - Ні).
Ознаки (X) (4C):

- carat_weight (Вага, ct)
- color_grade (Колір)
- clarity_grade (Чистота)
- cut_grade (Огранювання)

Оскільки ми використовуємо 4 ознаки (4D), ми не можемо побудувати 2D-межу класифікації, тому візуалізуємо Матрицю помилок (Confusion Matrix), яка показує точність роботи моделі.

На відміну від Дерева рішень (яке показало 100% точність), Нейронна мережа провалилася.

Хоча загальна точність (accuracy) склала 96.6%, це є ілюзією ("парадокс точності").

Аналіз Матриці помилок та звіту по класах (03_neural_network_metrics.txt) показує:

1. Модель не навчилася розпізнавати цільовий клас Investment Grade.
2. Метрики precision, recall та f1-score для класу Investment Grade дорівнюють 0.00.
3. Модель просто прогнозує Not Investment для всіх зразків, отримуючи високу точність за рахунок сильного дисбалансу класів (966 Not Investment проти 34 Investment Grade).

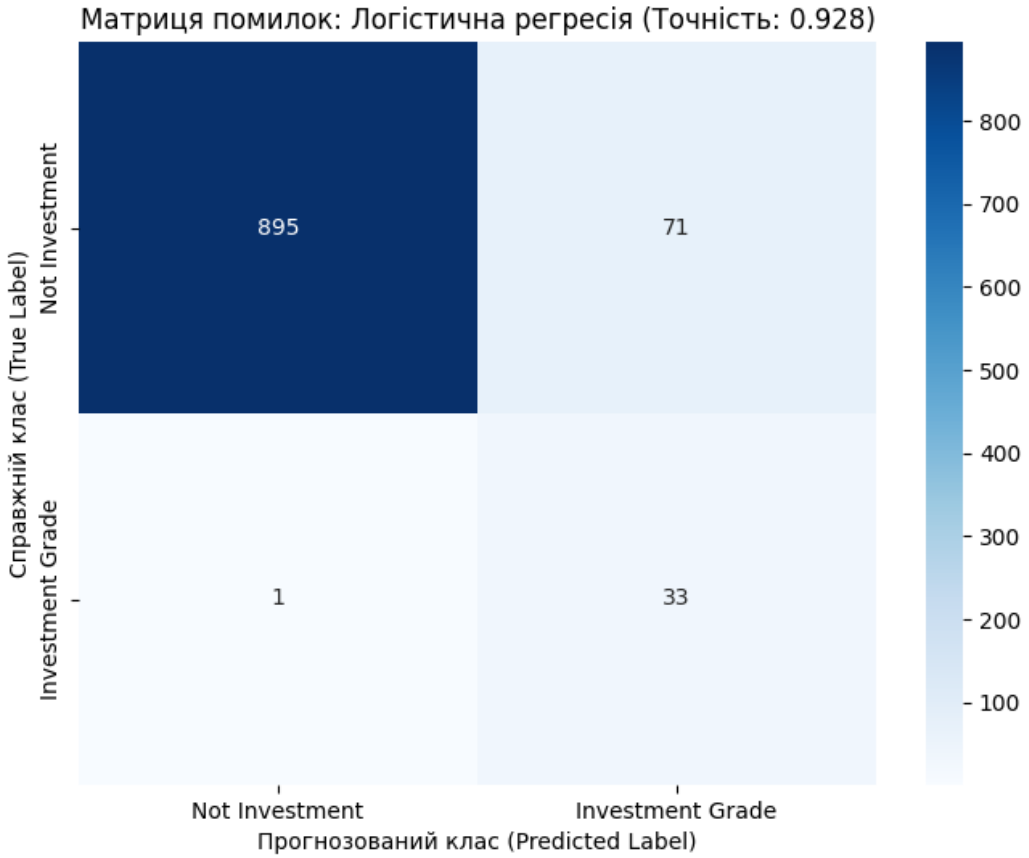
Навіть на 1000 рядках даних, але при сильному дисбалансі класів, проста нейронна мережа (MLPClassifier) з параметрами за замовчуванням виявилася неефективною. Це доводить, що для таких задач Дерева рішень є набагато більш надійним інструментом.

