

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА
АКАДЕМІЯ»

Кафедра інформатики факультету інформатики



**РОЗРОБКА ДОДАТКУ ПРОГНОЗУВАННЯ ЗАТРИМОК АВІАРЕЙСІВ
ЗА ДОПОМОГОЮ SCALA**

**Текстова частина до курсової роботи
за спеціальністю „Комп’ютерні науки” 122**

Керівник курсової роботи
к. ф.-м. н. Ющенко Ю.О.

(підпис)
“ ____ ” _____ 2021 р.

Виконав студент
Крещенко Т.О.
“ ____ ” _____ 2021 р.

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ
Зав. кафедри інформатики,
доцент, к. ф.-м. н.
_____ Гороховський С.С.
(підпис)
“ ____ ” _____ 2020 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на курсову роботу

студенту Крещенку Тарасу Олександровичу факультету інформатики 4-го
курсу

ТЕМА Розробка додатку прогнозування затримок авіарейсів за допомогою
Scala

Зміст ТЧ до курсової роботи:

Індивідуальне завдання
Календарний план виконання роботи
Анотація
Вступ
1 Задача прогнозування затримок авіарейсів
2 Аналіз та обробка даних
3 Навчання моделі штучного інтелекту
Висновки
Список літератури
Додатки

Дата видачі “ ____ ” _____ 2020 р. Керівник _____
(підпис)

Завдання отримав _____
(підпис)

Тема: Розробка додатку прогнозування затримок авіарейсів за допомогою Scala

Календарний план виконання роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Отримання теми курсової роботи	16.10.2020	
2.	Перегляд літератури	16.11.2020	
3.	Аналіз способів реалізації додатку	12.01.2021	
4.	Написання першого розділу ТЧ	01.02.2021	
5.	Створення практичної реалізації	15.03.2021	
6.	Написання другого та третього розділів	01.04.2021	
7.	Написання презентації та доповіді	05.04.2021	
8.	Корегування роботи згідно із зауваженнями керівника	12.04.2021	

Студент Крещенко Т.О.

Керівник Ющенко Ю.О.

“ ”

Зміст

Анотація	5
Вступ.....	6
1 ЗАДАЧА ПРОГНОЗУВАННЯ ЗАТРИМОК АВІАРЕЙСІВ	7
1.1 Огляд проблеми.....	7
1.2 Можливі варіанти реалізації	7
1.3 Обраний варіант реалізації.....	8
2 АНАЛІЗ ТА ОБРОБКА ДАНИХ	10
2.1 Опис завдання.....	10
2.2 Отримання початкових даних.....	10
2.3 Аналіз даних	10
2.3.1 Порівняння авіаперевізників.....	11
2.3.2 Порівняння аеропортів	13
2.3.3 Вплив часу вильоту на затримку рейсу	14
2.4 Обробка даних	15
3 НАВЧАННЯ МОДЕЛІ ШТУЧНОГО ІНТЕЛЕКТУ	17
3.1 Модель для одного перевізника та одного аеропорту	17
3.2 Модель для одного перевізника та багатьох аеропортів	17
Висновки	19
Список Використої Літератури	20
Додаток А Перелік прийнятих скорочень	21

Анотація

У роботі розглянуто проблему прогнозування затримок авіарейсів, запропоновано способи її реалізації. Розроблено додаток на мові Scala та фреймворку Apache Spark, що проводить аналіз та обробку даних, а також навчає модель лінійної регресії для прогнозування затримки авіарейсів.

Ключові слова: машинне навчання, прогнозування, затримки авіарейсів, Scala, Apache Spark, лінійна регресія, аналіз даних

Вступ

Авіап перевезення – найпопулярніший спосіб подорожування на великі відстані. Проте не секрет, що дуже часто плани руйнуються саме через відтермінування авіаперельоту. Саме тому ідея прогнозування затримок авіарейсів є дуже актуальною, оскільки запровадження такої системи має можливість зробити життя трішечки кращим.

