Аналіз моделей дата-майнингу в розробці додатків для онлайн-торгівлі

Виконав ст. гр. ІПЗ-111М Кононенко А.В.

Керівник д.е.н., проф. Левицький С.І.

Огляд роботи

Актуальність теми. Інформаційна аналітика в сучасному світі - це міждисциплінарна галузь, що використовує сукупність прийомів та методів аналізу та вирішення завдань у різних сферах людської діяльності. Зокрема, такі системи підтримують багато бізнес-рішень — від операційних до стратегічних.

Метою дослідження є розробка програми для формування аналітичного звіту торгівельної діяльності онлайн-магазину, використовуючи наданий масив даних про покупців, покупки, а також формування передбачень про розмір доходу за певний період у майбутньому.

Об'єктом дослідження у межах цієї роботи є застосування системи створення аналітичної звітності у онлайн-магазині з використанням інструментів дата-майнингу.

Об’єктом розробки є клієнт-серверний web-додаток, що буде надавати послуги аналітичної обробки даних, перетворюючи та агрегуючи надані дані продаж, інформації про магазин, користувачів, інформацію відвідувань та перегляду товарів.

Наукова новизна. Створення універсальної схеми даних для подальшого аналізу та розробка комплексного алгоритму аналізу даних онлайн-магазину за допомогою методів та моделей дата-майнину з використанням моделей кластеризації даних та регрессії для прогнозування доходу у майбутні періоди.

Методи аналізу в сфері онлайн-торгівлі

Питання для вирішення:

1. Хто є клієнтами магазину, на які сегменти вони розбиті та що відрізняє один сегмент від іншого?
2. Які чинники впливають поведінку клієнтів, яка структура споживання?
3. Через які канали на них можна вплинути, яка віддача від цього?
4. Як виміряти лояльність, які фактори говорять про зміну тенденцій?

Методи аналізу:

1. Аналіз товару за ціновими сегментами
2. ABC-аналіз
3. XYZ-аналіз
4. RFM аналіз

Сучасні рішення в сфері онлайн-аналітики

Популярні сервіси:

1. Convead
2. Roistat
3. Calltouch
4. Google Analytics

Переваги використання:

-доступ до багатьох аналітичних інструментів. Користувальницькі змінні, API-інструменти, зведення відвідуваності, візуалізація трафіку, підтримка доступу для співробітників тощо;

-величезна кількість звітів – дашборди, статистика з трафіку, звіти по клієнтам, подіям, менеджерам, періодам, сайтам-донорам та інші.

Недоліки використання:

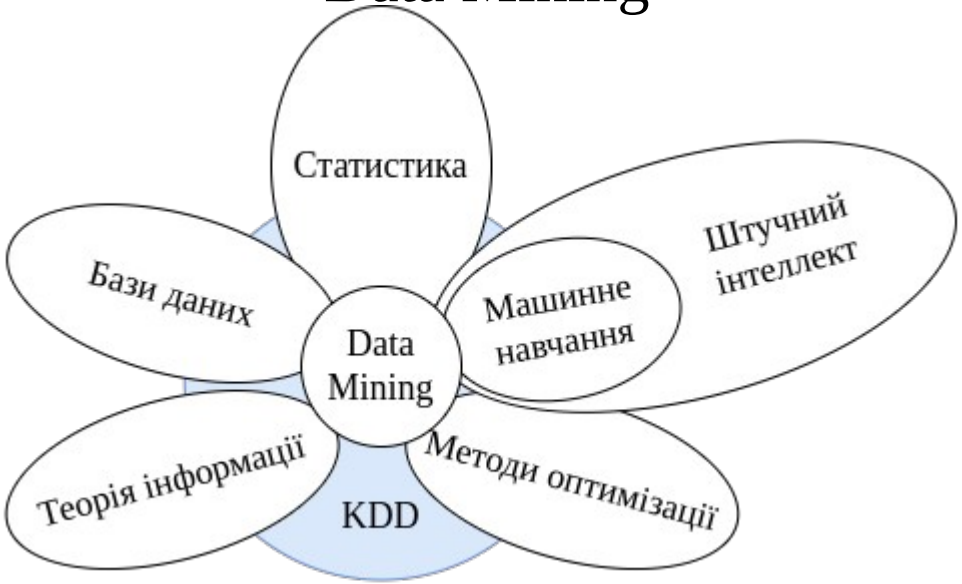
- досить складні налаштування та інтеграції;

- висока вартість користування.