4. Логістична регресія

Вхідні данні.
Цільова змінна (Y): is_investment_grade (Бінарна: 1 - Так, 0 - Ні).
Ознаки (X) (4C):

- carat_weight (Вага, ct)
- color_grade (Колір)
- clarity_grade (Чистота)
- cut_grade (Огранювання)

Оскільки ми використовуємо 4 ознаки (4D), ми візуалізуємо Матрицю помилок (Confusion Matrix).



Логістична регресія з параметром class_weight='balanced' показала високу якість.

Дерево рішень (СР-3) lосягло 100% точності, а Нейронна мережа (СР-3) повністю провалилася (f1-score=0.00) , не впоравшись з дисбалансом класів.

Логістична регресія показала точність 92.8% та, що важливіше, високі precision та recall для обох класів (на відміну від нейромережі). Це демонструє, що логістична регресія є набагато більш стійкою та надійною моделлю для роботи з незбалансованими даними, ніж MLPClassifier за замовчуванням.

5. Регресія Кокса

Вхідні данні.

Мета: Оцінити, як характеристики діаманта (4C) впливають на швидкість (ризик) його продажу.

Змінна тривалості (Duration): days_on_market (Днів на ринку).

Змінна події (Event): is_sold (1 - Продано, 0 - Цензуровано).

Ознаки (X): carat_weight (Вага, ct), color_grade (Колір), clarity_grade (Чистота), cut_grade (Огранювання).

Гіпотеза: Ми очікуємо, що кращі характеристики (менші числа у color_grade, clarity_grade, cut_grade) збільшують "ризик" (hazard) продажу, тобто дозволяють каменю продатися швидше. Це означає, що ми очікуємо побачити у звіті від'ємні коефіцієнти (coef < 0) для цих ознак.

Для наочного представлення результатів Cox-регресії побудовано графік Hazard Ratios (exp(coef)) з 95% довірчими інтервалами.

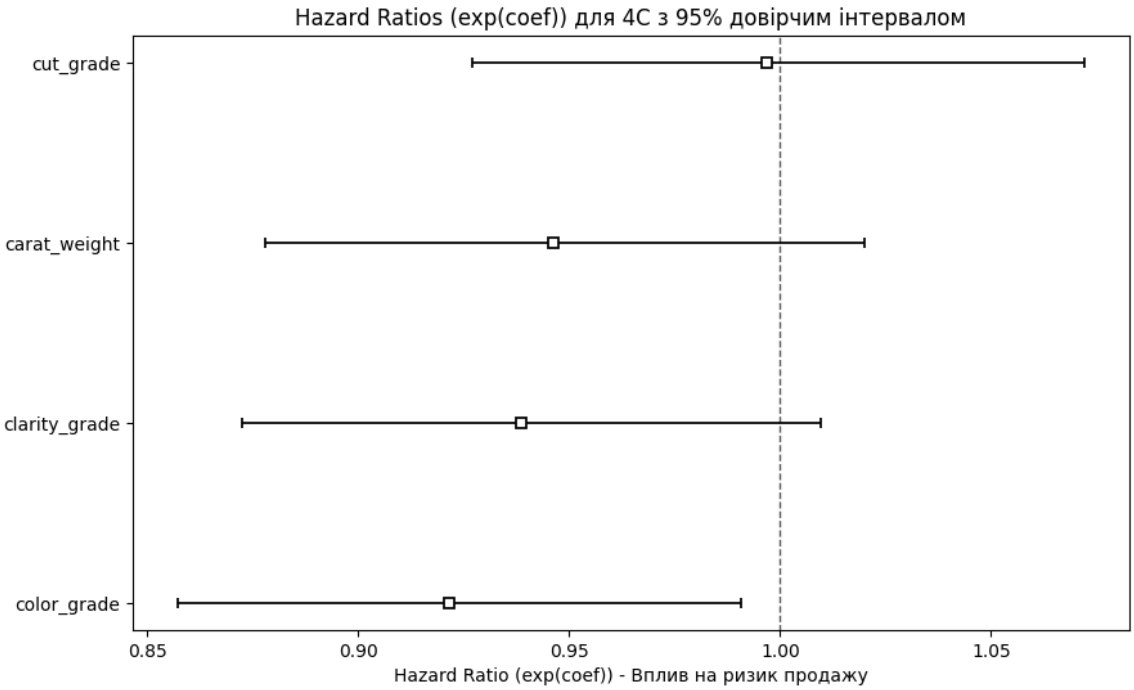
На графіку зображено чотири ознаки (наші 4C):

