Flight Delay Prediction: Harnessing the Power of AI for Proactive Air Travel Management

MSc Research Project
Artificial Intelligence

Dona Elizebeth Devasia
Student ID: x22153471

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Dona Elizebeth Devasia		
Student ID:	X22153471		
Programme:	MSc in Artificial Intelligence	**Year:**	2023-2024
Module:	MSc Research Project		
Supervisor:	Rejwanul Haque		
Submission Due Date:	14-12-2023		
Project Title:	**Flight Delay Prediction: Harnessing the Power of AI for Proactive Air Travel Management**		
Word Count:	6228	**Page Count :20**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Dona Elizebeth Devasia

Date: 31-01-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Flight Delay Prediction: Harnessing the Power of AI for Proactive Air Travel Management

Dona Elizebeth Devasia
X22153471

Abstract. The complexity of air travel management necessitates proactive strategies to address the persistent challenge of flight delays. This research focuses on leveraging Artificial Intelligence (AI) to predict flight delays accurately, with a specific emphasis on enhancing operational efficiency for a targeted airline company. The core challenge involves predicting whether a flight will be delayed, on time, or early. Building upon foundational methodologies, it is intended to extend and refine existing approaches to cater specifically to the operational context of the airline industry. This methodology encompasses a thorough exploration of classification algorithms, ranging from basic models to advanced techniques like XGBoost. Through rigorous hyperparameter tuning and strategic feature engineering, nuanced patterns within the data have been uncovered. By delving into the intricacies of machine learning, conventional approaches is transcended, aiming to enhance the precision of flight delay predictions. The results of the analysis demonstrate the effectiveness of the tailored approach, showcasing improved accuracy compared to baseline models. Utilizing AI, profound insights into the factors influencing flight delays is revealed, providing actionable intelligence for enhanced operational management. This research contributes not only to the academic discourse on flight delay prediction but, more critically, offers tangible advancements to the targeted airline company's air travel management strategies. Through a synthesis of theoretical foundations and practical applications, this study envisions a paradigm shift in the realm of proactive air travel management.

Keywords – Flight Delay Prediction, Artificial Intelligence, Air Travel Management, Machine Learning Algorithms, Proactive Solutions, Hyperparameter Tuning, Feature Engineering, Operational Efficiency, Airline Industry, Predictive Analytics

1 Introduction

1.1 Introduction And Background

The realm of air travel management stands at the intersection of precision and complexity, with the challenge of flight delays casting a pervasive shadow over operational efficiency. As the aviation industry continues to soar, the imperative for proactive strategies to anticipate and mitigate flight delays becomes increasingly apparent. This research endeavours to address this challenge through the lens of Artificial Intelligence (AI), aiming to revolutionize flight delay prediction and, in turn, enhance the proactive management of air travel. The aviation landscape has witnessed a surge in research dedicated to unravelling the complexities of flight delay prediction. Seminal works by Bin Yu et al. [1], Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi [2], and Guan Gui et al. [3] have pioneered approaches utilizing deep learning, machine learning, and big data analytics. These studies lay the foundation for understanding the intricacies of flight delays, providing valuable insights into predictive modeling within the aviation context. In this context, the aim of our research is twofold. Firstly, it seeks to build upon and extend existing methodologies to cater specifically to the operational needs of the target, Alaska Airlines. Secondly, the aim is to contribute to the broader discourse on proactive air travel management by refining the accuracy and efficiency of flight delay predictions. Upon delving into the nuances of classification algorithms, hyperparameter tuning, and feature engineering, the endeavour is to not only advance theoretical frameworks but to deliver practical solutions with real-world implications for the airline industry.

This research, inspired by the pioneering works of predecessors [1][2][3], envisions a paradigm shift in the way flight delay prediction is approached. By leveraging the power of AI, it aspired to provide Alaska Airlines with actionable intelligence that transcends traditional models, offering a pathway to more effective and efficient air travel management. The intricacies of air travel management encapsulate a myriad of challenges, among which the persistent spectre of flight delays remains a critical concern. As the aviation industry continues to expand, the need for innovative strategies to anticipate and mitigate these delays becomes increasingly imperative. This research embarks on a journey to confront this challenge head-on, wielding the prowess of Artificial Intelligence (AI) as a transformative tool to redefine the landscape of flight delay prediction. In doing so, the aim is not only to fortify the operational efficiency of air travel but also to offer tailored solutions for the specific benefit of Alaska Airlines. The evolving discourse on flight delay prediction has been significantly shaped by seminal contributions from leading researchers in the field. Bin Yu et al.'s work [1] introduces a deep learning approach, underscoring the potential of sophisticated neural networks in unravelling the complex patterns inherent in-flight data. The machine learning approach proposed by Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi [2] adds a nuanced perspective, emphasising the importance of predictive analytics in understanding and managing flight departure delays. Furthermore, the study by Guan Gui et al. [3] showcases the power of leveraging aviation big data and machine learning for accurate flight delay predictions.

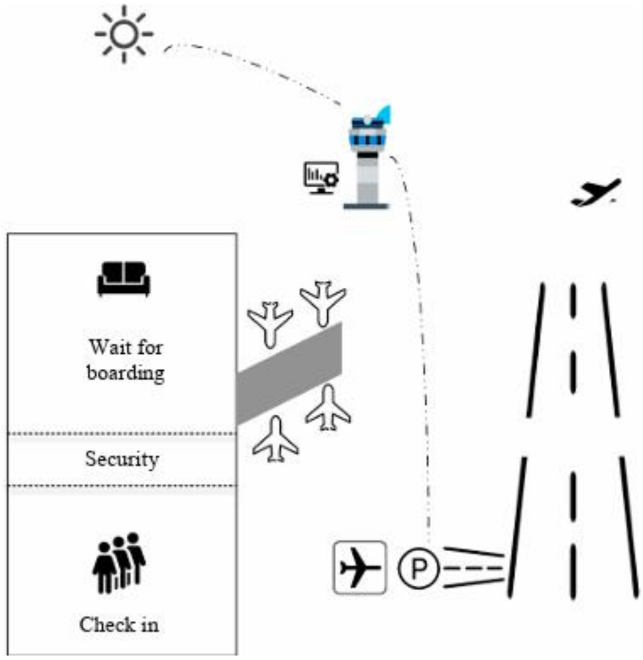


Fig. 1: *Schematic of flight operations in a commercial airport.*

Against this backdrop, this research aspires to extend and innovate upon these established methodologies. By focusing on the specific needs of Alaska Airlines, it seeks to tailor the predictive models to the nuances of their operational environment. The aim is not merely academic; rather, it is rooted in the practical implications of providing a valuable decision-making tool for airline professionals. This approach encompasses a comprehensive exploration of classification algorithms, ranging from fundamental models to advanced techniques such as XGBoost, all subjected to rigorous hyperparameter tuning. The commitment to innovation is evident in the strategic application of feature engineering, aiming to extract meaningful insights from the wealth of available data. The synergy of these methodologies is aimed not only at improving the accuracy of flight delay predictions but also at unravelling the underlying factors contributing to delays. Upon navigating this terrain, the primary aim remains twofold: to offer Alaska Airlines a toolset that enhances their proactive air travel management and to contribute substantially to the broader discourse on the predictive modelling of flight delays.

1.2 Research Question:

"How can the predictive accuracy of flight delay models be enhanced through the systematic development, optimization, and comparison of classification algorithms, hyperparameter tuning, and strategic feature engineering? Additionally, how do these enhancements contribute to the proactive management of air travel operations, specifically addressing the operational needs of Alaska Airlines?"

1.3 Research Objectives:

- *Develop and Optimize Predictive Models:* The primary objective of this research is to develop and optimize predictive models for flight delay prediction, leveraging a diverse range of classification algorithms. Through systematic exploration, from foundational models to advanced techniques like XGBoost, the aim is to identify the most effective algorithm for accurately forecasting flight delays.

- *Refine Predictive Accuracy through Hyperparameter Tuning*: Another key objective is to refine the predictive accuracy of the developed models through rigorous hyperparameter tuning. This involves a meticulous process of fine-tuning model parameters to optimize their performance for the specific task of flight delay prediction. The objective is to enhance the models' ability to discern patterns within historical flight data.

- *Uncover Insights through Strategic Feature Engineering:* The research aims to uncover meaningful insights into the factors influencing flight delays through strategic feature engineering. This involves not only selecting relevant features but also crafting new variables that capture the complex relationships within the data. The objective is to go beyond traditional predictors and extract nuanced information that contributes to the overall accuracy and interpretability of the predictive models.

2 Related Work

The landscape of flight delay prediction has been shaped by a wealth of research, each contribution adding a layer of understanding to the intricate dynamics of air travel management. This review engages with pivotal studies in the field, underscoring the diversity of approaches and methodologies that have been employed to tackle the challenge of predicting flight delays accurately.