- не підходить для невеликих рекламних бюджетів та стартапів;

- довгий час актуалізації статистики (від 4 годин).

Data Mining

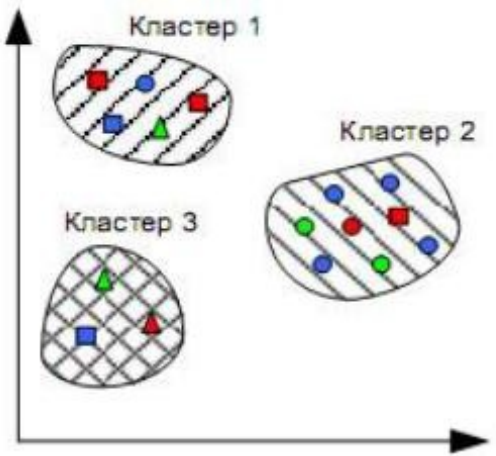


Data Mining (дата майнинг) - це методологія та процес виявлення у великих масивах даних, що накопичуються в інформаційних системах компаній, раніше невідомих, нетривіальних, практично корисних та доступних для інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності.

Процес Дата-майнингу

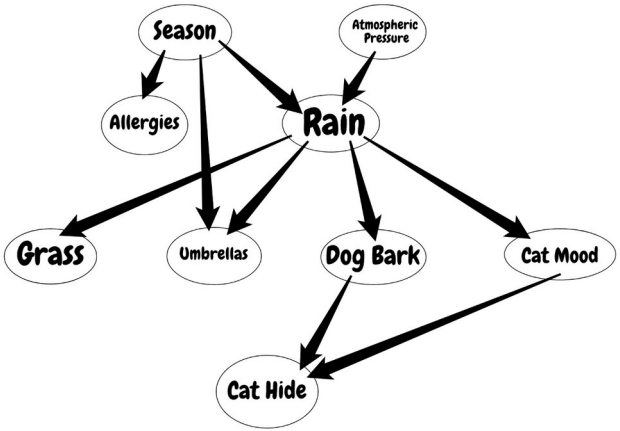
Методи Дата-майнингу

Кластерний аналіз



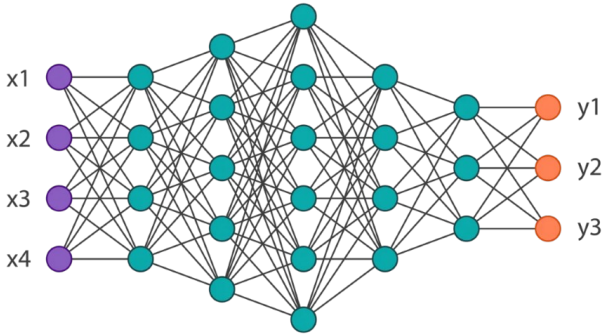
Кластери організують дані у візуальні структури

Байєсовські мережі



Структури графів представляють імовірнісні зв’язки між великою кількістю змінних

Штучні нейронні мережі



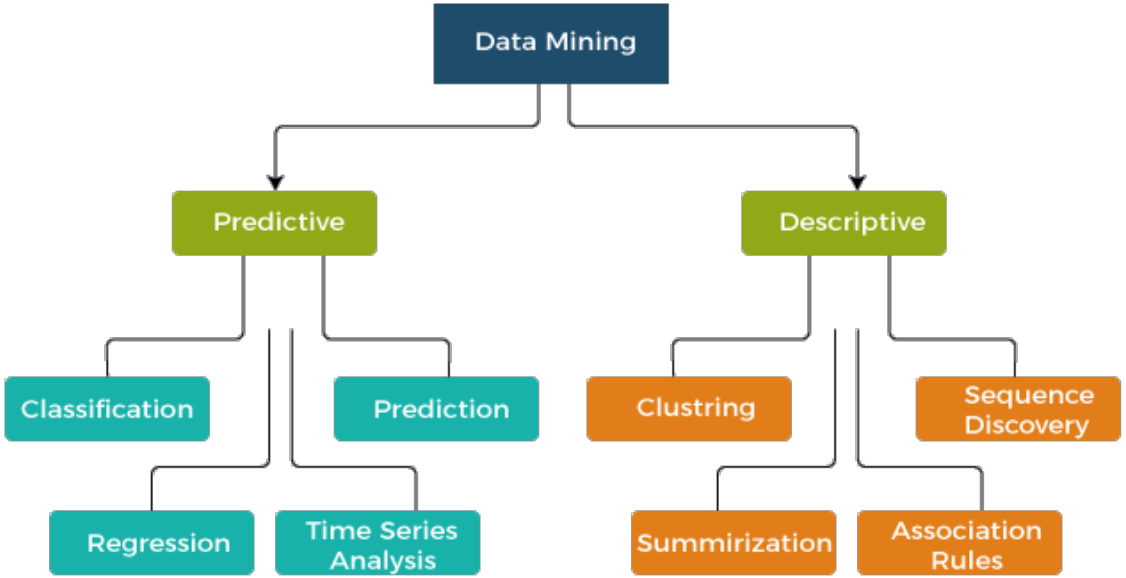
Дозволяє виконувати широкий спектр завдань аналітики після відповідного навчання