- carat_weight (Вага)
- color_grade (Колір)
- clarity_grade (Чистота)
- cut_grade (Огранювання)

Кожна ознака представлена точкою (оцінка exp(coef)) та горизонтальною лінією (довірчий інтервал). Вертикальна пунктирна лінія на рівні 1.0 відповідає нейтральному ефекту (відсутність впливу на ризик продажу).

Інтерпретація графіка та результатів:

- color_grade (Колір). Це єдина статистично значуща ознака (p = 0.027). Її точка (exp(coef) = 0.92) та весь довірчий інтервал [0.857, 0.991] знаходяться ліворуч від 1.0. Оскільки наш color_grade закодовано так, що 1=D (найкращий), а 10=M-Z (найгірший), це означає, що покращення кольору (зменшення числа) статистично значуще збільшує ризик (hazard) продажу – камені з кращим кольором продаються швидше.
- clarity_grade (Чистота). exp(coef) = 0.94, що також вказує на прискорення продажу при покращенні якості. Однак довірчий інтервал [0.873, 1.010] перетинає 1.0. Тому цей ефект не є статистично значущим (p = 0.089).
- carat_weight (Вага) та cut_grade (Огранювання). Довірчі інтервали обох цих ознак також перетинають 1.0 (p > 0.14 та 0.93). Їхній вплив на швидкість продажу в цій моделі статистично незначущий.


Hazard Ratios (exp(coef)) для 4C з 95% довірчим інтервалом
cut_grade
carat_weight
clarity_grade
color_grade
0.85
0.90
0.95
1.00
1.05
Hazard Ratio (exp(coef)) - Вплив на ризик продажу

12. Аналіз набору даних з використанням алгоритмів NetworkX

З датасету обрав для аналізу два типу вузлів:

- Тип 1. Експерти (5 експертів, expert_id).
- Тип 2. Категорії чистоти (10 категорій, clarity_grade).

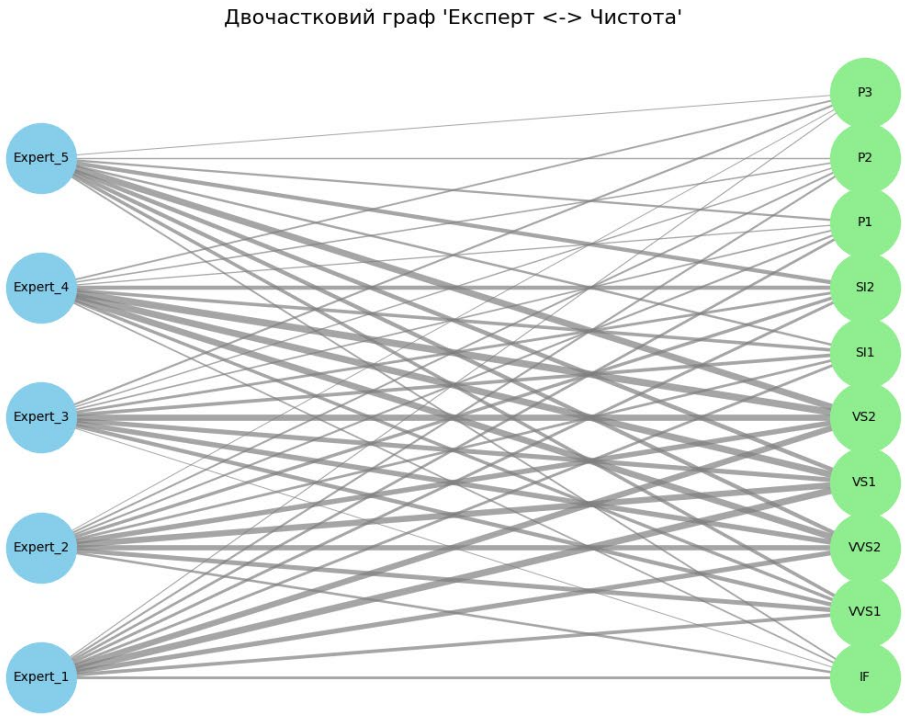
Визначив ребра (Edges). Зв'язок ("ребро") створюється, якщо експерт оцінив камінь з відповідною категорією чистоти.

