Тема: «Методи ймовірнісно-статистичного моделювання з використанням методики data mining»

Виконала студентка групи КА-44

Виходцева Олександра Олександрівна

Керівник: д.т.н., професор кафедри ММСА

Бідюк Петро Іванович

Об’єкт дослідження: розклад а/п Бориспіль (рейси, що обслуговуються компанією Interavia)

Предмет досліждення: ймовірнісно-статистичні методи моделювання з використанням data mining

Airline	FlightOut	Неделя	Год	День	Месяц	Code1	STDTime	AircraftType	ATDВремя1.2
Airline10	FlightN238	9	2014	1	3		7:00:22	737	6:52:00
Airline8	FlightN97	9	2014	1	3		10:55:58	320	10:45:00
Airline8	FlightN47	9	2014	1	3	41	17:30:26	738	18:04:00
Airline8	FlightN91	9	2014	1	3		21:15:54	738	21:04:00
Airline8	FlightN97	9	2014	2	3		10:55:00	321	10:54:00
Airline8	FlightN91	9	2014	2	3		21:15:55	737	21:00:00
Airline3	FlightN517	10	2014	3	3		10:25:00	A81	10:32:00
Airline8	FlightN97	10	2014	3	3		10:55:51	321	11:01:00
Airline8	FlightN91	10	2014	3	3		21:15:00	738	21:19:00

Постановка задачі

Потрібно:

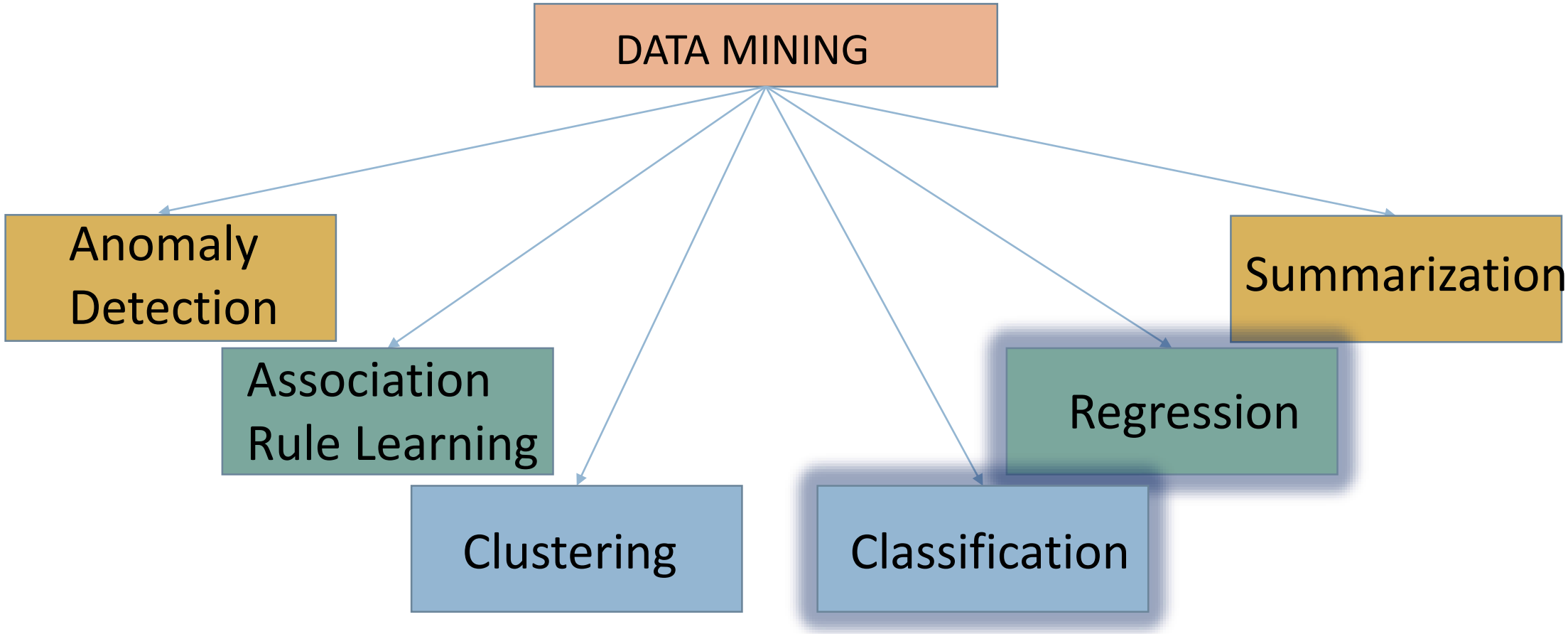
- Виявити можливі фактори, що впливають на затримку рейсів
- Зробити попередню обробку даних
- На основі даних за період 01.03.2014-28.02.2018 зпрогнозувати час затримки рейсів на 01.03.2018-05.03.2018