Метою курсової роботи є проаналізувати способи розв’язку задачі передбачення затримок авіарейсів, визначити переваги і недоліки використання мови Scala для задач машинного навчання, та розробити додаток, що прогнозує затримки авіарейсів.

Робота складається з трьох розділів.

У першому розділі проаналізовано проблему прогнозування затримок авіарейсів та доведено її актуальність. Запропоновано декілька можливих варіантів її розв’язку, та визначено обраний варіант розв’язку.

В другому розділі продемонстровано процес отримання, аналізу та обробки даних для використання у моделі машинного навчання. Обґрунтовано, які висновки було зроблено з результатів аналізу даних та чому.

Третій розділ присвячено навчанню моделі лінійної регресії. У ньому продемонстровано два кроки розробки моделі, та показано результат навчання.

1 ЗАДАЧА ПРОГНОЗУВАННЯ ЗАТРИМОК АВІАРЕЙСІВ

1.1 Огляд проблеми

У сучасному світі авіаперевезення відіграють неймовірно важливу роль. Такі речі як подорожі або бізнес-зустрічі є настільки великою частиною нашого життя саме завдяки можливості подорожувати швидко. Навіть під час світової пандемії щорічно відбуваються десятки мільйонів авіаперевезень [1].

При цьому, попри надзвичайну корисність та зручність, авіаперельоти мають ряд недоліків. Поміж них, завжди присутня можливість відтермінування рейсу. Це спричинено тим, що існує величезна кількість чинників, які можуть вплинути на переліт. Це такі речі як погодні умови, складність управління повітряним трафіком, велика кількість пасажирів, що можуть затриматися, та навіть присутність пташок на злітній смузі [2].

Водночас, доволі відомо, що планувати авіаподорож потрібно заздалегідь, і при цьому важливо мати чіткий розклад подій. Як же тоді планувати, якщо переліт може бути перенесений на 30 хвилин, годину, або навіть декілька годин?

Це сильно ускладнює задачу мандрівника. Тому, якби існувала система, що надає приблизні передбачення затримки авіарейсу, планувати подорожі було б хоча б трохи простіше. Розгляньмо декілька ідей щодо реалізації такої системи.

1.2 Можливі варіанти реалізації

Коли ми потребуємо наближення якогось невідомого значення, доволі часто корисними можуть бути деякі статистичні дані. Тому на думку може прийти ідея зібрати різні корисні дані про авіарейси, та проаналізувати з них ті, що стосуються конкретного випадку. Це може бути середня затримка від оператора перельоту, графік розподілу затримок всіх рейсів на поточному аеропорті, і т. п. Такий підхід є, мабуть, найпростішим, і досягає непоганого результату в передбаченні відтермінування. Проте він є не найкращим, бо

лише надає уявлення про можливу затримку, а не дає певне конкретне значення.

З математичної точки зору, ідеальним розв'язком було б створити певну формулу чи модель, що зможе врахувати всі впливові чинники та поєднати їх, маючи в результаті доволі точне передбачення. Але, на жаль, в природі існує «ефект метелика», що робить таку задачу практично неможливою. Можливо погані погодні умови можна передбачити, приєднавши систему до певної системи метеорологічної станції, все одно існують чинники, що неможливо відслідкувати та передбачити. Навіть поломка кавової машини на літаку може спричинити відтермінування польоту на декілька годин [2].

Проте існує підхід, що зможе досягти не набагато гірших результатів, ніж попередній, вимагаючи набагато меншої кількості зусиль. Це – машинне навчання. Тут все, що вимагається, – це зібрати потрібні дані, проаналізувати та обробити їх, та навчити модель.

1.3 Обраний варіант реалізації

В даній роботі було обрано реалізовувати модель машинного навчання. Зазвичай для такої задачі використовується мова Python. Справді, наявність величезної кількості якісних бібліотек для роботи з даними та машинним навчанням робить цю мову чи не найкращим варіантом.