Додав вагу (weight) кожному ребру. Вага дорівнює кількості разів, яку цей експерт оцінював камені цієї категорії чистоти.

Використав алгоритм на основі DFS (Обхід в глибину) для перевірки зв'язності графа (nx.connected_components).

Використав алгоритм на основі BFS (Обхід в ширину) для аналізу найкоротших шляхів між вузлами (nx.shortest_path_length).

Вузли було розділено на дві колонки (експерти зліва, чистота справа).



Отримали:

- Загальна кількість вузлів: 15 (5 + 10).
- Загальна кількість зв'язків: 50 (5 х 10).
- Знайдено 1 окремий компонент у графі – всі експерти та характеристики пов'язані в єдину мережу.
- Довжина шляху між експертами (наприклад, Expert_1 та Expert_5): 2.

Зведена таблиця (Аналог OLAP-куба):

Експерт	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	P1	P2	P3	Разом
Expert_1	15	21	31	45	39	15	16	14	10	3	209
Expert_2	13	29	33	40	33	14	19	10	9	2	202
Expert_3	2	25	33	30	41	21	15	7	5	10	189
Expert_4	7	19	42	44	47	20	25	4	6	8	222
Expert_5	8	20	25	27	43	12	25	11	5	2	178
Разом	45	114	164	186	203	82	90	46	35	35	1000

13. Практична робота зі зведеними таблицями

1. Середня ціна оцінених діамантів ($) за Експертом та Роком.

З датасету обрав для аналізу Міру:

- AVG(price)

У розрізі Вимірів:

- Експерти (5 експертів, expert_id – зробив з нього expert_name).
- Роки (3 роки, year – 2023, 2024, 2025).

Отримав зведену таблицю, яка аналізує Міру AVG(price) у розрізі Вимірів: expert_name та year.

expert_name	2023	2024	2025
Expert_1	7 101	5 557	7 000
Expert_2	8 163	6 686	7 263
Expert_3	6 042	6 143	5 124
Expert_4	6 356	5 747	6 351
Expert_5	6 597	8 922	7 101

Таблиця дозволяє швидко оцінити динаміку середньої вартості роботи кожного експерта. Наприклад, ми можемо побачити, чи зростає середня вартість каменів, що проходять через Expert_1, з року в рік, чи, можливо, Expert_3 спеціалізується на дорожчих каменях.

2. Загальна кількість продажів (шт.) за Кварталом та Походженням

З датасету обрав для аналізу Міру:

- SUM(is_sold)

У розрізі Вимірів:

- Квартал (витяг quarter з report_date).
- Походження діаманту (stone_origin – зробив з нього origin_name).