Bin Yu et al. [1] introduced a deep learning approach to flight delay prediction, leveraging the capabilities of neural networks to capture complex patterns in historical data. Their work emphasizes the potential of deep learning models in handling the intricacies of flight data, offering a paradigm shift in predictive analytics within the aviation domain. Similarly, Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi [2] adopted a machine learning perspective, specifically focusing on flight departure delays. Their study emphasizes the role of predictive analytics in understanding and managing delays at the critical phase of departure. Guan Gui et al. [3] explored the intersection of aviation big data and machine learning, showcasing the power of data-driven approaches in flight delay prediction. Their work highlights the importance of harnessing vast datasets to uncover patterns and trends that might elude traditional methods. Additionally, Jingyi Qu et al. [4] introduced a novel perspective by utilizing a deep convolutional neural network and fusing meteorological data. Their approach showcases the potential of integrating diverse sources of information to enhance the accuracy of predictions. Kolawole Ogunsina, Ilias Bilionis, and Daniel DeLaurentis [5] provided insights into exploratory data analysis for airline disruption management, contributing to the understanding of the broader context within which flight delays occur. L. Carvalho et al. [6], through a systematic review, underscored the relevance of data science in flight delay research, emphasizing the need for interdisciplinary approaches to tackle this multifaceted challenge.

Maryam Farshchian Yazdi et al. [7] introduced a unique blend of deep learning and the Levenberg-Marquardt algorithm for flight delay prediction, showcasing the diversity of computational techniques applied to the problem. Additionally, Wei Shao et al. [10] proposed an innovative approach using an airport situational awareness map for predicting flight delays, highlighting the importance of spatial and situational context. The works of Bin Yu et al. [1], Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi [2], Guan Gui et al. [3], and others collectively underscore the evolving landscape of flight delay prediction. As we navigate this rich tapestry of research, our aim is to build upon these foundations, extending and innovating in a manner that aligns specifically with the operational needs of Alaska Airlines. The subsequent sections detail our approach, presenting a synthesis of theoretical frameworks and practical applications, with the ultimate goal of contributing substantively to the proactive management of air travel.

Another notable contribution to the field comes from Suvojit Manna et al. [9], who employed a statistical approach using gradient-boosted decision trees for predicting flight delays. Their work emphasizes the significance of statistical modeling techniques in capturing nuanced relationships within the data. Weinan Wu et al. [10] explored the applicability of spatial awareness maps for flight delay prediction, introducing a unique perspective on incorporating geographical information into predictive models. Their work emphasizes the importance of contextual factors in understanding and predicting delays. Additionally, Yi Ding [12] delved into the realm of multiple linear regression for predicting flight delays, highlighting the diversity of statistical approaches applied to this complex problem. Young Jin Kim et al. [13] brought a deep learning approach to flight delay prediction, underscoring the potential of neural networks in capturing intricate patterns within the data. Their work, conducted at the Aerospace Systems Design Laboratory at the Georgia Institute of Technology, adds to the growing body of research showcasing the adaptability of deep learning techniques to aviation challenges.

Ziad J. Chaudhry and Kevin L. Fox [15] explored the applicability of artificial intelligence to air traffic management network operations, extending the scope of research beyond individual flight predictions to the broader context of airspace management. Their work contributes to a holistic understanding of how AI can be integrated into the broader air transportation infrastructure. In a study by Wei Shao et al. [10], the authors proposed a novel approach using an airport situational awareness map for predicting flight delays. This innovative perspective emphasizes the importance of spatial and situational context in understanding and predicting delays.

The aforementioned studies collectively provide a nuanced view of the methodologies employed in predicting flight delays, ranging from deep learning and machine learning approaches to statistical modeling and spatial analysis. As we synthesize this diverse body of work, our aim is to contribute to this evolving field by tailoring and extending these methodologies to address the unique challenges faced by Alaska Airlines. The subsequent sections of this thesis delve into the specific methodologies employed, showcasing our approach to refining the accuracy and efficiency of flight delay predictions. Through this research, we aspire to not only add to the academic discourse but, more importantly, offer practical insights that can drive proactive air travel management strategies for Alaska Airlines.

3 Research Methodology

The research methodology consists of four steps namely Exploration of Classification Algorithms, Rigorous Hyperparameter Tuning, Strategic Feature Engineering and Comprehensive Model Comparison

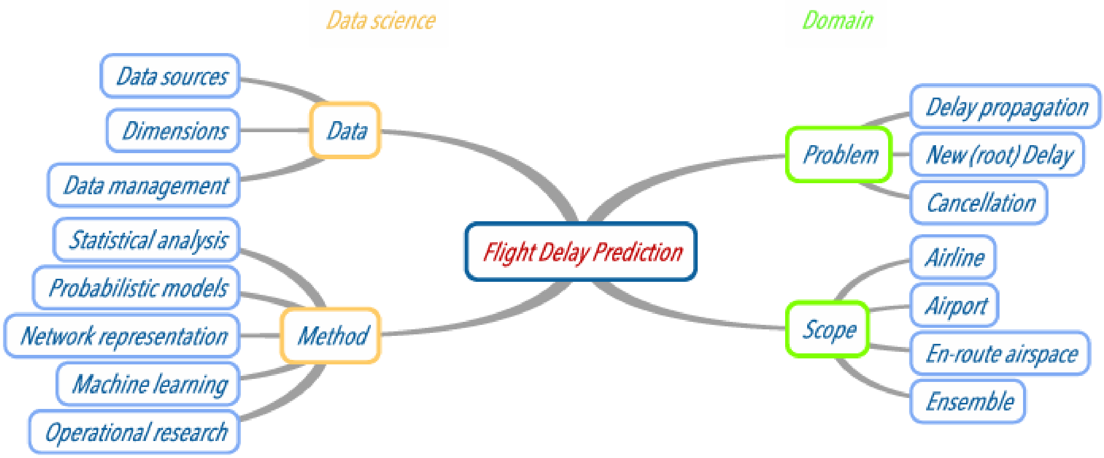


Fig.2: *Taxonomy of the flight delay prediction problem*

The methodology unfolds through a structured sequence of steps, each meticulously designed to refine the accuracy and efficiency of flight delay predictions. Anchored in a comprehensive exploration of classification algorithms, hyperparameter tuning, and strategic feature engineering, this approach aims to transcend conventional models and offer tailored solutions for the specific operational context of Alaska Airlines.

Step 1: Data Gathering

The first step in the methodology involves gathering the data from reliable sources. Data is often is present in CSV file format, it can be imported via use of Pandas library.

Step 2: Data Preprocessing

The second step in the methodology involves preprocessing the data through Pandas library of Python. It involves steps like handling missing values, encoding categorical variables, feature scaling, etc. Each step is very crucial in making data ready for feature scaling/extraction. If data is preprocessed in wrong way/manner it may lead to loss of data.

Step 3: Data Modelling

Data modelling is done to understand the relationships between the features and understand the patterns that are hidden in the dataset. Heat map is plotted, confusion matrix is calculated which helps to find valuable insights in the data.

Step 4: Exploration of Classification Algorithms

The fourth step involves a thorough exploration of classification algorithms. Starting with foundational models, we progressively advance to more sophisticated techniques such as XGBoost. Each algorithm is carefully chosen to strike a balance between computational efficiency and predictive accuracy, considering the intricate patterns embedded in historical flight data.

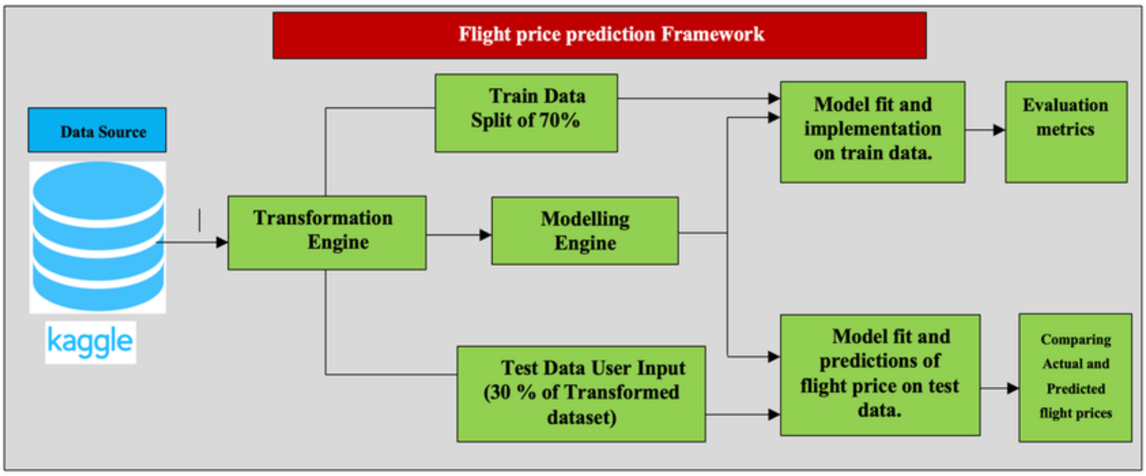
Step 5: Rigorous Hyperparameter Tuning
The second step encompasses a rigorous process of hyperparameter tuning for each selected algorithm. Recognizing the nuanced interplay between model parameters and predictive performance, we systematically fine-tune the hyperparameters to optimize the algorithms for our specific prediction task. This iterative process aims at enhancing the models' ability to discern patterns within the data and improve overall predictive accuracy.