Але і тут є свої недоліки, оскільки Python – інтерпретована мова з динамічною типізацією, через що інтенсивна робота з даними використовує більше ресурсів та займає більше часу, ніж це було б на компільованій мові зі статичною типізацією.

Тут приходить в нагоді Scala. По-перше, вона є компільованою мовою зі статичною типізацією, що призводить до більшої швидкодії. По-друге, найпопулярніший фреймворк для роботи з розподіленими обчисленнями Apache Spark був написаний саме на Scala, що робить її ідеальним вибором для роботи з дуже великою кількістю даних.

Отже, було використано мову Scala, фреймворк Apache Spark, бібліотека `evilplot` для генерації графіків та діаграм, а також бібліотека `MLlib` для тренування моделі машинного навчання.

2 АНАЛІЗ ТА ОБРОБКА ДАНИХ

2.1 Опис завдання

Перед тим, як навчати модель, потрібно виконати такі кроки: завантажити «сирі» дані з достовірного джерела, проаналізувати їх, визначити поля, що можуть бути корисними для навчання, обробити ці поля, та видалити непотрібні.

2.2 Отримання початкових даних

Для отримання даних було обрано сайт Бюро транспортної статистики США, де надається доступ до баз даних зі статистикою різних видів транспорту. Було обрано базу даних «Airline On-Time Performance Data», що містить дані про кожен авіарейс у США, а саме: дата та час відправлення, ID перевізника, місце відправлення, місце прибуття, затримка відправлення та прибуття, та інше. Далі дані було завантажено у форматі CSV.

Initial Dataset Summary

summary	year	month	day	airline	origin	dest	depTime	depDelay	arrTime	arrDelay	distance
count	638649	638649	638649	638649	638649	638649	19550	619547	618710	617649	638649
mean	2019.0	1.0	15.95	null	12690.34	12690.21	1332.73	9.99	1483.04	4.71	765.25
stddev	0.0	0.0	8.98	null	1520.77	1520.7	493.55	49.56	520.81	52.09	579.91
min	2019	1	1	9E	10135	10135	1	-47.0	1	-85.0	31.0
25%	2019	1	8	null	11292	11292	924	-6.0	1104	-16.0	337.0
50%	2019	1	16	null	12889	12889	1329	-3.0	1515	-7.0	604.0
75%	2019	1	24	null	14057	14057	1737	5.0	1917	7.0	1008.0
max	2019	1	31	ZW	16218	16218	2400	1690.0	2400	1707.0	4983.0

Рисунок 2.2.1 – Опис датасету

На рисунку 2.2.1 можна детальніше побачити, в якому форматі містяться дані. Під час розробки використовувалися дані лише січня 2019 року, бо надмірна кількість даних має властивість сильно сповільнювати процес розробки.

2.3 Аналіз даних

Мета проведення аналізу даних – зрозуміти, від яких даних залежить затримка рейсу.

2.3.1 Порівняння авіаперевізників

Щоб отримати загальне уявлення, як відрізняються затримки рейсів в різних перевізників, було згенеровано графік середніх затримок (див. Рисунок 2.3.1.1).