Актуальність дослідження

Прогнозування затримок рейсів може бути використано:

- **Компанією-хендлером** (компанією з обслуговування літаків) при плануванні робочого графіка працівників для своєчасного виявлення ситуацій, в яких потрібно долучати додатковий персонал
- **Людьми, які подорожують**, для планування маршруту з урахуванням затримок на пересадочних етапах

Класи задач Data Mining



Власний аналіз задачі

В ході дослідження предметної області було зроблено припущення, що **на час затримки рейсів впливають такі додаткові дані:**

- **Погодні умови** (загалом 10 факторів, серед яких температура повітря, видимість, напрям вітру)
- **Час доби** (ранок: 6:00-12:00, день: 12:00-17:00, вечір: 17:00-00:00, ніч: 00:00-6:00)

Обрані методи

В роботі було отримано прогнози наступними методами:

- **Дерева регресії**
- **Дерева класифікації**
- **Байесівський класифікатор**

Байесівський класифікатор застосовано двічі:

1) до набору даних, з якого виключено усі категоріальні змінні

2) до набору даних, до якого було застосовано dummy кодування

Для бінарної задачі класифікації виділено такі класи:

-Клас 0: рейс вилетів завчасно/вчасно

-Клас 1: рейс запізнився

Для задачі класифікації з трьома класами:

-Клас 1: рейс вилетів завчасно/вчасно

-Клас 2: рейс запізнився не більше ніж на 20 хвилин

-Клас 3: рейс запізнився більше ніж на 20 хвилин

Проміжні критерії якості прогнозу

Для прогнозування неперервних змінних:

- **RMSE**

$$RMSE = \sqrt{(y - \hat{y})^2}$$

Для прогнозування класу

- **Logloss**

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

- **AUC (Area Under the ROC Curve)**

Площа під ROC-кривою – графіком, що відображає залежність кількості правильно класифікованих позитивних прикладів від кількості неправильно класифікованих негативних прикладів.