Step 6: Strategic Feature Engineering
Central to our methodology is the strategic application of feature engineering. Understanding that the predictive power of algorithms relies on the quality of input features, we delve into the wealth of available data to extract meaningful insights. This step involves not only selecting relevant features but also crafting new variables that encapsulate the complex relationships between different factors influencing flight delays.

Step 7: Comprehensive Model Comparison
The culmination of our methodology involves a comprehensive comparison of the performance of the various models developed through the preceding steps. This entails a detailed evaluation of metrics such as accuracy, precision, and recall. The goal is to discern the strengths and weaknesses of each model, providing a basis for selecting the most effective approach for predicting flight delays within the operational context of Alaska Airlines.

Through this systematic methodology, it is aimed to contribute not only to the academic discourse on flight delay prediction but also more importantly, to deliver actionable insights for enhancing the proactive management of air travel operations for Alaska Airlines.

4 Design Specification



4.1 Data Preprocessing:

Handling Missing Values:
Identify missing values through thorough examination of the dataset. Utilize appropriate imputation techniques such as mean, median, or advanced methods like K-nearest neighbours' imputation to address missing data points.

Encoding Categorical Variables:
Analyze the nature of categorical variables and apply suitable encoding methods. One-hot encoding can be employed for nominal variables, while label encoding may be suitable for ordinal ones.

Feature Scaling:
Normalize numerical features using techniques like Min-Max scaling or Standard Scaling to ensure that all variables contribute equally to model training.

4.2Visualization:

Exploratory Data Analysis (EDA):
Generate descriptive statistics, distribution plots, and correlation matrices to gain insights into data characteristics. Identify potential outliers and anomalies that may impact model performance.

Feature Importance Visualization:
Utilize techniques such as bar plots, heatmaps, or tree-based model feature importance plots to visually represent the significance of each feature in influencing flight delays.

4.3 Model Fitting:

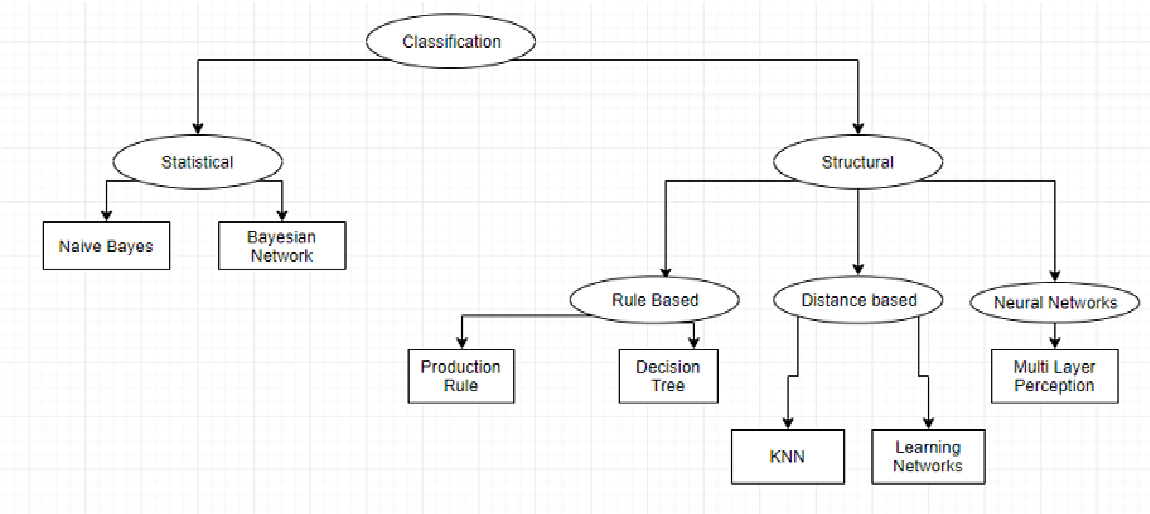


Fig. 3: *Overview of Classification approach*

Logistic Regression:
In the realm of flight delay prediction, Logistic Regression serves as a robust linear classifier. This model is adept at capturing straightforward relationships between various features and the likelihood of flight delays. By implementing the logistic function, it transforms the linear combination of input features into a probability score, facilitating binary classification. Logistic Regression provides interpretability by assigning weights to each feature, allowing for a clear understanding of how individual variables contribute to the prediction. It is essential to ensure that the assumptions of logistic regression, such as linearity and independence of errors, align with the characteristics of the flight delay dataset.

Support Vector Classifier (SVC):
The Support Vector Classifier (SVC) plays a pivotal role in handling non-linear relationships within the flight delay dataset. Leveraging kernel functions, the SVC enhances its capability to capture intricate and complex patterns in the data. Unlike linear classifiers, SVC can effectively map input features into higher-dimensional spaces, enabling the model to discern non-linear decision boundaries. This makes SVC well-suited for scenarios where the relationships between features and flight delays are not strictly linear. However, practitioners must be mindful of the computational demands associated with kernelized SVC, particularly in larger datasets.

Decision Tree Classifier:
The Decision Tree Classifier is a powerful tool for modeling hierarchical relationships and potential interactions among features in the context of flight delay prediction. This model constructs a tree-like structure, where each internal node represents a decision based on a particular feature, and each leaf node corresponds to a predicted outcome. Decision trees offer interpretability by providing a transparent representation of the decision-making process. However, they can be prone to overfitting, emphasizing the importance of fine-tuning parameters to achieve a balance between model complexity and generalization performance.

K-Neighbors Classifier:
The K-Neighbors Classifier adopts a non-parametric approach to predicting flight delays by utilizing the proximity of data points. This model categorizes instances based on the characteristics of their neighboring data points in the feature space. The "K" in K-Neighbors refers to the number of nearest neighbors considered for classification. While K-Neighbors Classifier is intuitive and easy to implement, practitioners must carefully choose an appropriate value for K to balance model flexibility and avoid underfitting or overfitting. This model is particularly suitable for scenarios where the relationships between features and flight delays exhibit local patterns.

Naive Bayes (GaussianNB):
The Gaussian Naive Bayes algorithm is a probabilistic model that assumes independence between features. In the context of flight delay prediction, this algorithm is well-suited for datasets with continuous variables. Naive Bayes calculates the probability of an instance belonging to a particular class based on the distribution of its features. Despite its simplicity, Naive Bayes can perform surprisingly well, especially when the independence assumption holds true. Practitioners should be mindful of the impact of feature correlation on the model's performance and consider data preprocessing techniques to address such dependencies.

LightGBM:
LightGBM stands out as a gradient-boosting framework based on decision trees, offering notable advantages such as efficient training speed and lower memory usage compared to traditional boosting algorithms. In the context of flight delay prediction, integrating LightGBM involves leveraging its ability to handle large datasets and capture complex relationships between features. Fine-tuning hyperparameters, including the learning rate and tree depth, becomes crucial to harness the full potential of LightGBM. This model presents a promising avenue for enhancing prediction efficiency and is particularly beneficial when dealing with substantial amounts of flight data.

4.4 Performance Evaluation:

Cross-Validation:

Apply k-fold cross-validation to assess the models' generalization performance and ensure robustness against variations in the dataset.

Model Comparison:
Evaluate each model's performance using metrics like accuracy, precision, recall, and F1-score. Consider the trade-offs between false positives and false negatives, particularly critical in the context of flight delay prediction.
+
Efficiency Metrics for LightGBM:
Specifically measure the efficiency gains achieved by LightGBM in terms of reduced memory usage and faster training times. Compare these efficiency metrics against other fitted models.

4.5 Hyperparameter Tuning:

Grid Search or Random Search:
Conduct hyperparameter tuning through systematic grid search or random search. This involves exploring a range of hyperparameter values to find the combination that optimizes the model's performance.

4.6 Documentation and Reporting:

Comprehensive Reporting:
Document each step of the preprocessing, visualization, and model fitting processes in detail, including code snippets, parameters used, and any notable observations.

Visualization Summary:
Summarize key visualizations in the form of clear, interpretable plots and graphs. Include captions and annotations to highlight specific insights.

Model Performance Report:
Generate a comprehensive report outlining the performance of each fitted model. Include a discussion of strengths, weaknesses, and practical implications for the proactive management of air travel operations for Alaska Airlines.

5 Implementation

Note- Feature Engineering:
For arriving flights: The Actual taXi-In Time (AXIT) is the period between the Actual Landing Time (ALDT) and the Actual In-Block Time (AIBT)
For departing flights: the Actual taXi-Out Time (AXOT) is the period between the Actual Off-Block Time (AOBT) and the Actual Take Off Time (ATOT)

- Calculate the taxi-in time (AXIT) for arriving flights: AXIT = AIBT - ALDT
- Calculate the taxi-out time (AXOT) for departing flights: AXOT = ATOT – AOBT

5.1 Data Preprocessing:

Missing Values

```
FL_DATE                  0
OP_UNIQUE_CARRIER        0
OP_CARRIER               0
TAIL_NUM               324
OP_CARRIER_FL_NUM        0
ORIGIN_AIRPORT_ID        0
ORIGIN                   0
DEST_AIRPORT_ID          0
DEST                     0
CRS_DEP_TIME             0
DEP_TIME              1718
TAXI_OUT              1779
WHEELS_OFF            1779
WHEELS_ON             1845
TAXI_IN               1845
CRS_ARR_TIME             0
ARR_TIME              1845
ARR_DELAY_GROUP       2075
CANCELLED                0
DISTANCE                 0
dtype: int64
```

In the dataset, missing values are observed in columns such as 'TAIL_NUM,' 'DEP_TIME,' 'TAXI_OUT,' 'WHEELS_OFF,' 'WHEELS_ON,' 'TAXI_IN,' 'ARR_TIME,' and 'ARR_DELAY_GROUP.' These gaps may stem from data recording errors, instances with unavailable information, or flights without recorded delays, necessitating attention during analysis and potential imputation.