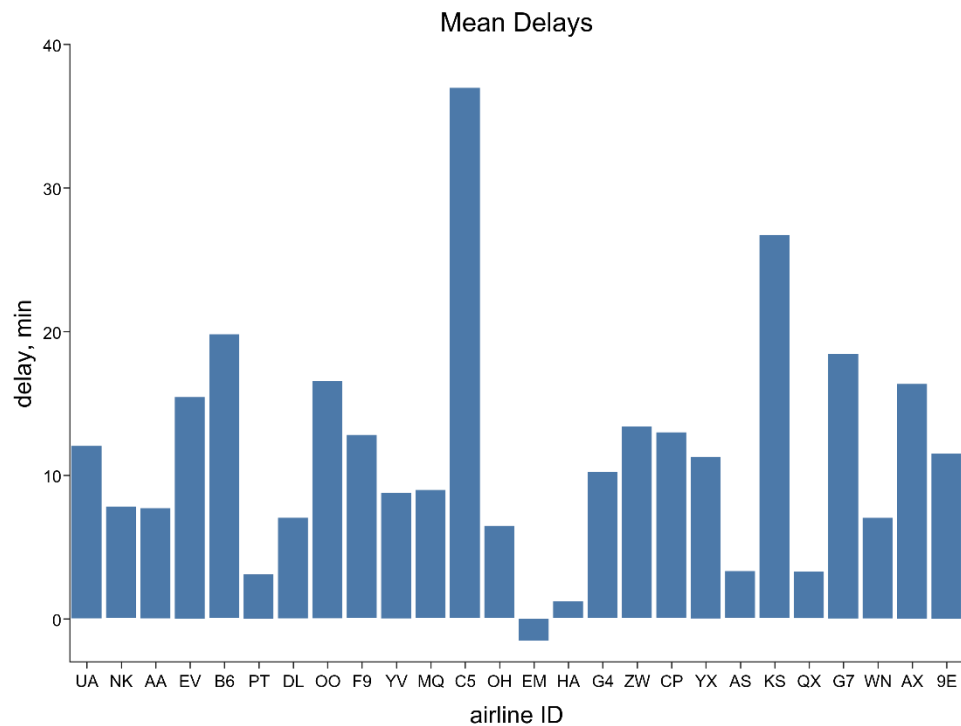


Рисунок 2.3.1.1 – Середні затримки авіаперевізників

Можемо бачити, що середні затримки варіюються від -3 до 37 хвилин, причому більшість з них менші 15 хвилин. Наявність перевізників, у яких затримка кардинально відрізняється від інших, говорить про те, що перевізники по-різному відносяться до затримання рейсів, а отже, ці дані є корисними.

