Big Data Analytics for Enhancing Flight Delay Forecasting Based on Weather Observations:

A Mixed Methods Study

Reham Tousoun

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2023

Supervisor: James Garza

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission.*

|  |  |
| --- | --- |
| **Module Title:** | MSc Data Analytics |
| **Assessment Title:** | Capstone Project |
| **Supervisor Name:** | James Garza |
| **Student Full Name:** | Reham Tousoun |
| **Student Number:** | Sba22483 |
| **Assessment Due Date:** | 22/09/2023 |
| **Date of Submission:** | /09/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Acknowledgments and Dedication**

I begin with heartfelt gratitude to ALLAH, the most Merciful and Compassionate, for the countless blessings bestowed upon me.

I extend my deepest appreciation to my supervisor, James Garza, for his unwavering patience and continuous support throughout this journey.

My sincere thanks go to all the lecturers at CCT, especially David Mcquaid, Muhammad Iqbal, Marina Iantorno, Sam Weiss, and Rory Byrne, for their exceptional guidance and mentorship.

I am profoundly grateful to Springboard for providing me with the funded opportunity to enrich my knowledge and skills.

I am indebted to CCT College for the excellent facilities and the supportive environment provided during my academic pursuits.

Forever and always, my deepest debt of gratitude remains with my parents May they rest in peace and my siblings Hesham, Seham, Wesam, Ahmed, and Mohammed. Special thanks are extended to my friends for their support and encouragement pivotal in reaching this academic milestone.

Finally, I express my deepest love and gratitude to my husband, Amr, and my children, Mohammed, Abdurahman, and Annas, for their constant support, love, and patience, enabling me to pursue and achieve my career goals.

**Reham Adel Abd El-Kareem Tousoun**

*I dedicate this dissertation in loving memory of my dear parents,*

*Adel AbdelKareem Tousoun & Faten Mostafa Radwan.*

**Abstract**

Every minute around the world many flights are cancelled or delayed for many reasons, such as weather conditions, airport traffic, fuel, and other circumstances. According to the United States Department of Transportation, 75.48% of flights were delayed due to weather between June 2017 and May 2022.

**Table of Contents**

[1. Introduction 7](#_Toc146202504)

[1.1. Overview 7](#_Toc146202505)

[1.2. Problem Statement 7](#_Toc146202506)

[1.3. Research Questions and Hypothesis 8](#_Toc146202507)

[1.4. Research Objectives 8](#_Toc146202508)

[1.5. Thesis Organization 9](#_Toc146202509)

[2. Background Literature review and theory 11](#_Toc146202510)

[3. Research methodology 12](#_Toc146202511)

[4. Implementation, Result and Evaluation 13](#_Toc146202512)

[5. Conclusion and Future Work 14](#_Toc146202513)

[6. Appendix 15](#_Toc146202514)

[7. References 16](#_Toc146202515)

# Introduction

## Overview

Each year, around 20% of airline flights experience delays or cancellations, primarily attributed to factors such as adverse weather conditions, carrier equipment issues, and technical problems at airports. These delays incur substantial costs for both airlines and passengers. For example, in 2007, flight delays were estimated to have cost the US economy $32.9 billion, with more than half of the financial burden borne by passengers (Ball et al., 2010). Flight delays pose a significant challenge in the civil aviation industry, resulting in both direct and indirect costs, such as maintenance expenses, additional fees, and decreased passenger satisfaction (Wang et al., 2020) (Zheng et al., 2020). …….

…………………………………..

## Problem Statement

Flight delays due to adverse weather conditions can have significant economic and operational consequences for airlines, passengers, and the entire air transport system. Most of the research conducted on predicting flight delays has primarily focused on short-term predictions. Short-term predictions refer to providing flight status information to passengers on the same day of travel, typically when they are already at the airport. These predictions consider various factors such as weather conditions, airport congestion, and current flight delays. However, the limitation of short-term predictions is that they may not adequately meet passengers' expectations and satisfaction levels, as they are already at the airport when they receive the delay information. In this work we focus on the prediction of the flight delay due to only weather conditions. Imagine **What would happen if we could properly predict whether the flight would be delayed or not at the time of ticket purchase?** If this dream came true, it would allow passengers to save both time and money by scheduling connecting flights with plenty of time to spare. Accurate predictions for long-term flight delays, for example at the travel date using the power of machine learning, deep learning, and big data approaches can be considered a magic key to overcoming the short-term flight delay issues (Smith and Sherry, 2008).