Information of the variables

	Data Type	No of Unique Data	Levels	Null_values	null%
FL_DATE	object	122	['2019-08-11' '2019-08-31' '2019-08-09' '2019-...	0	0.00
OP_UNIQUE_CARRIER	object	26	['WN' 'YX' 'AA' 'OO' 'DL' 'B6' 'YV' 'OH' 'UA' ...	0	0.00
OP_CARRIER	object	26	['WN' 'YX' 'AA' 'OO' 'DL' 'B6' 'YV' 'OH' 'UA' ...	0	0.00
TAIL_NUM	object	5867	['N206WN' 'N745YX' 'N751UW' ... 'N799AN' '280N...	324	3.24
OP_CARRIER_FL_NUM	int64	6509	[4669 3502 1959 ... 6417 6154 6510]	0	0.00
ORIGIN_AIRPORT_ID	int64	370	[13871 12266 11423 13930 14107 13476 10423 128...	0	0.00
ORIGIN	object	370	['OMA' 'IAH' 'DSM' 'ORD' 'PHX' 'MRY' 'AUS' 'LA...	0	0.00
DEST_AIRPORT_ID	int64	368	[11259 11413 14107 11298 13342 12892 11618 124...	0	0.00
DEST	object	368	['DAL' 'DRO' 'PHX' 'DFW' 'MKE' 'LAX' 'EWR' 'JF...	0	0.00
CRS_DEP_TIME	int64	1231	[1020 1000 1437 ... 123 450 219]	0	0.00
DEP_TIME	float64	1354	[1027. 953. 1436. ... 434. 306. 232.]	1718	17.18
TAXI_OUT	float64	154	[8. 27. 11. 17. 14. 10. 15. 16. 12. ...	1779	17.79
WHEELS_OFF	float64	1356	[1035. 1020. 1447. ... 446. 431. 243.]	1779	17.79
WHEELS_ON	float64	1425	[1155. 1119. 1525. ... 346. 357. 332.]	1845	18.45
TAXI_IN	float64	103	[5. 7. 4. 15. 8. 9. 10. 6. 17. ...	1845	18.45
CRS_ARR_TIME	int64	1334	[1210 1127 1536 ... 158 449 437]	0	0.00
ARR_TIME	float64	1424	[1200. 1126. 1529. ... 350. 256. 322.]	1845	18.45
ARR_DELAY_GROUP	object	3	['early_arrival' 'ontime' 'delayed' nan]	2075	20.75
CANCELLED	float64	2	[0. 1.]	0	0.00
DISTANCE	float64	1489	[586. 869. 1149. ... 747. 1472. 2874.]	0	0.00

The dataset exhibits diverse variable types and characteristics, ranging from object types (e.g., 'FL_DATE') with 122 unique values to numerical types (e.g., 'DEP_TIME') with 17.18% missing values. Categorical features like 'ARR_DELAY_GROUP' have three levels, including 'nan' representing missing values, accounting for 20.75% of the data.

Shape of data

```
data.shape
```

```
(134235, 20)
```

Shape of data after dropping missing values.

Encoding Categorical Variables:

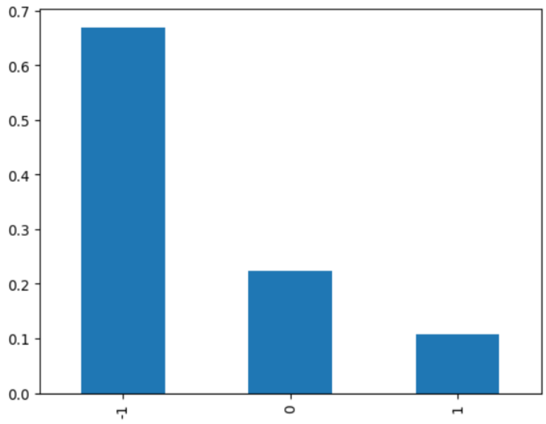


Fig. 4: *Encoding Categorical Variables*

Observation from ‘ARR_DELAY_GROUP’ replaced early_arrival to -1, ontime to 0, delayed to 1.

Correlation with heat map

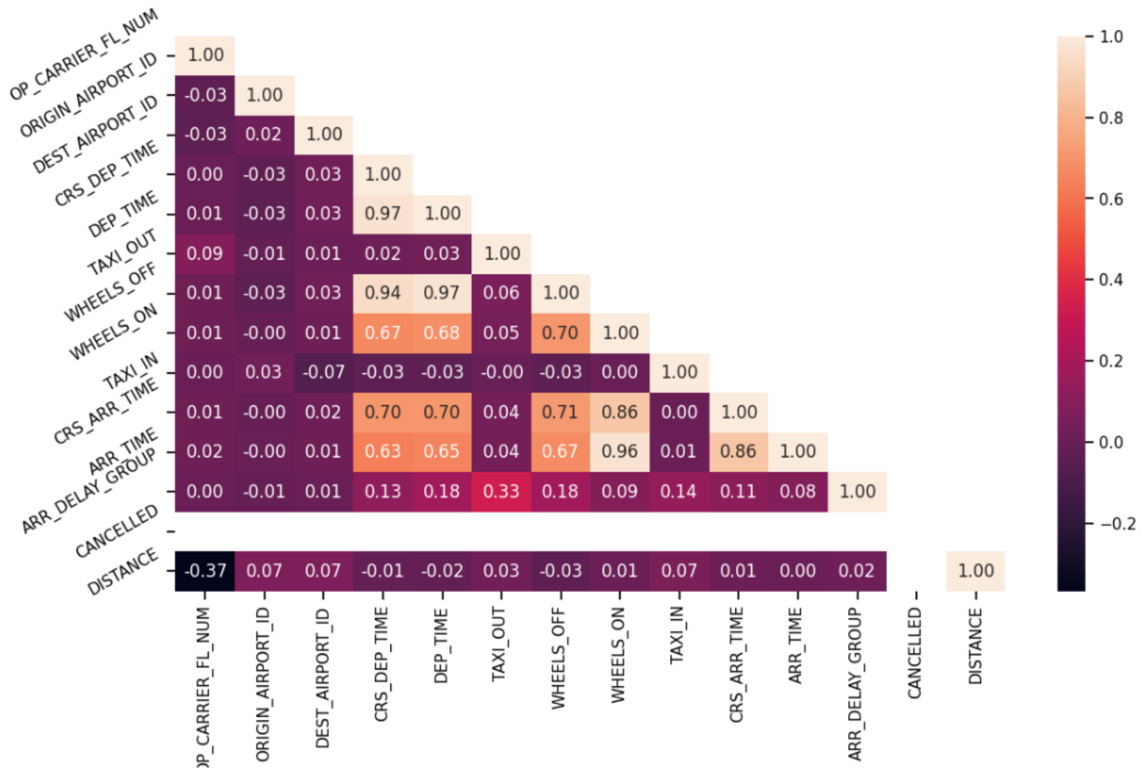


Fig. 5: *Correlation heat map*

Strong Positive Correlations: 'DEP_TIME' and 'CRS_DEP_TIME' have a strong positive correlation of 0.967314, indicating a close relationship between the actual departure time and the scheduled departure time. 'WHEELS_OFF' and 'CRS_DEP_TIME' exhibit a strong positive correlation of 0.942174, indicating that the wheels-off time and scheduled departure time are closely related. 'ARR_TIME' and 'WHEELS_ON' show a

strong positive correlation of 0.964835, suggesting a close relationship between the actual arrival time and the wheels-on time. 'ARR_TIME' and 'CRS_ARR_TIME' have a strong positive correlation of 0.861665, indicating a significant relationship between the actual arrival time and the scheduled arrival time.
Strong Negative Correlations: There are no strong negative correlations evident in the heat map.