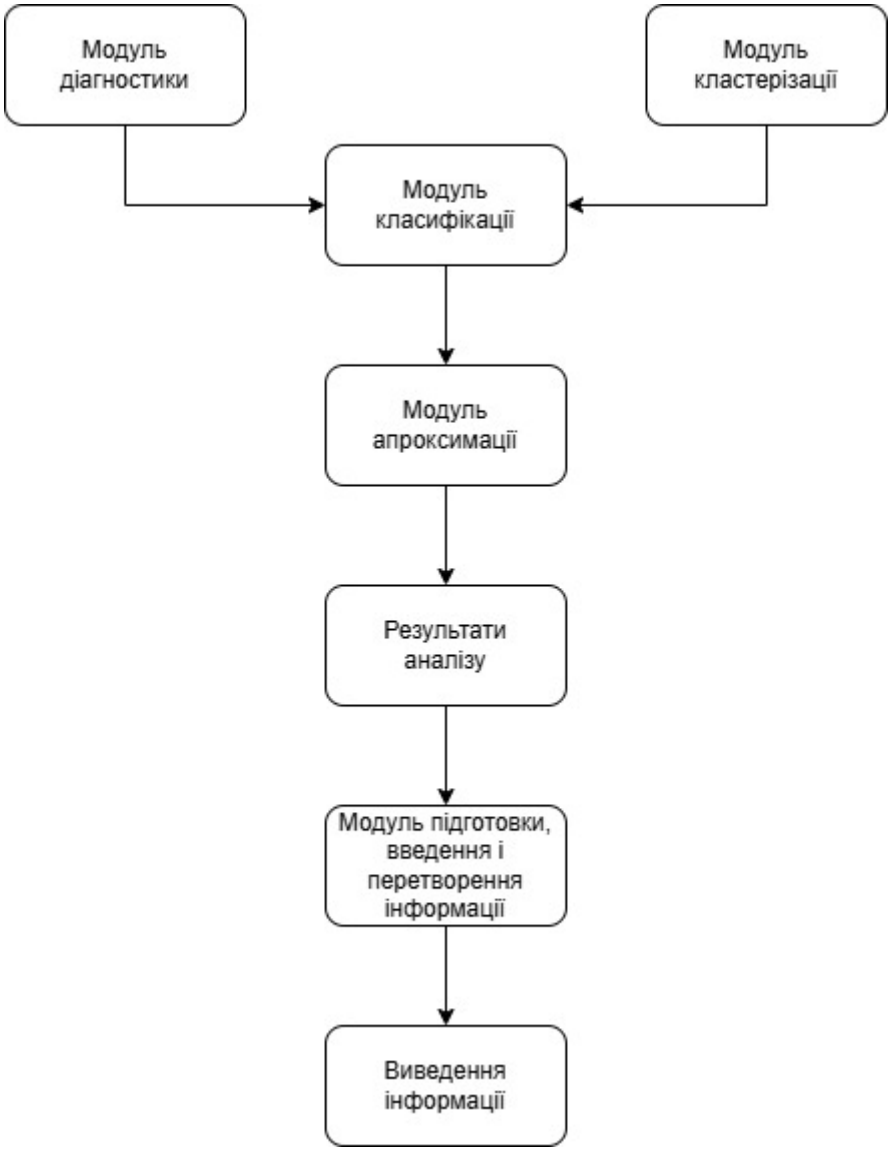
Отримав зведену таблицю, яка аналізує Міру SUM(is_sold) у розрізі Вимірів: quarter та origin_name.

quarter	Natural	Simulant	Synthetic	Treated
1	123	12	30	6
2	131	8	42	8
3	139	4	39	8
4	128	9	22	6

Таблиця демонструє сезонність продажів. Ми можемо чітко побачити, в які квартали продається найбільше Natural діамантів (очікуємо піки, наприклад, у 4-му кварталі перед святами) та як розподілені продажі Synthetic каменів протягом року.

14. Підготовка реляційної БД як джерела даних для сховища даних

Схематично процес (Flow) буде виглядати:



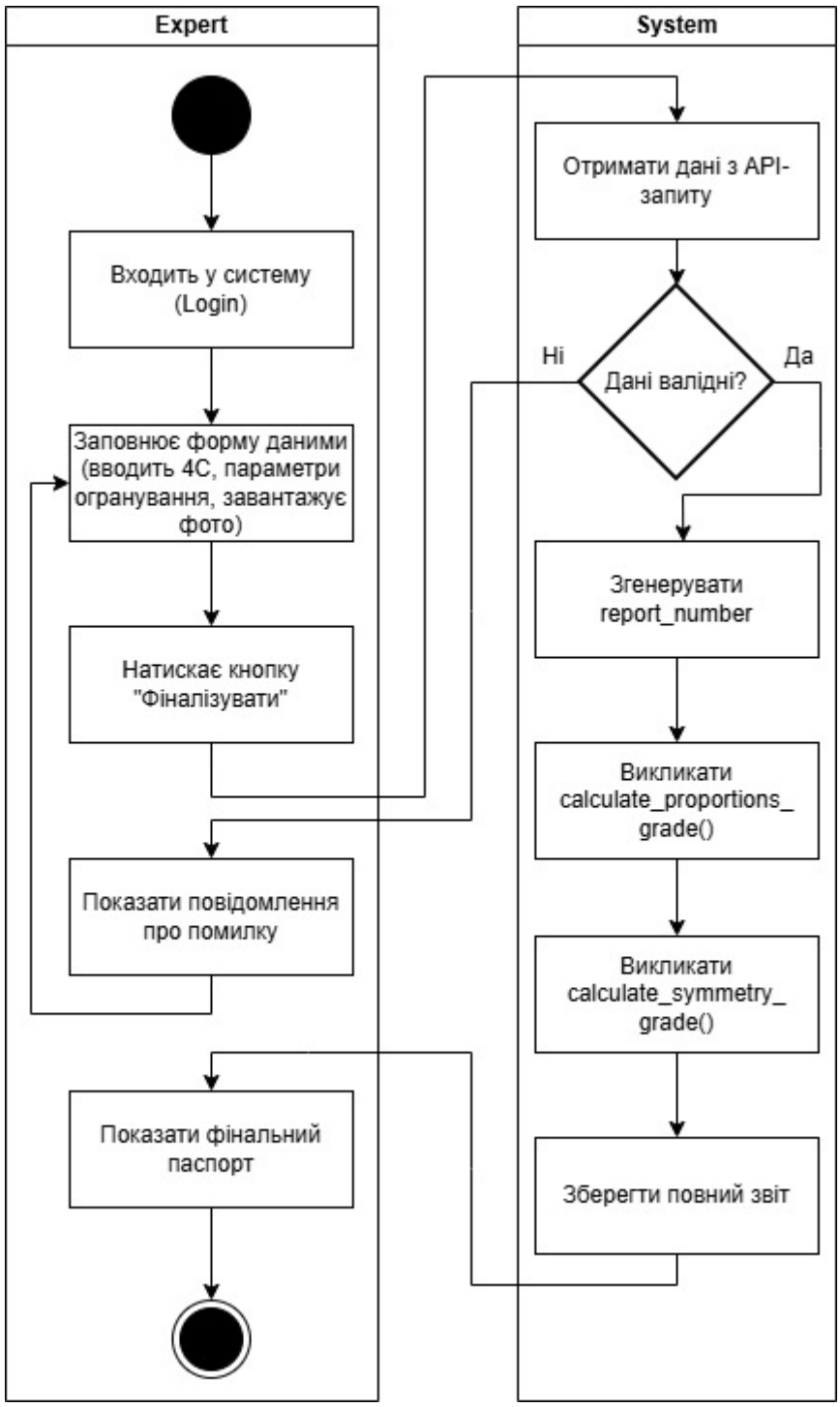
Роздивимося подрібніше Етап 1.

"Модуль діагностики" – це система на основі правил. Ми використовуємо:

- Міжнародні стандарти (IDC-Rules): Для визначення color_grade, clarity_grade.
- Розрахунковий модуль (назвемо його ReportCalculator): Логіка if/else, яка на основі crown_angle, table_width_percent тощо, присвоює proportions_grade та symmetry_grade.

Це детерміновані алгоритми (не стохастичні), вони завжди дають однаковий результат для однакових вхідних даних.