Моделі Дата-майнингу

У інтелектуальному аналізі даних математичний аналіз використовується для визначення закономірностей і тенденцій у даних. Ці закономірності та тенденції можна зібрати разом і визначити як **моделі аналізу даних**.

Моделі аналізу вирішують задачі:

1. Прогнозування
2. Ризику і ймовірності
3. Пропозиції
4. Пошуку послідовностей
5. Групування



Класифікація

Завдання щодо розробки додатку

Має бути створений веб-серверний додаток, який:

1.Має змогу прийняти масив даних онлайн-магазина про користувачів, продукти та покупки в одному з реалізованих підтримуємих додатком форматі.

2.Перетворити дані до стандартної моделі даних, до якої будуть застосовані аналітичні та статистичні алгоритми.

3.Описати алгоритм розрахування необхідних метрик та моделей дата-майнингу. Проаналізувати дані в зазначеному періоді часу, розрахувавши важливі статистичні дані, такі як: кількість покупок, топ товарів за виручкою, найбільш активні користувачі, середній дохід за день, спрогнозувати прибуток магазину на 2 місяці у майбутньому.

4.Сформувати модель даних результатів аналізу та зберегти до бази даних.

5.Створити візуальний звіт, представивши проаналізовану інформацію у вигляді графіків, таблиць та числових значень.

Інструменти розробки

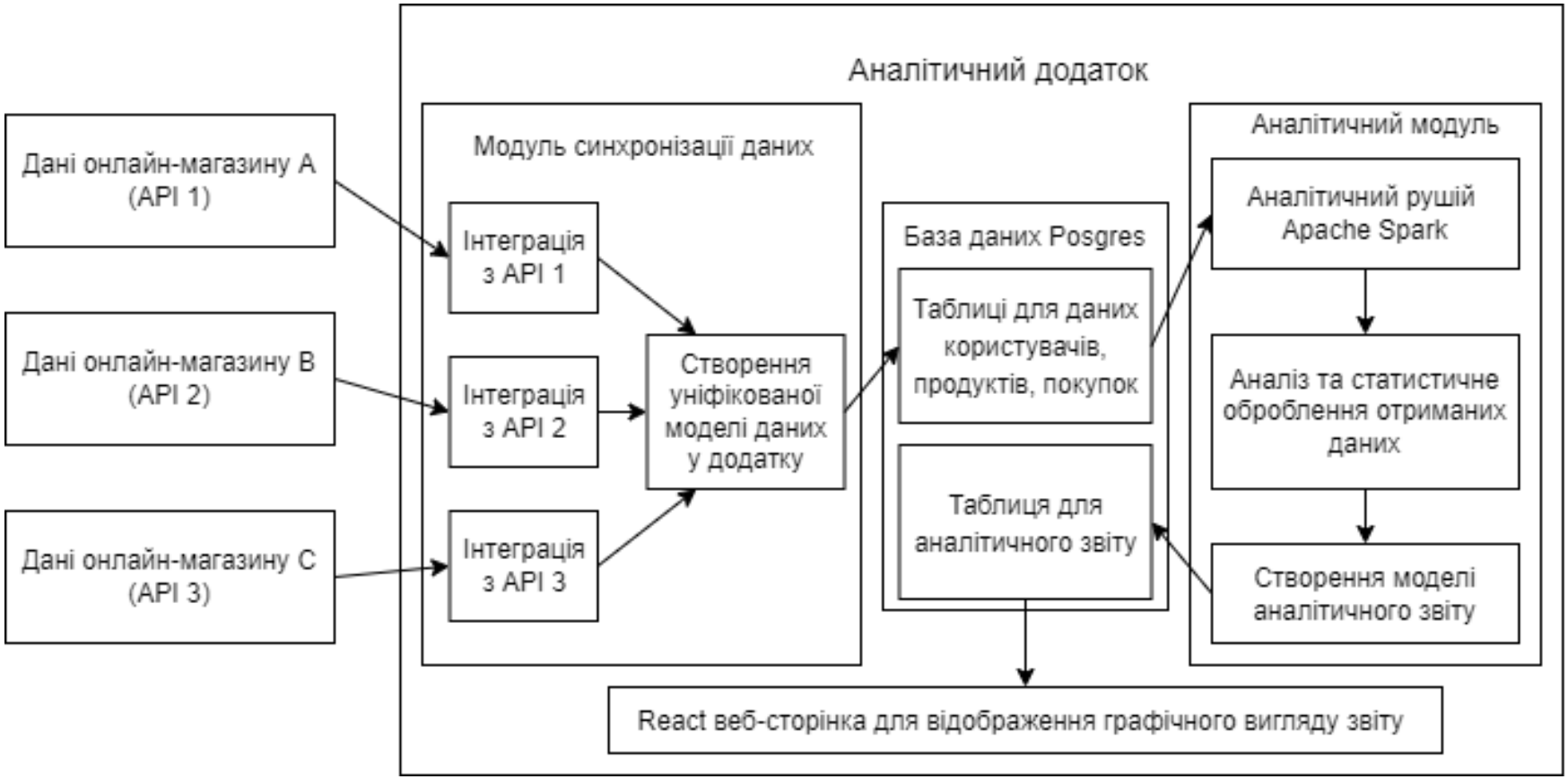
У інтелектуальному аналізі даних математичний аналіз використовується для визначення закономірностей і тенденцій у даних. Ці закономірності та тенденції можна зібрати разом і визначити як **моделі аналізу даних**.