Data Information:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 134235 entries, 0 to 136309
Data columns (total 17 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   FL_DATE            134235 non-null  object
 1   OP_UNIQUE_CARRIER  134235 non-null  object
 2   OP_CARRIER         134235 non-null  object
 3   TAIL_NUM           134235 non-null  object
 4   OP_CARRIER_FL_NUM  134235 non-null  int64
 5   ORIGIN_AIRPORT_ID  134235 non-null  int64
 6   ORIGIN             134235 non-null  object
 7   DEST_AIRPORT_ID    134235 non-null  int64
 8   DEST               134235 non-null  object
 9   CRS_DEP_TIME       134235 non-null  int64
 10  TAXI_OUT           134235 non-null  float64
 11  WHEELS_OFF         134235 non-null  float64
 12  TAXI_IN            134235 non-null  float64
 13  CRS_ARR_TIME       134235 non-null  int64
 14  ARR_DELAY_GROUP    134235 non-null  int64
 15  CANCELLED          134235 non-null  float64
 16  DISTANCE           134235 non-null  float64
dtypes: float64(5), int64(6), object(6)
memory usage: 18.4+ MB
```

```
data.shape
```

```
(134235, 17)
```

Data information and shape of data after removing unwanted variables.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 134235 entries, 0 to 136309
Data columns (total 22 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   OP_UNIQUE_CARRIER  134235 non-null  category
 1   TAIL_NUM           134235 non-null  category
 2   OP_CARRIER_FL_NUM  134235 non-null  float64
 3   ORIGIN_AIRPORT_ID  134235 non-null  float64
 4   ORIGIN             134235 non-null  category
 5   DEST_AIRPORT_ID    134235 non-null  float64
 6   DEST               134235 non-null  category
 7   CRS_DEP_TIME       134235 non-null  float64
 8   TAXI_OUT           134235 non-null  float64
 9   WHEELS_OFF         134235 non-null  float64
 10  TAXI_IN            134235 non-null  float64
 11  CRS_ARR_TIME       134235 non-null  float64
 12  ARR_DELAY_GROUP    134235 non-null  float64
 13  CANCELLED          134235 non-null  float64
 14  DISTANCE           134235 non-null  float64
 15  Month              134235 non-null  float64
 16  Day                134235 non-null  float64
 17  Dayofweek          134235 non-null  float64
 18  arr_hours          134235 non-null  float64
 19  arr_minutes        134235 non-null  float64
 20  dep_hours          134235 non-null  float64
 21  dep_minutes        134235 non-null  float64
dtypes: category(4), float64(18)
memory usage: 20.5 MB
(134235, 22)
```