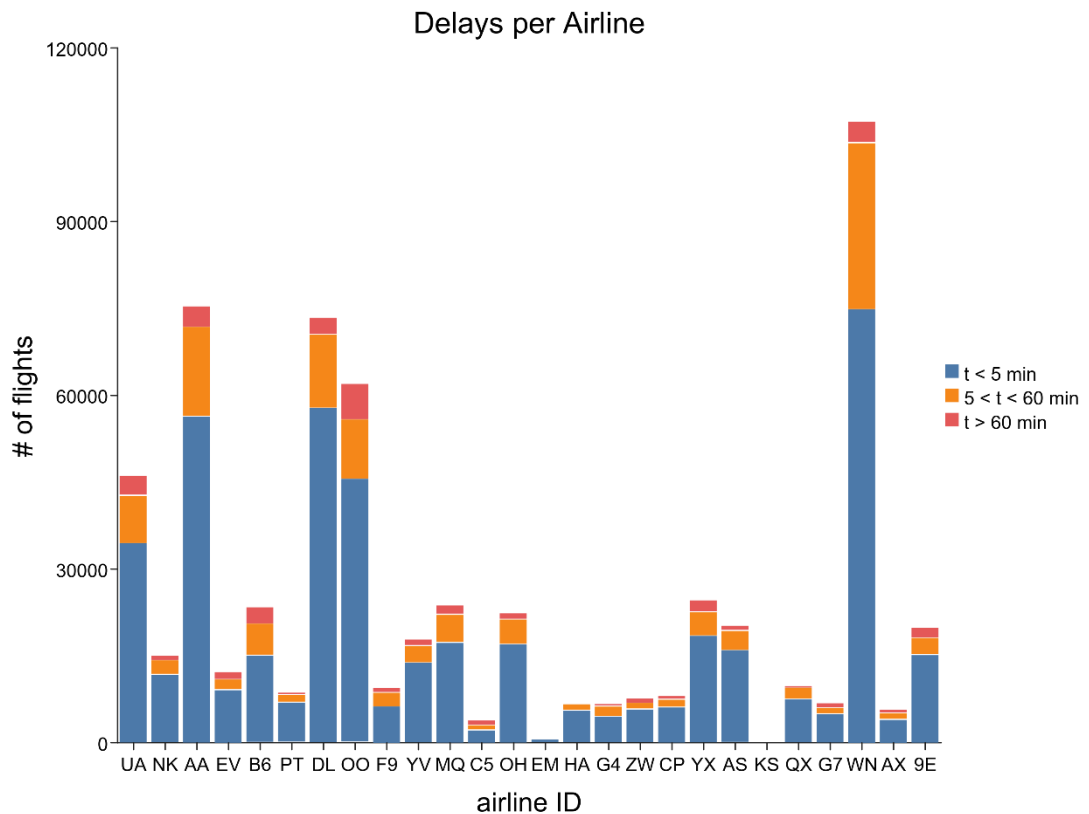


Рисунок 2.3.1.2 – Кількість різних затримок авіаперевізників

На рисунку 2.3.1.2 зображена кількість малих затримок до 5 хвилин, середніх – від 5 до 60 хвилин, та великих затримок – від 60 хвилин. Варто звернути увагу, що різні перевізники можуть мати дуже велику різницю у кількості рейсів. Можемо бачити, що різні перевізники мають різні відношення малих, середніх та великих затримок. А саме: перевізник «WN» має приблизно в 8 раз менше великих затримок, ніж середніх, а перевізник «OO» – вдвічі менше. Це лише підтверджує припущення, що дані різних перевізників відрізняються, а тому є корисними.

Порівнявши з попереднім графіком, можна зробити висновок, що мала кількість даних може спричинити нестандартну середню затримку. Це видно по перевізникам «C5», «EM» та «KS». Це, насамперед, спричинено тим, що існує невеликий шанс, що рейс буде затримано на кілька десятків годин, що дуже відрізняється від інших даних, та може кардинально змінити середнє значення. Це означає, що при навчанні моделі потрібно видалити дані, що містять такі затримки.

2.3.2 Порівняння аеропортів

Тепер спробуємо порівняти різні аеропорти та встановити, чи залежить від них затримка авіарейсів.