## Research Questions and Hypothesis

**Research Questions:**

Can we properly build a robust binary classification model that utilizes machine learning, deep learning, and big data analytics approaches to predict if the flight will delay more than 15 minutes from the scheduled departure time related to weather conditions?

If so, what is the most appropriate algorithm used to predict flight departure delays related to weather observation?

**Research Hypothesis:**

***Null Hypothesis (H0):*** The use of big data analytics tools significantly improves the classification of flight departure delays exceeding 15 minutes related to weather conditions compared to machine learning and deep learning.

***Alternative Hypothesis (H1):*** The use of big data analytics tools does not improve the classification of flight departure delays related to weather conditions.

## Research Objectives

The objectives of this study are:

* **RO1. Build Accurate Robust Binary Classification Models:** To successfully build accurate robust binary classification models using machine learning, deep learning, and big data analytics to predict flight delays exceeding 15 minutes based on weather conditions.
* **RO2. Optimize Hyperparameters:** To assess and optimize Hyperparameters by tunning the hyperparameters of the chosen algorithms to improve the model's performance and prediction accuracy.
* **RO3. Leverage Big Data for Weather Dataset:** To Study the importance of leveraging big data to handle the weather dataset.
* **RO4.** **Select Appropriate ML Techniques and Big Data Analytics Tools:** To select the appropriate ML techniques and big data analytics tool/platform and for forecasting flight delays related to weather.
* **RO5. Clarify the challenges and limitation:** Clarify the challenges and limitations to extend different machine learning classifiers like K-Nearest Neighbour (KNN), Support Vector Machines (SVM), and Random Forest(RF) in this study.
* **RO6. Evaluate Each Classifier:** To evaluate the performance of each flight delays forecasting model by computing the quality of service factors (QoS) for the predictor models includes accuracy, precision, and recall.
* **RO7. Handle the imbalanced dataset:** To study the effect on the real-time imbalanced dataset on the model’s performance.
* **RO8.** **Evaluate Model Performance after Applying Random Undersampling:** To evaluate the Performance of each model after applying the Random Undersampling technique to the current dataset.
* **RO9.** **Justification for Using Random Undersampling**: Why the Random Undersampling approach is the only technique used to imbalance the dataset.

## Thesis Organization

The thesis is structured as follows: In Chapter 2, we delve into the literature review and associated research within the flight delay prediction domain. The focus is on methodologies like Decision trees, Gaussian Naive Bayes, Logistic Regression, Deep Neural Networks, and a range of big data platforms and tools. In chapter 3 provides a thorough explanation of the research methodologies utilized to fulfil the project objectives. While Chapter 4 is dedicated to discussing the application of the research methodology in the project, showcasing the outcomes of each model, and evaluating model performance using quality of service factors (QoS) such as accuracy, precision, and recall. Lastly, Chapter 5 encapsulates the primary conclusions of this thesis work and suggests future research directions.

# Background Literature review and theory

Several research teams have conducted extensive studies on the analysis and prediction of flight delays, considering the significant costs associated with these delays for airlines, passengers, and society as a whole. In this literature review, we elaborate on the most research conducted in the recent years.

Overall, there has been insufficient research conducted especially in utilizing the big data platforms and tools in the prediction of flight delay.

Weather

Adverse weather is a challenge for the aviation industry leading to poor visibility. It can increase aircraft accident rates and even damage aircraft parts. This adverse weather has crucially significant economic and operational consequences for airlines, passengers, and the entire air transport system

# Research methodology

# Implementation, Result and Evaluation

# Conclusion and Future Work

# Appendix

# References

Ball, M. et al. (2010) Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the United States. Available at: https://rosap.ntl.bts.gov/view/dot/6234 (Accessed: 24 May 2023).

Wang, Y. et al. (2020) ‘A Comparative Study on Flight Delay Networks of the USA and China’, Journal of Advanced Transportation, 2020, p. e1369591. Available at: https://doi.org/10.1155/2020/1369591.

Zheng, Z. et al. (2020) ‘How Late Does Your Flight Depart? A Quantile Regression Approach for a Chinese Case Study’, Sustainability, 12(24), p. 10553. Available at: https://doi.org/10.3390/su122410553.