Загальні критерії якості прогнозу

- **Точність, %**

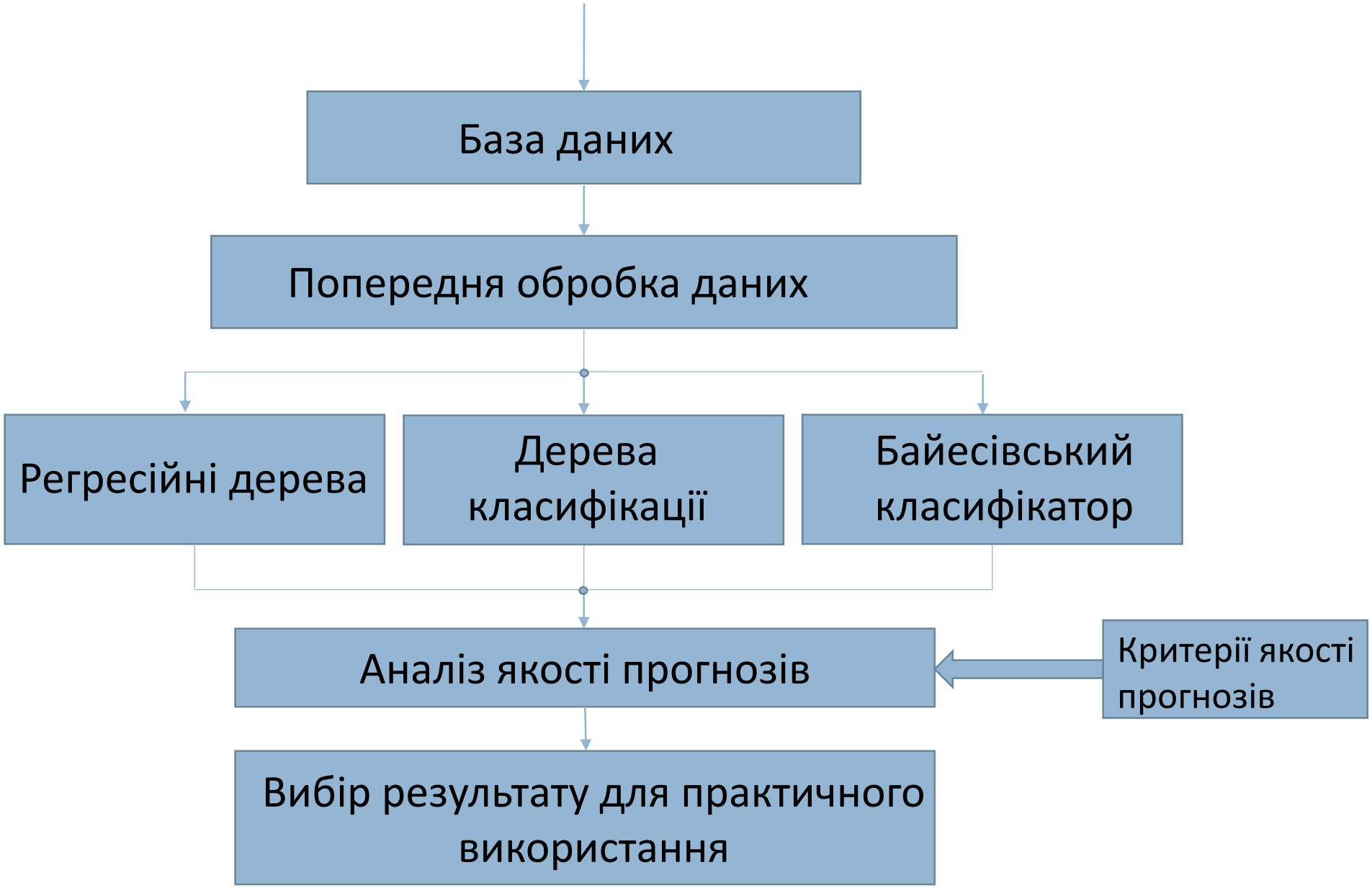
$$\text{Точність} = \frac{\text{кількість правильно класифікованих спостережень}}{\text{загальна кількість класифікованих спостережень}} * 100\%$$

- **«Погані» помилки, %**

$$\text{"Погані" помилки} = \frac{\text{к–ть спостережень, де зрогнозований клас "менше" реального}}{\text{загальна к–ть класифікованих спостережень}} * 100\%$$

- **- «Хороші» помилки, %**

$$\text{"Хороші" помилки} = \frac{\text{к–ть спостережень, де зрогнозований клас "більше" реального}}{\text{загальна к–ть класифікованих спостережень}} * 100\%$$


База даних
Попередня обробка даних
Регресійні дерева
Дерева класифікації
Байесівський класифікатор
Аналіз якості прогнозів
Критерії якості прогнозів
Вибір результату для практичного використання

Порівняння результатів прогнозування для задачі бінарної класифікації

Метод	Точність, %	«Погані» помилки, %	«Хороші» помилки, %
Регресійне дерево	90,89	9,11	0
Дерево класифікації	88,02	2,86	9,11
Байєсівський класифікатор (без категоріальних змінних)	90,1	9,9	0
Байєсівський класифікатор (з категоріальними змінними)	12,5	0	87,5

Порівняння результатів прогнозування для трьох класів

Метод	Точність, %	«Погані» помилки, %	«Хороші» помилки, %
Регресійне дерево	51,56	14,06	34,38
Дерево класифікації	30,47	3,9	65,63
Байєсівський класифікатор (без категоріальних змінних)	62,5	26,04	11,46
Байєсівський класифікатор (з категоріальними змінними)	30,99	6,5	62,5

Висновки

- Виконано огляд методів, які можуть бути використані для прогнозування затримок рейсів
- Реалізована СППР для обробки початкових даних та побудови прогнозів обраними методами.
- Для обраних методів побудовані прогнози, проведено їх порівняльний аналіз, обрано метод, який забезпечує обчислення кращого прогнозу.

Подальше дослідження

- Збір додаткових даних, які за думкою експертів можуть впливати на затримки рейсів
- Збільшення періодичності замірів погодних показників
- Групування значень категоріальних змінних за певними правилами
- Повторне впровадження в систему планування робочого графіка хендлінгової компанії

Дякую за увагу!