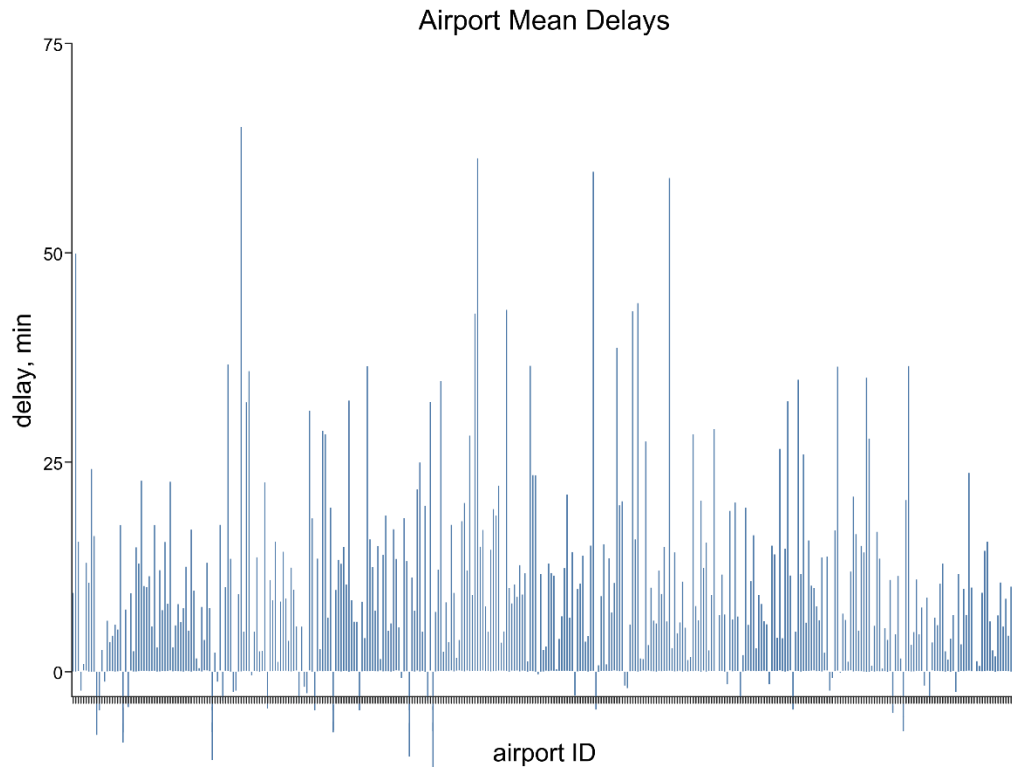


Рисунок 2.3.2.1 – Середні затримки на різних аеропортах

На рисунку 2.3.2.1 можемо спостерігати результат, схожий на результат рисунку 2.3.1.1. А саме, різні аеропорти мають радикальну різницю в середніх затримках. Тому робимо висновок, що ці дані також треба врахувати під час навчання моделі.

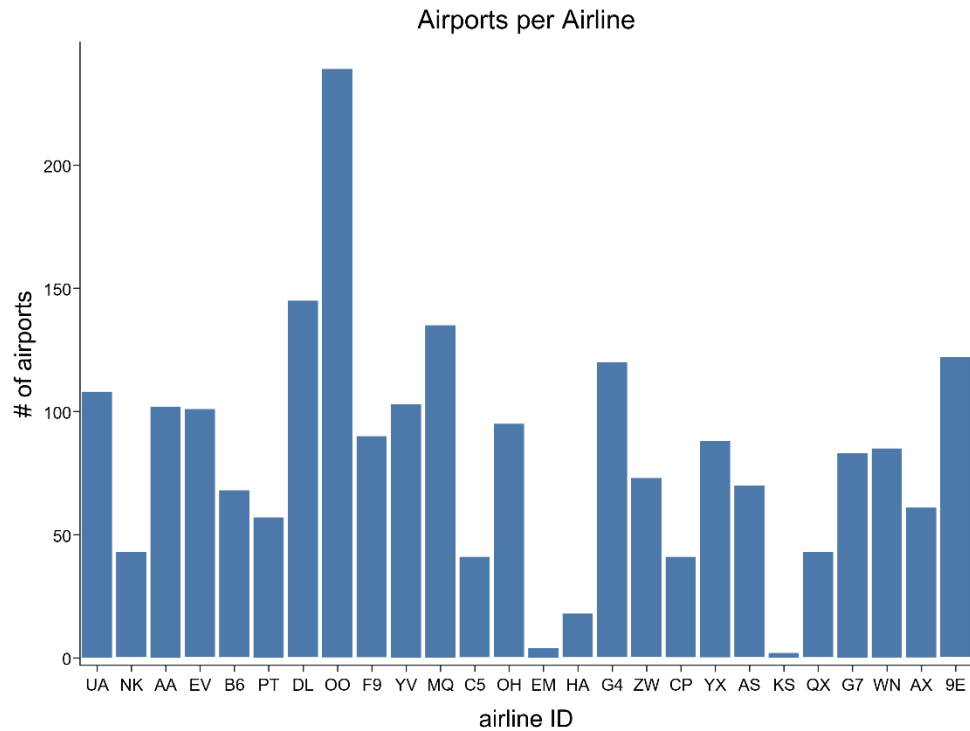


Рисунок 2.3.2.2 – Кількості аеропортів, на яких оперують перевізники

На рисунку 2.3.2.2 можна побачити, що різні перевізники оперують на доволі різних кількості аеропортів.

З цих спостережень можна зробити висновок, що затримка рейсів сильно варіюється і залежно від перевізників, і залежно від аеропортів. Робимо висновок, що щоб точно передбачити відтермінування польотів, потрібно мати окрему модель для кожної пари авіаперевізників та аеропортів. Проте з рисунку 2.3.2.2 видно, що для деяких пар може бути недостатня кількість рейсів, щоб точно навчити модель.