Java

Apache Maven

Apache Spark

Структура проекту



Процес завантаження даних

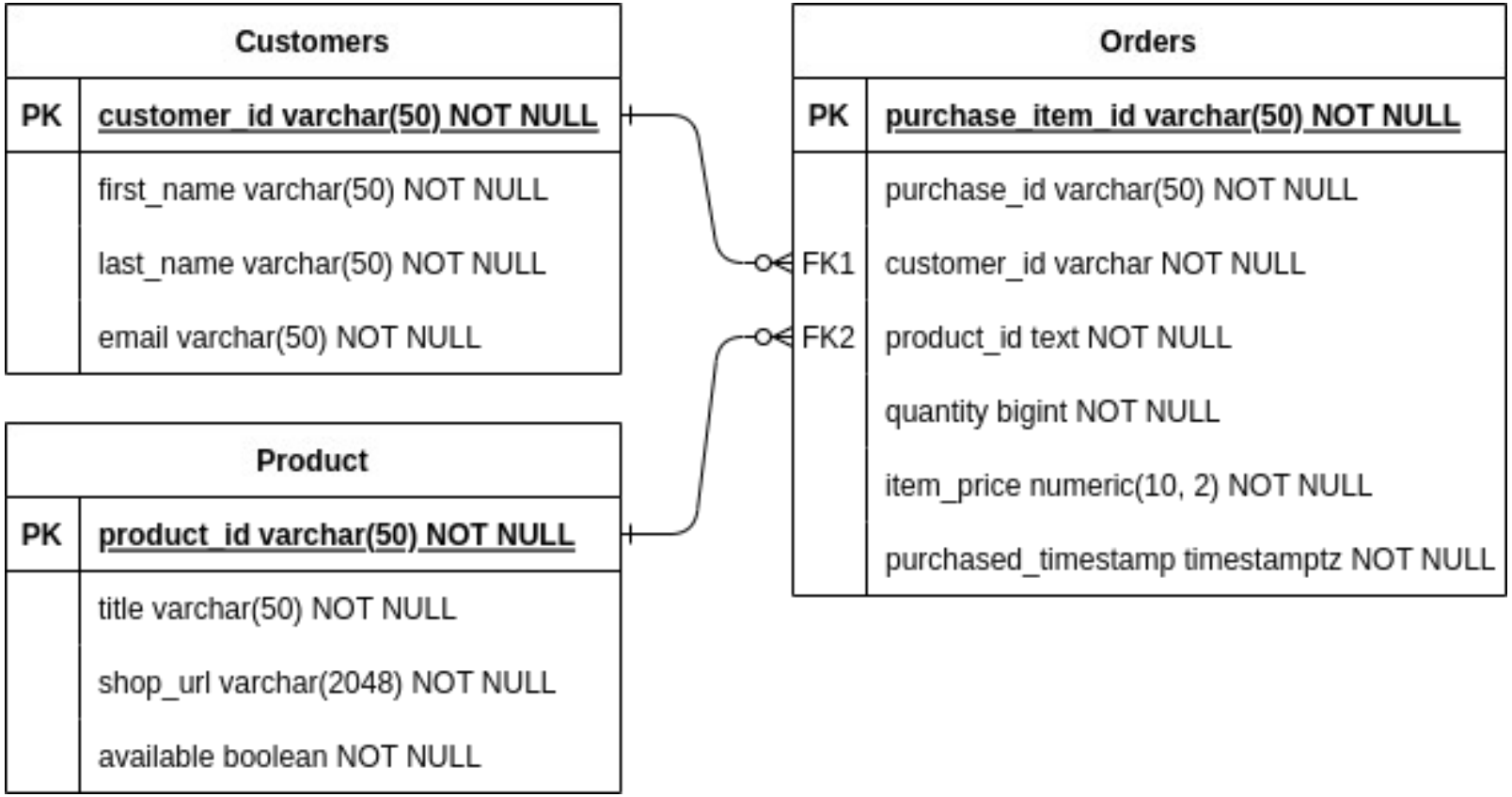
```java
@RequiredArgsConstructor
public abstract class DataSync<C extends ConnectionConfiguration, S extends AbstractConnection> {
    1 usage
    private final CustomerService customerService;
    1 usage
    private final ProductService productService;
    1 usage
    private final PurchaseItemService purchaseService;

    1 usage
    public void syncExternalData(C connectionConfiguration) {
        try (S connection = createConnection(connectionConfiguration)) {
            if (!connection.isAvailable()) {
                throw new IllegalStateException("Can't connect to external data source");
            }

            connection.syncCustomerData(customerService::saveData);
            connection.syncProductData(productService::saveData);
            connection.syncPurchaseData(purchaseService::saveData);
        } catch (Exception e) {
            throw new IllegalStateException("Data sync error!", e);
        }
    }

    1 usage  1 implementation
    protected abstract S createConnection(C connectionConfiguration);
}
```

Схема бази даних для аналізу



Метрики для розрахунку та моделі дата-майнингу

1. Кількість унікальних покупців за весь час.
2. Кількість унікальних активних користувачів за період, що аналізується.
3. Кількість унікальних найменувань товарів у магазині.
4. Кількість унікальних найменувань товарів, що були придбані за період, що аналізується.
5. Загальна кількість одиниць купленого товару.
6. Кількість успішних продажів за весь час.
7. Кількість успішних продажів за даний період.
8. Загальний дохід за даний період від продажів.
9. Середнє значення доходу від кожного продажу.
10. Середня кількість унікальних найменувань товарів у кожному продажу.

1. ABC-аналіз продуктів за правилом Парето:
   - сегмент A (80% доходу);
   - сегмент B (15% доходу)
   - сегмент C (5% доходу).
2. RFM-аналіз покупців:
   2.1 Recency (Рівні сегменти A, B, C).
   2.2 Frequency (Рівні сегменти A, B, C).
   2.3 Monetary (Рівні сегменти A, B, C).

3. Регресійна модель для прогнозування доходу:

3.1. Визначимо величину продаж за кожен день у даному періоді.

3.2. Застосуємо алгоритм лінійної регресії, аби сформувати найбільш пасуючу до даних лінію та отримати очікувані значення на майбутній період.