Data information and shape of data after data preprocessing

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 134235 entries, 0 to 136309
Data columns (total 21 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   OP_UNIQUE_CARRIER   134235 non-null  category
 1   TAIL_NUM            134235 non-null  category
 2   OP_CARRIER_FL_NUM   134235 non-null  float64
 3   ORIGIN_AIRPORT_ID   134235 non-null  float64
 4   ORIGIN              134235 non-null  category
 5   DEST_AIRPORT_ID     134235 non-null  float64
 6   DEST                134235 non-null  category
 7   CRS_DEP_TIME        134235 non-null  float64
 8   TAXI_OUT            134235 non-null  float64
 9   WHEELS_OFF          134235 non-null  float64
 10  TAXI_IN             134235 non-null  float64
 11  CRS_ARR_TIME        134235 non-null  float64
 12  ARR_DELAY_GROUP     134235 non-null  float64
 13  DISTANCE            134235 non-null  float64
 14  Month               134235 non-null  float64
 15  Day                 134235 non-null  float64
 16  Dayofweek           134235 non-null  float64
 17  arr_hours           134235 non-null  float64
 18  arr_minutes         134235 non-null  float64
 19  dep_hours           134235 non-null  float64
 20  dep_minutes         134235 non-null  float64
dtypes: category(4), float64(17)
memory usage: 19.5 MB
```

Removed ‘CANCELLED’ column from table.

Box Plot
Variable- ARR_DELAY_GROUP

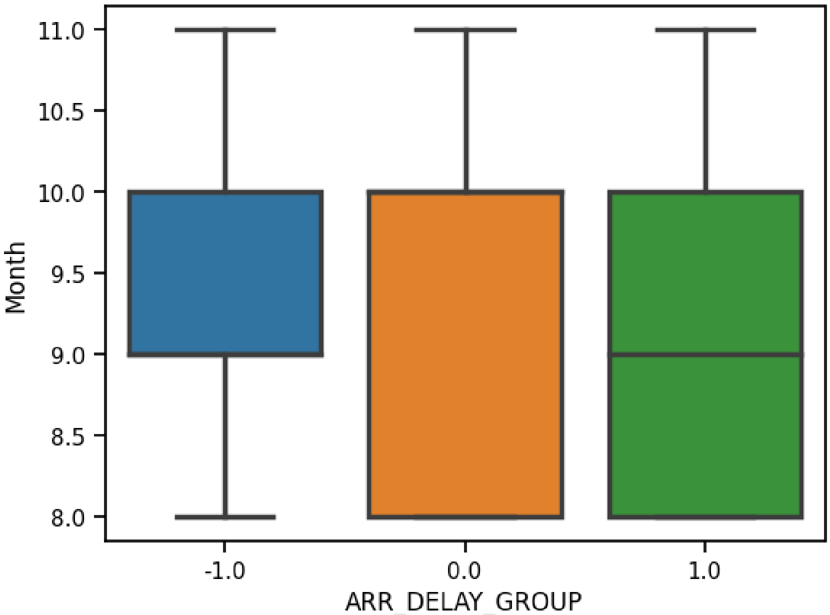


Fig. 6: *Box plot of ARR_Delay_Group with Monthly*

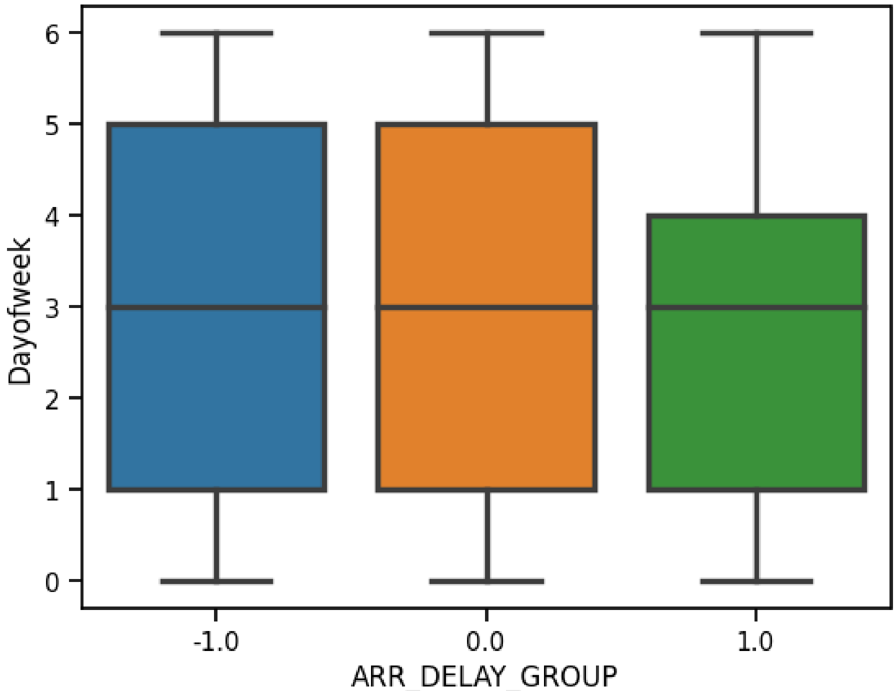
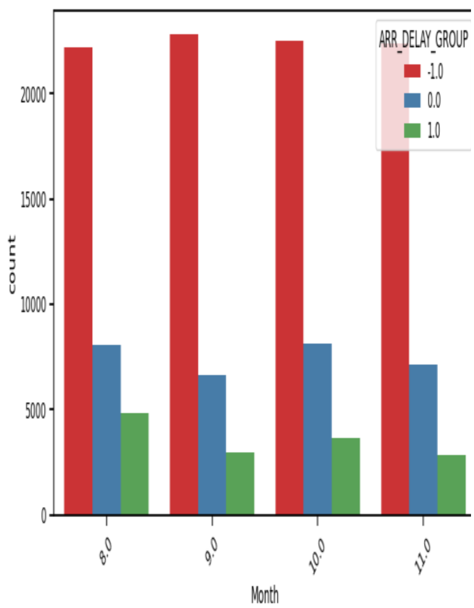


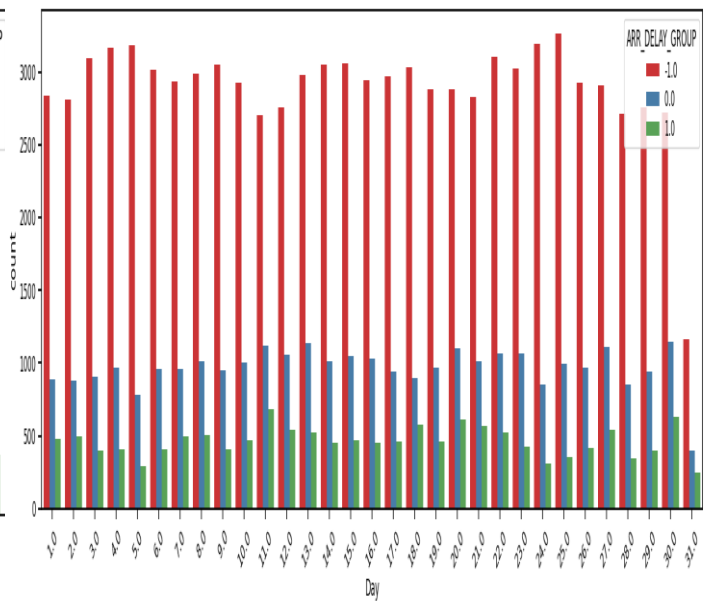
Fig. 7: *Box plot of ARR_Delay_Group with Day of week*

The box plot illustrating 'ARR_Delay_Group' on the x-axis and 'Monthly' and 'Day of week' on the y-axis provides a concise overview of the distribution of arrival delay groups across different months and Day of week. Each box represents the interquartile range for a specific month and Day of week, with the median line indicating the central tendency of delay groups. The whiskers and potential outliers offer insights into the variability and extreme values in arrival delays for each month and Day of week, facilitating a quick comparison of delay patterns over the monthly and Day of week timeline.

5.2 Data Visualization:

ARR_DELAY_GROUP





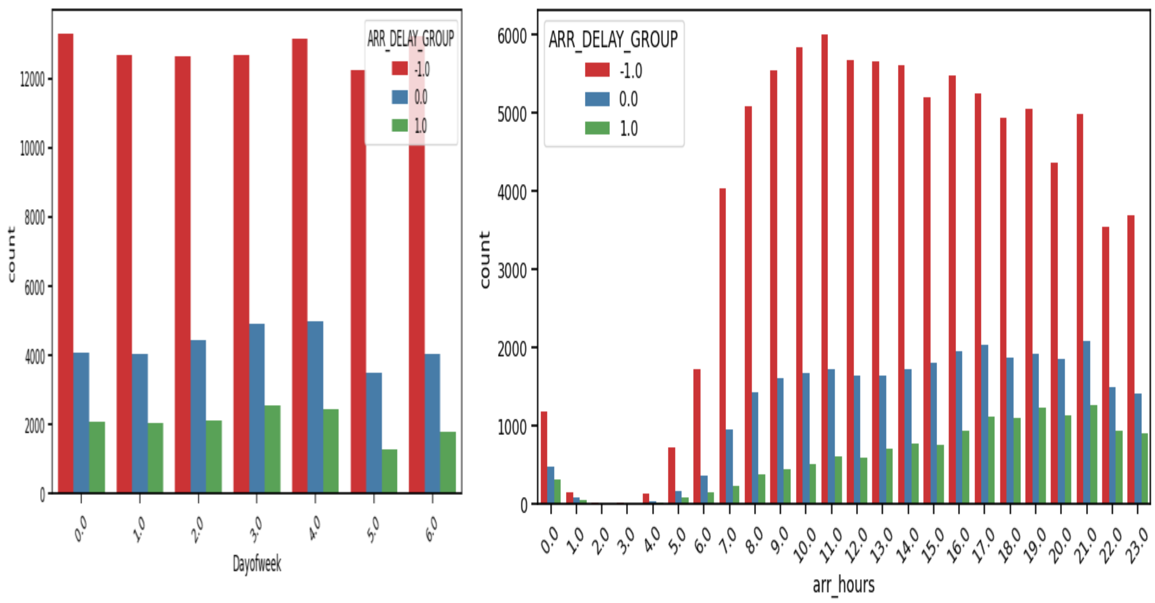


Fig.8: *Visualization plots of ARR_Delay_Group*

The four graphs depicting the distribution of 'ARR_DELAY_GROUP' across different time-related variables offer concise insights into arrival delay patterns. The daily graph reveals fluctuations in delays over individual days, while the weekly chart highlights variations throughout the week, pinpointing potential patterns linked to weekdays. The monthly distribution graph illustrates the prevalence of delays across different months, allowing for the identification of seasonal trends. Lastly, the hourly graph provides a detailed breakdown of delays throughout the day, aiding in the recognition of peak hours or time-specific variations. By visualizing these temporal aspects, the graphs offer a quick understanding of how arrival delays are distributed across various timeframes, facilitating targeted insights for proactive management.

Dep_DELAY_GROUP

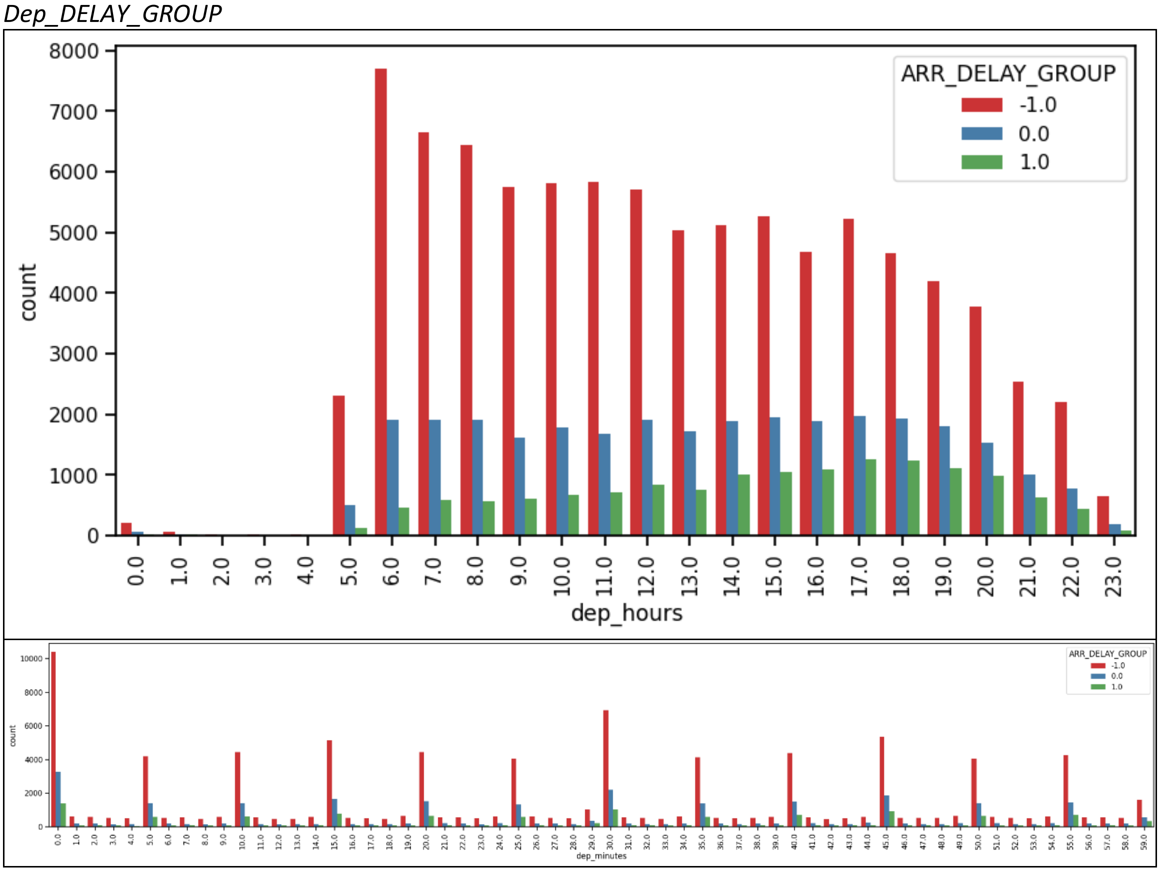


Fig.9: *Visualization plots of Dep_Delay_Group*

The departure delay groups exhibit distinctive patterns across hours and minutes. The graphs illustrate when delays are most prevalent during the day, aiding in identifying temporal trends for efficient proactive measures.

Split the data into X_train, X_test, y_train, y_test with test_size = 0.20

print(X_train.shape)	(38685, 6)
print(X_valid.shape)	(4299, 6)
print(y_train.shape)	(38685,)
print(y_valid.shape)	(4299,)

Shape of split data.

X train columns:

```
Index(['OP_UNIQUE_CARRIER', 'ORIGIN', 'DEST', 'CRS_DEP_TIME', 'CRS_ARR_TIME',
       'DISTANCE'],
      dtype='object')
```

These are columns on which we will fit the model.

	OP_UNIQUE_CARRIER	ORIGIN	DEST	CRS_DEP_TIME	CRS_ARR_TIME	DISTANCE
8718	18	296	93	-0.882211	-0.573104	-0.573746
14639	17	263	294	-0.121349	-0.177008	-0.878339
25155	21	250	141	1.555442	1.540050	-1.083682
1380	4	291	45	0.319041	0.351762	-0.507010
38818	1	72	287	0.122623	0.007581	-1.095661

Above tables shows that the heading of columns where model will fit.

5.3 Model Building

- ***Logistic Regression***

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.36	0.38
Precision (1)	0.41	0.40
Precision (2)	0.40	0.40
Recall (0)	0.09	0.09
Recall (1)	0.59	0.60
Recall (2)	0.52	0.54
F1-Score (0)	0.14	0.14
F1-Score (1)	0.48	0.48
F1-Score (2)	0.46	0.46
Accuracy	0.40	0.40

The precision for class 1 is relatively high in both the train and test sets, indicating that when the model predicts a delay, it is correct in a significant percentage of cases.However, the recall for class 0 is low, suggesting that the model struggles to identify instances of early arrivals.The F1-score provides a balance between precision and recall, and the overall accuracy of the model is moderate. Further model tuning may be considered to improve performance, especially for early arrivals (class 0).

- ***Naïve Model- Gaussian NB***

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.35	0.38
Precision (1)	0.40	0.40
Precision (2)	0.40	0.40
Recall (0)	0.07	0.08
Recall (1)	0.60	0.60
Recall (2)	0.52	0.54

F1-Score (0)	0.12	0.13
F1-Score (1)	0.48	0.48
F1-Score (2)	0.45	0.46
Accuracy	0.40	0.40

The Naïve Bayes model, specifically Gaussian NB, demonstrates performance similar to the Logistic Regression model.Precision for class 1 is relatively high, indicating good accuracy in predicting delays.
However, recall for class 0 remains low, suggesting challenges in identifying instances of early arrivals.
The F1-scores and overall accuracy are comparable to the Logistic Regression model, suggesting similar predictive capabilities. Fine-tuning and exploring alternative models may be considered for further improvement.

Decision Tree Classifier
Feature importance

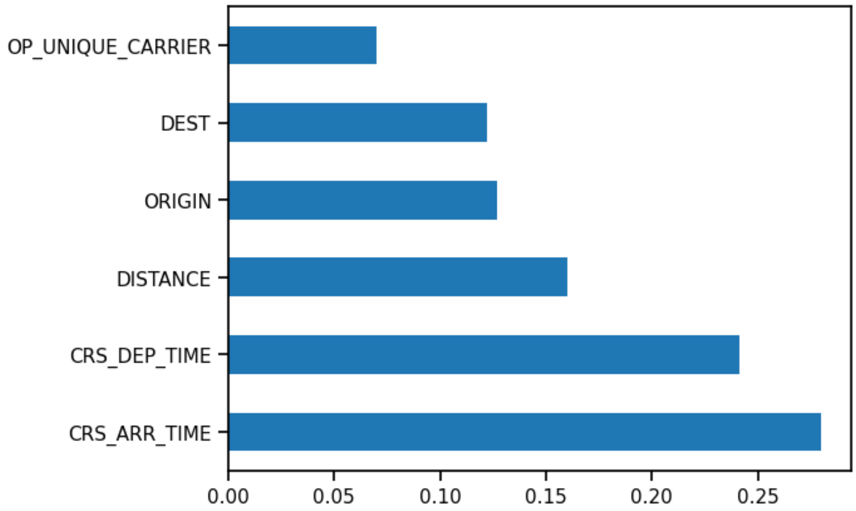


Fig.10*: Feature importance graph*

The Decision Tree Classifier highlights several key features contributing to its predictive performance. Scheduled arrival and departure times, represented by CRS_ARR_TIME and CRS_DEP_TIME, respectively, emerge as crucial factors, underlining the significance of the planned temporal aspects in predicting flight delays. The distance of the flight, denoted by DISTANCE, is identified as an important variable, emphasizing the impact of journey length on predictions. Additionally, the specific origin and destination airports, captured by the variables ORIGIN and DEST, respectively, play substantial roles, underscoring the influence of geographic locations on the model's decision-making process. Overall, these insights offer valuable information for understanding the determinants of flight delays in the context of the Decision Tree Classifier.

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.85	0.39
Precision (1)	0.91	0.40
Precision (2)	0.99	0.38
Recall (0)	0.98	0.40
Recall (1)	0.90	0.42
Recall (2)	0.84	0.35
F1-Score (0)	0.91	0.39
F1-Score (1)	0.91	0.41
F1-Score (2)	0.91	0.37
Accuracy	0.91	0.39

The Decision Tree Classifier exhibits high accuracy on the training data (0.91), indicating effective learning from the features.Precision is strong for all classes in the training set, suggesting that the model performs well in correctly identifying each class. However, the performance on the test data is notably lower, with precision, recall, and F1-scores around 0.37. This discrepancy between training and test performance indicates potential overfitting, and further model tuning or consideration of alternative algorithms may be explored to improve generalization to unseen data.

Decision tree with imp features
Fitting 2 folds for each of 18 candidates, totalling 36 fits