2.3.3 Вплив часу вильоту на затримку рейсу

Далі, було проаналізовано, чи існує взаємозв'язок між часом вильоту та затримкою.

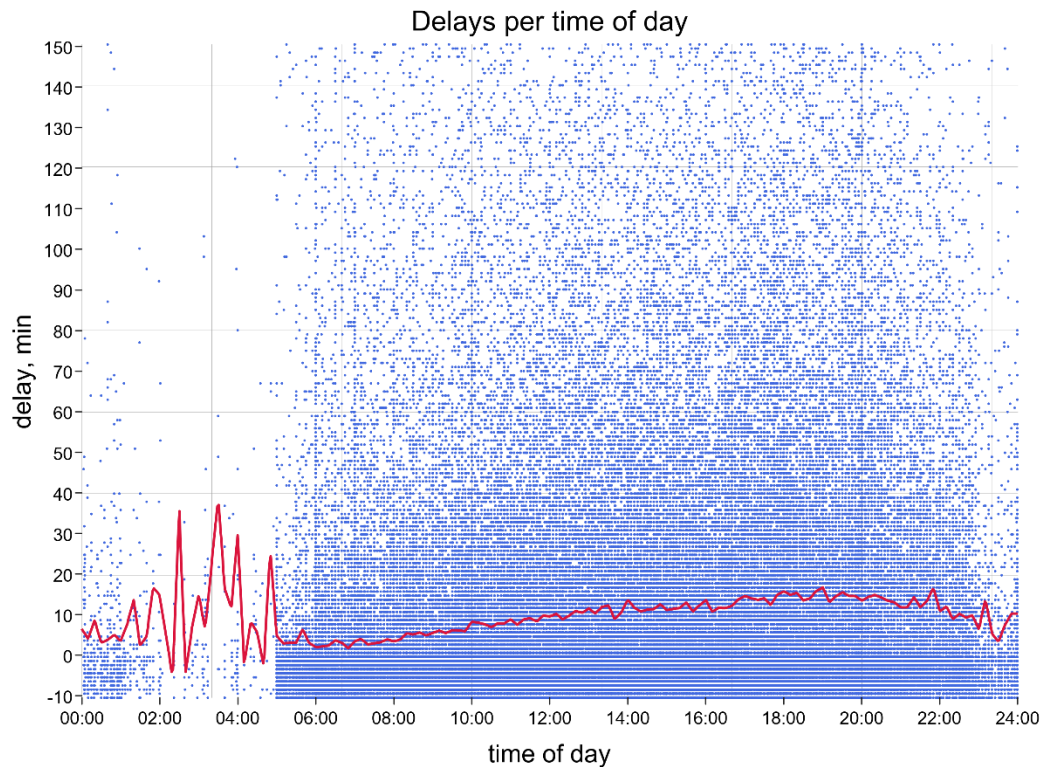


Рисунок 2.3.3.1 – Середня затримка від часу дня

На рисунку 2.3.3.1 синім зображено окремі рейси, а червоним – середня затримка. Якщо проігнорувати затримки рейсів в періоді від 1 до 5 години, де недостатньо даних для точного аналізу, можемо зробити висновок, що затримка рейсів збільшується зі збільшенням часу вильоту. Отже, ці дані також є корисними.