Реалізація розрахунку метрик

```
long uniqueCustomerCount = customerData.count();
long uniqueProductCount = productData.count();
long uniquePurchasesCount = purchaseData.select(countDistinct(PURCHASE_ITEM_ORDER_ID))
        .collectAsList().get(0).getLong( i: 0);
```

```
long activeCustomersCount = allJoinedData.select(countDistinct(PURCHASE_ITEM_CUSTOMER_ID))
        .collectAsList().get(0).getLong( i: 0);
long uniqueBoughtProductCount = allJoinedData.select(countDistinct(PURCHASE_ITEM_PRODUCT_ID))
        .collectAsList().get(0).getLong( i: 0);
long uniquePurchasesInSpecifiedPeriod = allJoinedData.select(countDistinct(PURCHASE_ITEM_ORDER_ID))
        .collectAsList().get(0).getLong( i: 0);
long boughtProductItemCount = allJoinedData.select(sum(PURCHASE_ITEM_QUANTITY))
        .collectAsList().get(0).getLong( i: 0);
Double totalValue = getDouble(allJoinedData.select(sum(PURCHASE_LINE_VALUE))
        .collectAsList().get(0), index: 0);
Double averagePurchaseValue = getDouble(allJoinedData.select(avg(PURCHASE_LINE_VALUE))
        .collectAsList().get(0), index: 0);
Double averagePurchaseItemsCount = getDouble(allJoinedData.select(avg(PURCHASE_ITEM_QUANTITY))
        .collectAsList().get(0), index: 0);
```

Реалізація розрахунку моделей

```
.withColumn(PRODUCT_SEGMENT,
        when(col(PRODUCT_VALUE_RANK).$less$eq( other: 0.8),  o: "A")
                .otherwise(when(col(PRODUCT_VALUE_RANK).$less$eq( other: 0.95),  o: "B")
                        .otherwise( value: "C"))
)
```

```
customersWithPurchasesInfo = customersWithPurchasesInfo
        .withColumn(PURCHASE_RECENCY_SEGMENT,
                when(col(LAST_PURCHASE_RECENCY).$greater$eq( other: 0.67),  o: "A")
                        .otherwise(when(col(LAST_PURCHASE_RECENCY).$greater$eq( other: 0.33),  o: "B")
                                .otherwise( value: "C"))
        )
        .withColumn(PURCHASE_FREQUENCY_SEGMENT,
                when(col(PURCHASE_FREQUENCY_RANK).$greater$eq( other: 0.67),  o: "A")
                        .otherwise(when(col(PURCHASE_FREQUENCY_RANK).$greater$eq( other: 0.33),  o: "B")
                                .otherwise( value: "C"))
        )
        .withColumn(PURCHASE_MONETARY_SEGMENT,
                when(col(PURCHASE_MONETARY_RANK).$greater$eq( other: 0.67),  o: "A")
                        .otherwise(when(col(PURCHASE_MONETARY_RANK).$greater$eq( other: 0.33),  o: "B")
                                .otherwise( value: "C"))
        );
```

```java
private List<AnalyticsReport.ChartPoint> collectChartPoints(Dataset<Row> orderDataset) {
    List<AnalyticsReport.ChartPoint> existingPoints = orderDataset
            .withColumn(DAY_TIMESTAMP_FIELD, date_trunc( s: "DAY", col(PURCHASE_ITEM_PURCHASE_TIMESTAMP)))
            .groupBy(col(DAY_TIMESTAMP_FIELD)) RelationalGroupedDataset
            .agg(sum(col(PURCHASE_LINE_VALUE)).as(PRODUCT_TOTAL_VALUE)) Dataset<Row>
            .orderBy(col(DAY_TIMESTAMP_FIELD))
            .collectAsList() List<Row>
            .stream() Stream<Row>
            .map(this::mapRowToChartPoint) Stream<ChartPoint>
            .collect(Collectors.toList());

    // calculate regression
    SimpleRegression simpleRegression = new SimpleRegression();

    for (int i = 0; i < existingPoints.size(); i++) {
        simpleRegression.addData(i, existingPoints.get(i).getValue());
    }

    LocalDate lastExistingDate = existingPoints.get(existingPoints.size() - 1).getDate();

    for (int i = 0; i < DAYS_TO_PREDICT; i++) {
        existingPoints.add(new AnalyticsReport.ChartPoint(
                lastExistingDate.plusDays( l: i + 1),
                simpleRegression.predict( x: existingPoints.size() + (double) i),
                forecast: true));
    }

    return existingPoints;
}
```