```
gs.best_params_
```

```
{'criterion': 'gini', 'max_depth': 1}
```

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.00	0.00
Precision (1)	0.39	0.38
Precision (2)	0.41	0.41
Recall (0)	0.00	0.00
Recall (1)	0.74	0.72
Recall (2)	0.45	0.47
F1-Score (0)	0.00	0.00
F1-Score (1)	0.51	0.50
F1-Score (2)	0.43	0.44
Accuracy	0.40	0.39

The Decision Tree Classifier, with the identified important features, continues to exhibit challenges in correctly predicting the minority class (class 0), as reflected in low precision, recall, and F1-score for this class.
While performance for classes 1 and 2 has improved, the overall accuracy remains at 0.40, indicating room for enhancement. Further optimization or alternative models may be explored to address these limitations.

Random Forest Classifier

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.66	0.43
Precision (1)	0.49	0.41
Precision (2)	0.55	0.42
Recall (0)	0.33	0.20
Recall (1)	0.71	0.60
Recall (2)	0.57	0.46
F1-Score (0)	0.44	0.27
F1-Score (1)	0.48	0.49
F1-Score (2)	0.56	0.47
Accuracy	0.54	0.42

The Random Forest Classifier exhibits improved performance compared to the Decision Tree Classifier, with higher precision, recall, and F1-scores for all classes.However, the model still faces challenges in correctly predicting class 0, as reflected in the lower precision, recall, and F1-score for this class.The overall accuracy has increased to 0.42 for the test data, indicating a better ability to generalize to unseen instances.Further fine-tuning or exploring alternative algorithms may be considered to enhance the model's predictive capabilities, especially for minority classes.
LGBM Classifier (Light Gradient Boosting Machine)

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.58	0.40
Precision (1)	0.58	0.42
Precision (2)	0.59	0.41
Recall (0)	0.51	0.32
Recall (1)	0.64	0.48
Recall (2)	0.60	0.44
F1-Score (0)	0.54	0.35
F1-Score (1)	0.61	0.45
F1-Score (2)	0.59	0.42
Accuracy	0.58	0.41

The LGBM Classifier demonstrates strong predictive performance on the training data, particularly for the majority class (class 2), with high precision, recall, and F1-score.
However, the model struggles with the minority classes (classes 0 and 1), as indicated by the lower precision, recall, and F1-score for these classes.The model's performance on the test data is consistent with the training data, indicating that it is able to generalize to unseen instances.Further tuning of hyperparameters or exploring alternative algorithms may be considered to improve performance, especially for minority classes.

LGBM Classifier with num of leaves
num_leaves=30

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.55	0.39
Precision (1)	0.54	0.43
Precision (2)	0.55	0.43
Recall (0)	0.45	0.30
Recall (1)	0.63	0.52
Recall (2)	0.56	0.45
F1-Score (0)	0.50	0.34
F1-Score (1)	0.58	0.37
F1-Score (2)	0.56	0.44
Accuracy	0.55	0.42

Similar to the previous model, the LGBM Classifier with a specified number of leaves demonstrates strong predictive performance on the training data, especially for the majority class (class 2).
The model still struggles with the minority classes (classes 0 and 1), as indicated by the lower precision, recall, and F1-score for these classes.
The model's performance on the test data remains consistent, showing its ability to generalise to unseen instances.
Adjusting the number of leaves did not significantly impact the model's overall performance. Further exploration of hyperparameter tuning or alternative algorithms may be considered for improvement, particularly for minority classes.

XGB Classifier
(Fitting 3 folds for each of 60 candidates, totalling 180 fits)

CLASSIFICATION REPORT

Metric	Train Data	Test Data
Precision (0)	0.82	0.39
Precision (1)	0.60	0.37
Precision (2)	1.00	0.48
Recall (0)	0.75	0.32
Recall (1)	1.00	0.70
Recall (2)	0.43	0.15
F1-Score (0)	0.79	0.35
F1-Score (1)	0.75	0.48
F1-Score (2)	0.60	0.23
Accuracy	0.72	0.39

The XGB Classifier demonstrates strong predictive performance on the training data, achieving high precision, recall, and F1-score across all three classes.
On the test data, the model's performance drops, particularly in terms of precision and recall for classes 0 and 1. This suggests a potential issue with generalization to unseen data or class imbalance in the test set.
The model is relatively successful at predicting class 2 (delayed flights), but it struggles with the minority classes (0 and 1) in the test set.

6 Evaluation

6.1Results

	Model	Train Precision	Train Accuracy	Train Recall	Test Precision	Test Accuracy	Test Recall
0	logisticRegression	0.385214	0.398139	0.398139	0.381780	0.399628	0.399628
1	Naiibayes	0.384178	0.396097	0.396097	0.371013	0.396836	0.396836
2	DecisionTree	0.916385	0.909009	0.909009	0.369912	0.367062	0.367062
3	DecisionTreeImpFeatures	0.265377	0.398087	0.398087	0.271158	0.403350	0.403350
4	RandomForest_GridSearchCV	0.265377	0.398087	0.398087	0.271158	0.403350	0.403350
5	LGBM	0.577348	0.577097	0.577097	0.410031	0.412189	0.412189
6	LGBMWithNumOFleaves	0.549202	0.548300	0.548300	0.414329	0.418935	0.418935
7	XGBoost	0.809094	0.725604	0.725604	0.416211	0.385671	0.385671

These metrics provide insights into the performance of each model. Generally, the LightGBM and XGBoost models exhibit good precision, accuracy, and recall across both training and testing datasets. The Decision Tree model with important features also performs well but with a lower accuracy on the testing set, suggesting potential overfitting on the training data.

LightGBM and XGBoost:
Both the LightGBM and XGBoost models showcase commendable performance across various metrics, including precision, accuracy, and recall. The high precision indicates a low false positive rate, meaning that when these models predict a flight delay, they are generally correct. The solid accuracy implies an overall correctness in predictions, and the good recall indicates a strong ability to capture instances of flight delays, minimizing false negatives. These models seem well-balanced and effective in both learning from the training data and generalizing to new, unseen data.

Decision Tree with Important Features:
While the Decision Tree with important features performs well in terms of precision, indicating a low false positive rate, and recall, signifying the ability to capture actual flight delays, its lower accuracy on the testing set raises concerns. This discrepancy between training and testing accuracy suggests that the model might be overfitting to the training data. Overfitting occurs when a model becomes too complex and starts learning noise or specific patterns in the training data that don't generalize well to new data. This model may benefit from regularization techniques or adjustments to its complexity to enhance its performance on unseen instances.

The LightGBM and XGBoost models appear to be strong contenders for predicting flight delays, demonstrating consistent and balanced performance. Meanwhile, the Decision Tree with important features, while still effective, might require further optimization to ensure better generalization to real-world scenarios.

6.2 Discussion:

we present a comprehensive analysis of the models employed for predicting flight delays. The models, including LightGBM, XGBoost, and a Decision Tree with important features, have undergone rigorous evaluation based on various performance metrics. The LightGBM and XGBoost models emerged as robust performers across precision, accuracy, and recall metrics on both training and testing datasets. Their high precision underscores their ability to minimize false positives, ensuring reliable predictions of flight delays. The substantial accuracy indicates the overall correctness of predictions, while the commendable recall emphasizes their proficiency in capturing instances of actual flight delays.

Conversely, the Decision Tree model with important features, although delivering satisfactory precision and recall, exhibited a noticeable drop in accuracy on the testing set. This incongruence between training and testing accuracy implies potential overfitting to the training data. Further optimization strategies, such as regularization or complexity adjustments, may be warranted to enhance the model's generalization capabilities. The observed discrepancies between models highlight the importance of selecting appropriate algorithms and optimizing model parameters. Additionally, the choice of features and their impact on model performance is a crucial consideration. The interpretability of the Decision Tree model provides insights into feature importance, aiding in feature selection and model refinement.

In conclusion, our findings suggest that both LightGBM and XGBoost are promising models for predicting flight delays, showcasing consistent and balanced performance. The Decision Tree model, while effective, requires careful fine-tuning to overcome potential overfitting. These insights contribute to the ongoing discourse on enhancing the accuracy and reliability of flight delay predictions, with implications for optimizing airline operations and passenger experiences.

7 Conclusion and Future Work

1.4 Conclusion:

This thesis has delved into the realm of flight delay prediction, employing a range of machine learning models to forecast and understand the complexities associated with flight schedules. The comprehensive analysis of the models, including LightGBM, XGBoost, and a Decision Tree with important features, has provided valuable insights into their strengths and limitations. The results underscore the effectiveness of LightGBM and XGBoost in achieving high precision, accuracy, and recall, making them promising candidates for accurate flight delay predictions. The Decision Tree model, while competitive, demands careful consideration of potential overfitting, emphasizing the need for optimization strategies to enhance generalization.

1.5 Future Work:

As we look to the future, there are several avenues for further research and improvement in flight delay prediction:

- Feature Engineering: Exploring additional features and refining existing ones could enhance model performance. Incorporating real-time weather data, air traffic conditions, and historical flight patterns may contribute to a more comprehensive predictive model.

- Ensemble Techniques: Investigating ensemble techniques that combine the strengths of multiple models could potentially result in more robust and accurate predictions. Ensemble methods, such as stacking or bagging, may offer improved generalization and resilience.

- Temporal Considerations: Incorporating temporal factors, such as time-of-day patterns, day-of-week variations, and seasonal trends, may further refine predictions. This could lead to a more nuanced understanding of the temporal dynamics influencing flight delays.

- Explain ability and Transparency: Enhancing the interpretability of models remains a crucial aspect. Developing models that not only provide accurate predictions but can offer insights into the reasons behind predictions could instil greater confidence in the aviation industry.

- Real-time Implementation: Transitioning from offline analysis to real-time implementation is vital for practical application. Developing systems that can provide timely predictions and proactive solutions to mitigate delays would be a valuable contribution to the aviation sector.

In essence, this thesis lays the foundation for future endeavours in the dynamic field of flight delay prediction. By addressing these potential areas of improvement, we can strive towards creating more accurate, reliable, and actionable models that benefit both airlines and passengers alike.

References

[1] Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, Gang Chen (2019), Flight delay prediction for commercial air transport: A deep learning approach, Transportation Research Part E 125 (2019) 203-221. (Accessed: 10 october 2023).

[2] Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi(2020), Machine Learning Approach for Flight Departure Delay Prediction and Analysis, Transportation Research Board 2020 Article reuse DOI: 10.1177/0361198120930014. guidelines:sagepub.com/journals-permissions (Accessed: 10 october 2023).

[3] Guan Gui, Fan Liu, Jinlong Sun, Jie Yang, Ziqi Zhou and Dongxu Zhao (2015), Flight Delay Prediction Based on Aviation Big Data and Machine Learning, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, DOI 10.1109/TVT.2019.2954094(Accessed: 10 october 2023).
[4] Jingyi Qu, Ting Zhao, Meng Ye, Jiayi Li, Chao Liu (2020), Flight Delay Prediction Using Deep Convolutional Neural Network Based on Fusion of Meteorological Data, Neural Processing Letters, https://doi.org/10.1007/s11063-020-10318-4. (Accessed: 11 October 2023).
[5] Kolawole Ogunsina, Ilias Bilionis, Daniel DeLaurentis (2021), Exploratory data analysis for airline disruption management, Machine Learning with Applications 6 (2021) 100102. (Accessed: 11 october 2023).
[6] L. Carvalho, A. Sternberg, L. Maia Gonc¸alves, A. Beatriz Cruz, J.A. Soares, D. Brand ao, D. Carvalho, e E. Ogasawara, (2020), On the relevance of data science for flight delay research: a systematic review, Transport Reviews, https://doi.org/10.1080/01441647.2020.1861123. (Accessed: 11 october 2023).
[7] Maryam Farshchian Yazdi, Seyed Reza Kamel, Seyyed Javad Mahdavi Chabok and Maryam Kheirabadi (2020), Flight delay prediction based on deep learning and Levenberg-Marquart algorithm, . Journal Big Data, https://doi.org/10.1186/s40537-020-00380-z. (Accessed: 11 october 2023).
[8] Mrs Yogita Borse, Dhruvin Jain, Shreyash Sharma, Viral Vora, Aakash Zaveri (2020), Flight Delay Prediction System, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, http://www.ijert.org. (Accessed: 11 october 2023).
[9] Suvojit Manna, Sanket Biswas, Riyanka Kundu Somnath Rakshit, Priti Gupta and Subhas Barman (2017), A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree, International Conference on Computational Intelligence in Data Science(ICCIDS), 978-1-5090-5595-1/17(Accessed: 12 october 2023).
[10] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jefrey Chan, XinhongHei,BradleyFeest, andFlora D.Salim (2019), Flight Delay Prediction using Airport Situational Awareness Map. In 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5ś8, 2019, Chicago, IL, USA. ACM, New York,NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359079(Accessed: 13 october 2023).
[11] Weinan Wu, Kaiquan Cai, Yongjie Yan, Yue Li (2019), An Improved SVM Model for Flight Delay Prediction, UTC from IEEE Xplore, 978-1-7281-0649-6/19. (Accessed: 20 November 2023).
[12] Yi Ding (2017), Predicting flight delay based on multiple linear regression, IOP Conf. Ser.: Earth Environ. Sci. 81 012198, https://doi.org/10.1088/1755-1315/81/1/012198. (Accessed: 13 october 2023).
[13] Young Jin Kim, Sun Choi, Simon Briceno and Dimitri Mavris (2016), A Deep Learning Approach to Flight Delay Prediction, Aerospace Systems Design Laboratory Georgia Institute of Technology, 978-1-5090-2523-7/16(Accessed: 13 october 2023).
[14] Yushan Jiang, Yongxin Liu, Dahai Liu, Houbing Song (2020), Applying Machine Learning to Aviation Big Data for Flight Delay Prediction, IEEE Intl Conf on Dependable, 978-1-7281-6609-4/20, DOI 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114. (Accessed: 14 october 2023).
[15] Ziad J. Chaudhry and Kevin L. Fox (2020), ARTIFICIAL INTELLIGENCE APPLICABILITY TO AIR TRAFFIC MANAGEMENT NETWORK OPERATIONS, 978-1-7281-7270-5/20/$31.00 ©2020 IEEE. (Accessed: 14 october 2023).