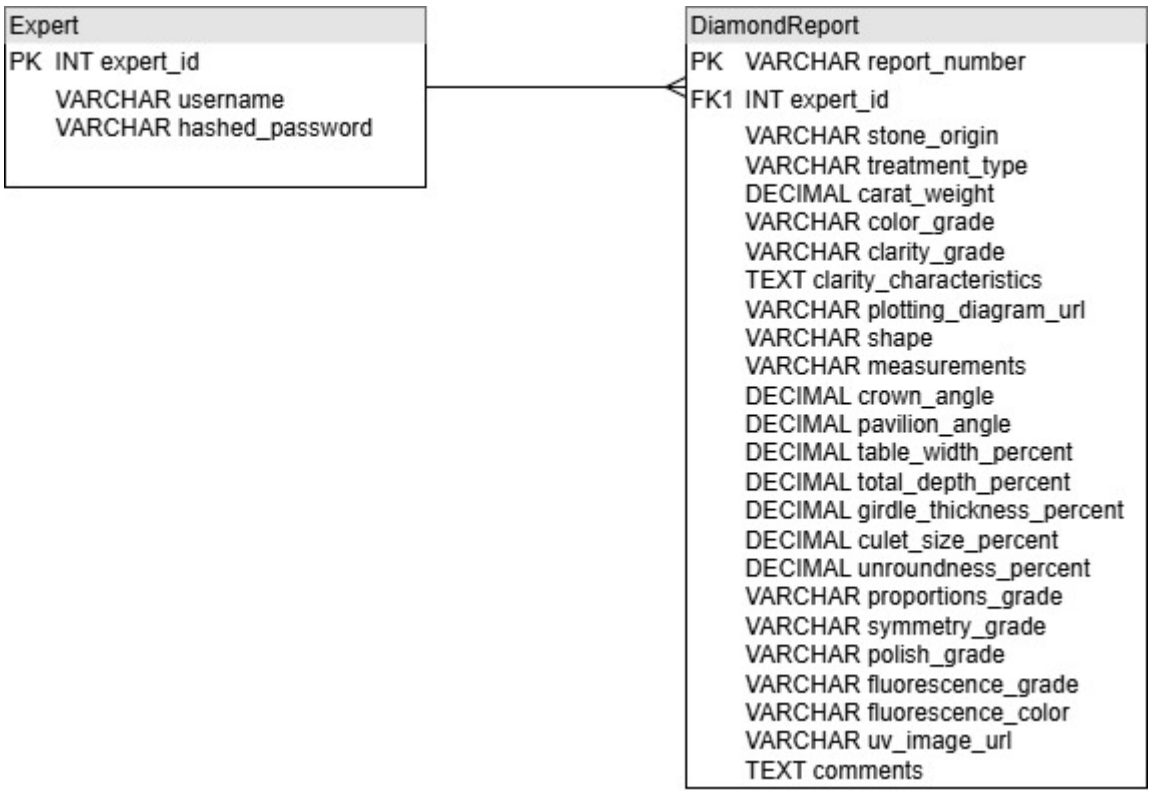
Ось як виглядає Діаграма Активності (Activity Diagram) для Етапу 1:



Можна чітко побачити OLTP-процес: "Заповнює форму даними …" -> "Зберегти повний звіт".

А ось так виглядає Діаграма сутностей (Entity-Relationship Diagram, ERD) для Етапу 1 (нормалізована структура OLTP):

- Сутність 1: Expert (Довідник експертів)
- Сутність 2: DiamondReport (Журнал звітів/транзакцій)
- Зв'язок: Один Expert може створити багато (*) DiamondReport (зв'язок "один-до-багатьох").


Expert
PK INT expert_id
VARCHAR username
VARCHAR hashed_password
DiamondReport
PK VARCHAR report_number
FK1 INT expert_id
VARCHAR stone_origin
VARCHAR treatment_type
DECIMAL carat_weight
VARCHAR color_grade
VARCHAR clarity_grade
TEXT clarity_characteristics
VARCHAR plotting_diagram_url
VARCHAR shape
VARCHAR measurements
DECIMAL crown_angle
DECIMAL pavilion_angle
DECIMAL table_width_percent
DECIMAL total_depth_percent
DECIMAL girdle_thickness_percent
DECIMAL culet_size_percent
DECIMAL unroundness_percent
VARCHAR proportions_grade
VARCHAR symmetry_grade
VARCHAR polish_grade
VARCHAR fluorescence_grade
VARCHAR fluorescence_color
VARCHAR uv_image_url
TEXT comments