Створення об’єкту аналітичного звіту та збереження до БД

```
AnalyticsReport report = AnalyticsReport.builder()
        .customerCount(uniqueCustomerCount)
        .activeCustomerCount(activeCustomersCount)
        .uniqueProductCount(uniqueProductCount)
        .uniqueBoughtProductCount(uniqueBoughtProductCount)
        .boughtProductItemCount(boughtProductItemCount)
        .purchasesCount(uniquePurchasesCount)
        .uniquePurchasesInSpecifiedPeriod(uniquePurchasesInSpecifiedPeriod)
        .totalValue(totalValue)
        .averagePurchaseValue(averagePurchaseValue)
        .averagePurchaseUniqueItemsCount(averagePurchaseItemsCount)
        .customersWithStats(customersWithStats)
        .productsWithStats(productsWithStats)
        .revenueValueChartByDay(revenueValueChartByDay)
        .build();

analyticsReportService.save(
        new AnalyticsReportEntity(System.currentTimeMillis(),
                request.getShopId(),
                report,
                request.getPeriodStart(),
                request.getPeriodEnd()
        )
);
```

Презентація результатів

	Назва	Дохід	Сегмент
1	Дитячий конструктор Roo Crew Блоки 50 деталей	90986	C
2	Електронний конструктор Znatok Перші кроки в електроніці	133796	B
3	Конструктор BitKit Боксер 52 елементи	61238	C
4	Конструктор LEGO ART Квіткове мистецтво 2870 деталей	178134	B
5	Конструктор LEGO City Farm Тварини на фермі та у хліві 230 деталей	262654	B
6	Конструктор LEGO City Missions Детективні місії водної поліції 278 деталей	101867	C
7	Конструктор LEGO City Space Космодром 1010 деталей	558502	A
8	Конструктор LEGO City Stuntz Подвійна петля 598 деталей	706676	A
9	Конструктор LEGO City Trains Залізнична станція	251916	B
10	Конструктор LEGO City Перевезення гелікоптера 215 деталей	148958	B
11	Конструктор LEGO City Поліцейська машина 244 деталей	79827	C
12	Конструктор LEGO Classic Коробка кубиків 484 деталі	154297	B
13	Конструктор LEGO Creator Expert Будинок з привидами 3231 деталь	1586862	A
14	Конструктор LEGO Creator Expert Букет 756 деталей	455202	A
15	Конструктор LEGO Creator Expert Райський птах 1173 деталей	510769	A