2.4 Обробка даних

Далі була проведена обробка даних, а саме: видалено дані, які містять значення null, видалено непотрібні дані, а також час дня перетворений на кількість хвилин з опівночі. На рисунку 2.4.1 можна побачити перші 8 елементів результуючого датасету.

```

+-----+-----+-----+-----+
|airline|origin|schedDepTime|depDelay|
+-----+-----+-----+-----+
|      AA| 14100|           640|    -9.0|
|      AA| 14100|           640|    -6.0|
|      AA| 14100|           640|    -1.0|
|      AA| 14100|           640|    -5.0|
|      AA| 14100|           640|    90.0|
|      AA| 14122|           421|    -4.0|
|      AA| 14122|           421|    -7.0|
|      AA| 14122|           421|   895.0|
+-----+-----+-----+-----+

```

Рисунок 2.4.1 – Вигляд датасету після обробки даних

3 НАВЧАННЯ МОДЕЛІ ШТУЧНОГО ІНТЕЛЕКТУ

3.1 Модель для одного перевізника та одного аеропорту

Першим кроком було зробити найпростіший варіант моделі, для одного перевізника та одного аеропорту. Було обрано пару, яка має найбільшу кількість рейсів у датасеті. Для навчання була обрана модель лінійної регресії. Датасет розділяється довільно на дві частини: навчальний та тестовий. Відповідно до загального правила, 70% даних належатиме навчальному датасету, а решта 30% – тестовому. Після тренування, на екран виводяться параметри, а також оцінки помилок.

Як було припущено раніше, дані з дуже великим значенням затримки може нашкодити моделі.

```
With delays > 60 mins
RMSE: 29.042032357346418
MSE: 843.4396434451563
r2: 0.002474642657214643

Without delays > 60 mins
RMSE: 10.230900091650096
MSE: 104.67131668532593
r2: 0.003036749398253602
```

Рисунок 3.1.1 – Порівняння моделі з видаленням великих затримок і без видалення

На рисунку 3.1.1 можемо бачити, що середня квадратична помилка (MSE) зменшується у 8 разів після видалення великих затримок. Отже, це припущення виявилось вірним.

В результаті вийшло натренувати модель з середньою помилкою у 10 хвилин, при чому помилка у більше, ніж 15 хвилин зустрічалася у 8.3% випадків.

3.2 Модель для одного перевізника та багатьох аеропортів

Наступним кроком було створити модель, що включає рейси певного перевізника на всіх аеропортах. Як зазначалося раніше, деякі аеропорти можуть мати доволі малу кількість рейсів. У таких ситуаціях ця модель зможе передбачити затримку з більшою точністю, бо враховуватимуться дані рейсів цього ж перевізника з інших аеропортів.

Для представлення різних аеропортів було використано метод «One-hot encoding», який полягає в тому, що для кожного аеропорту створюється новий стовбець зі значенням 0 або 1.

```
Error > 15 min: 11.437%  
RMSE: 11.94332474626656  
MSE: 142.64300599478318  
r2: 0.061807237144060445
```

Рисунок 3.2.1 – Результат навчання моделі №2

В результаті було навчено модель з середньою помилкою у 12 хвилин. У 11.4% випадків модель помилилася на більше, ніж 15 хвилин.

Надалі можливо навчити моделі для кожного перевізника, та передати параметри цих моделей у певний інтерфейс, з яким буде взаємодіяти користувач-мандрівник.

Висновки

Створити додаток, що прогнозує затримки авіарейсів цілком можливо, і розв’язання цієї проблеми може спростити процес планування подорожей для мільйонів людей.

З іншого боку, Scala та Spark – дуже потужне поєднання технологій для роботи з великою кількістю даних. Створений додаток продемонстрував, що це – прекрасне поєднання для розв’язання задач машинного навчання.

Додаток можна покращити, створивши загальнодоступний інтерфейс для користувача, наприклад, веб-сайт.

Список Використаної Літератури

1. Air Traffic By The Numbers. [Електронний ресурс] – Режим доступу :
https://www.faa.gov/air_traffic/by_the_numbers/
2. Why is My Flight Delayed? The 20 Main Reasons for Flight Delays.
[Електронний ресурс] – Режим доступу :
<https://www.claimcompass.eu/blog/why-is-my-flight-delayed/>

Додаток А

Перелік прийнятих скорочень

США	– Сполучені Штати Америки;
ID	– identifier;
CSV	– comma separated values;
MSE	– mean squared error.