1 2 3 4

Сегментація продуктів

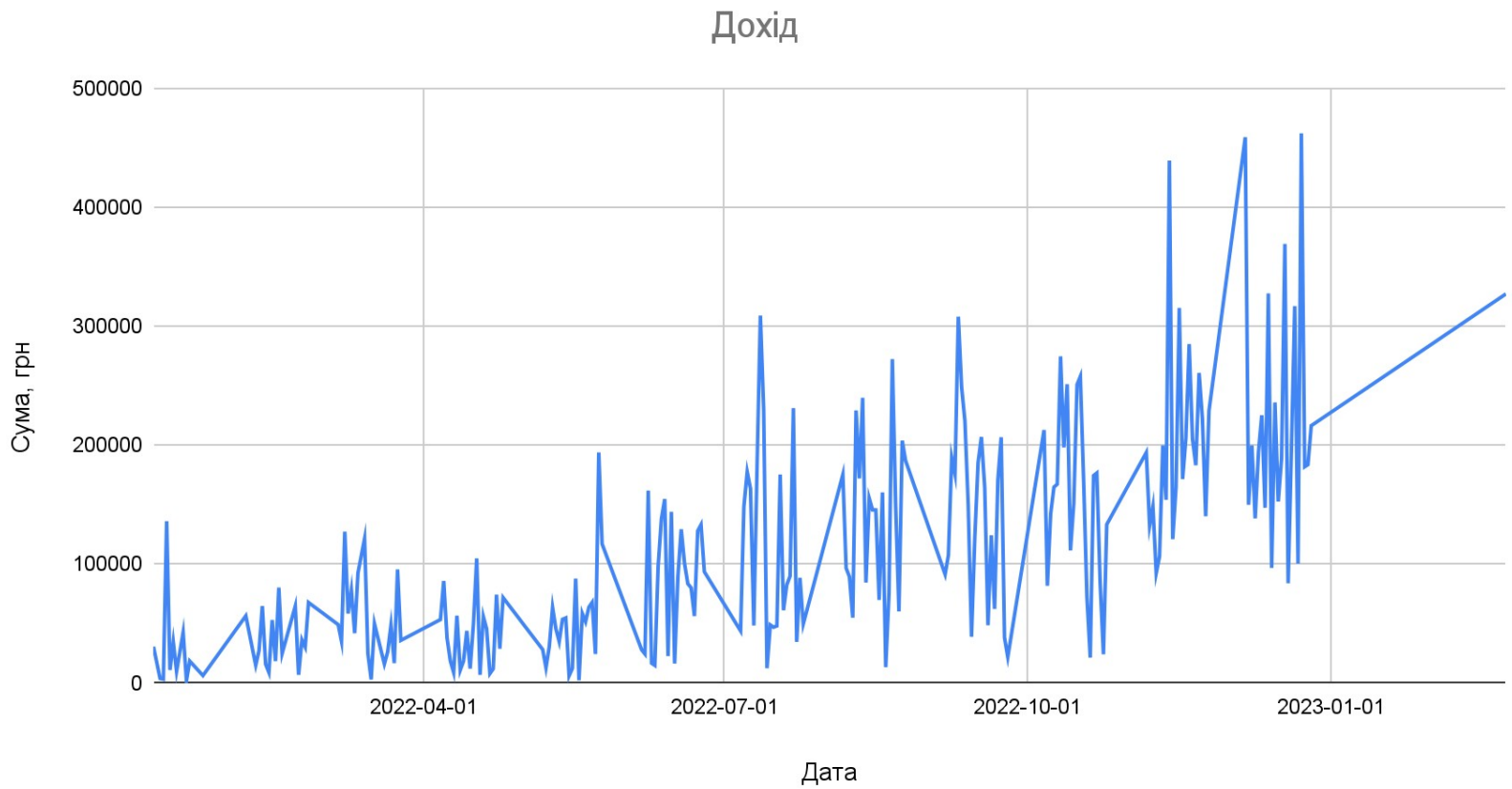
Презентація результатів

Ім'я	Прізвище	e-mail	Дохід від покупця	кількість купівель	сегмент R	сегмент F	сегмент M
Йошка	Ярешко	yoshkayareshko@gmail.com	266317	14	A	B	A
Чесмил	Янішевський	chesmylyanishevskyi@gmail.com	216568	5	B	B	B
Христина	Янушевич	khrystynayanushevych@gmail.com	109324	8	A	A	B
Яволод	Ягупольський	yavolodyahupolskyi@gmail.com	310138	13	C	C	A
Назарій	Юрченко	nazariiyurchenko@gmail.com	1037	1	A	A	C
Святолюба	Шеренгова	sviatoliubasherenhova@gmail.com	98683	4	B	A	C
Назарій	Шейко	nazariisheiko@gmail.com	213103	13	A	C	B
Тур	Чубенко	turchubenko@gmail.com	63167	4	B	A	C
Шаміль	Чикаленко	shamilchykalenko@gmail.com	118254	7	A	A	B
Ігорина	Царук	ihorynatsaruk@gmail.com	181671	11	C	B	B
Зборислав	Хряпа	zboryslavkhriapa@gmail.com	70646	7	A	B	C
Юліанія	Устенко	yulianiiaustenko@gmail.com	268264	17	B	B	A
Явір	Уляницький	yavirulianytskyi@gmail.com	202198	11	B	C	B
Захарій	Удовиченко	zakhariiudovychenko@gmail.com	114805	8	C	C	B
Кузьма	Турула	kuzmaturula@gmail.com	2880	1	A	A	C

1 2 3 4 5 6 7 8 9 10

Сегментація покупців

Презентація результатів



Доходи онлайн магазину та передбачення на наступні 2 місяці

Висновки

1. Виконано **аналіз предметної області**.
2. Проведено дослідження теоретичних основ, сучасних методик та підходів в сфері дата-майнингу.
3. Проаналізовано та дано **оцінку сучасному стану сфери послуг онлайн-аналітики**.
4. Створено **додаток на мікросервісній основі**, що генерує аналітичні звіти за визначений період, оперуючи масивами даних про покупців, товари та покупки у онлайн-магазині.
5. На основі існуючих методів та моделей дата-майнингу було розроблено **алгоритми для розрахунку важливих метрик діяльності магазину**.
6. Реалізована **модель кластеризації продуктів магазину** відповідно до ABC-аналізу та **модель кластеризації покупців** відповідно до RFM-аналізу.
7. Реалізовано **регресивну модель для прогнозування доходу магазину** у майбутні періоди.
8. Було представлено роботу алгоритму на масиві згенерованих даних.
9. Виконано аналіз отриманих результатів.