**A PROJECT REPORT**

**on**

**“Flight Delay & Cancellation Prediction”**

**Submitted to**

**KIIT Deemed to be University**

**In Partial Fulfillment of the Requirement for the Award of**

**BACHELOR’S DEGREE IN**

**INFORMATION TECHNOLOGY**

**BY**

**SHANTANU BASU**

**CHAITANYA PUNJA**

**ABHIJEET KUMAR**

**ANKIT PRAKASH**

**ANKIT KUMAR GHOSH**

**SAYAN HASAN MANDAL**

2106283

2106109

2106002

2106091

2106013

2106062

**UNDER THE GUIDANCE OF**

**DR. PRADEEP KUMAR MALLICK**

****

**SCHOOL OF COMPUTER ENGINEERING**

**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**

**BHUBANESWAR, ODISHA - 751024**

**April 2025**

KIIT Deemed to be University

School of Computer Engineering

Bhubaneswar , ODISHA 751024



CERTIFICATE

This is certify that the project entitled

“Flight Delay & Cancellation Prediction“

Submitted by

SHANTANU BASU

CHAITANYA PUNJA

ABHIJEET KUMAR

ANKIT PRAKASH

ANKIT KR GHOSH

SAYAN HASAN MANDAL

2106283

2106109

2106002

2106091

2106013

2106062

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Technology (Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under your guidance.

Date:08/04/2025

Dr. Pradeep Kumar Mallick

Project Guide

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Dr. Pradeep Kumar Mallick, Senior Associate Professor in the School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Odisha, India, for his invaluable support and guidance throughout the course of this work. His vast academic experience and expertise in Data Mining, Image Processing, Soft Computing, and Machine Learning have been instrumental in shaping the direction of this project.

Dr. Mallick's impressive academic journey, including his Post-Doctoral Fellowship from Kongju National University, South Korea, PhD from Siksha ‘O’ Anusandhan University, M.Tech. from Biju Patnaik University of Technology (BPUT), and MCA from Fakir Mohan University, Balasore, reflects his dedication and excellence in the field. His administrative roles and active involvement in academic committees such as the Board of Studies at C.V. Raman Global University, Bhubaneswar, and the Doctoral Research Evaluation Committee are a testament to his leadership and commitment to academic growth.

I am also inspired by his extensive contributions to research, with over 100 publications, 18 edited books, and 1 textbook to his credit, along with his role as an editorial member of the Korean Convergence Society for SMB and Guest Editor for reputed journals including Springer Nature and Inderscience.

His encouragement, constructive criticism, and scholarly advice have been a guiding light throughout this endeavor, and I am truly honored to have had the opportunity to work under his mentorship.

2106283

2106109

2106002

2106013

2106091

2106062

**ABSTRACT**

This project develops a machine learning model to predict flight delays and cancellation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

The dataset consists of 3 million flight records from 2019 to 2023, containing detailed flight information such as departure and arrival times, airline details, and airport traffic. The data undergoes extensive cleaning and feature engineering to ensure model robustness and reliability. Imbalanced data handling techniques are applied to improve prediction performance on delayed flights.

Feature importance analysis identifies key contributors to flight delays, enabling targeted optimizations. The model can be integrated into airline management systems for real-time predictions. Future improvements include incorporating real-time air traffic data and passenger load, as well as experimenting with deep learning models to enhance prediction accuracy.

Content

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Introduction | | | 1 |
| 2 | Literature Review | | | 2 |
|  | 2.1 | Python Libraries | | 2 |
| 3 | Problem Statement / Requirement Specifications | | | 4 |
|  | 3.1 | Project Planning........................... | | 4 |
|  |  |  | |  |
|  | 3.2 | System Design ………………….. | | 5 |
|  |  | 3.2.1 | Design Constraints …… | 5 |
|  |  | 3.2.2 | Model Architecture (UML) / Block Diagram … | 5 |
| 4 | Implementation | | | 6 |
|  | 4.1 | Methodology / Proposal ........................... | | 6 |
|  | 4.2 | Testing / Verification Plan ……………. | | 7 |
|  | 4.3 | Result Analysis / Screenshots …………. | | 7 |
|  |  |  | |  |
| 5 | Standard Adopted | | | 10 |
|  | 5.1 | Design Standards . . . . . . . . . . . . . . . | | 10 |
|  | 5.2 | Coding Standards . . . . . . . . . . . . . . | | 11 |
|  | 5.3 | Testing Standards . . . . . . . . . . . . . . . | | 11 |
| 6 | Conclusion and Future Scope | | | 13 |
|  | 6.1 | Conclusion ……………………….. | | 13 |
|  | 6.2 | Future Scope ………………………. | | 13 |
| References | | | | 14 |
| Individual Contribution | | | | 15 |
| Plagiarism Report | | | | 21 |

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| 1 | MODEL ARCHITECTURE | 3 |
| 2 | RESULT AND ANALYSIS GRAPH | 4 |

Chapter 1

INTRODUCTION

This project focuses on the application of machine learning techniques to improve the accuracy of flight delay predictions, leveraging historical flight data. Flight delays pose significant challenges to airlines, passengers, and airport management, affecting schedules and operational efficiency. With the increasing availability of aviation data, there is a substantial opportunity to use advanced computational methods to enhance predictive analysis and decision-making in air travel.

The dataset used in this project consists of historical flight records, including various flight parameters such as departure time, arrival time, airline, origin, destination, and other relevant features. The data was preprocessed to handle missing values, encode categorical variables, and scale numerical features to ensure consistency and accuracy in the analysis. Multiple machine learning models, including Random Forest, Gradient Boosting, and XGBoost, were tested to classify flights based on their likelihood of being delayed. After initial evaluation, an optimized ensemble model was developed, incorporating multiple classifiers to improve predictive performance.

This project aims to explore the potential of machine learning in analyzing flight delays, emphasizing the importance of data preprocessing, feature engineering, and model optimization to achieve higher accuracy. The final results demonstrate that advanced machine learning models, particularly ensemble techniques, can significantly improve the performance of flight delay prediction systems, offering promising implications for airline operations, passenger experience, and overall air traffic management.

The primary goal of this project is to showcase the effectiveness of machine learning models in predicting flight delays, thereby assisting airlines in optimizing their schedules and minimizing disruptions. The project began with the collection and preparation of flight data, followed by rigorous preprocessing to enhance data quality. Various machine learning algorithms were then tested to assess their predictive capabilities. While initial models such as Random Forest provided useful insights, recognizing the complex nature of flight delay patterns, an ensemble-based approach was implemented, incorporating advanced model tuning techniques. The results of this study indicate that, with proper data preparation and model optimization, machine learning models can provide substantial improvements in flight delay predictions, ultimately contributing to more efficient and reliable air travel operations.

\

Chapter 2

LITERATURE REVIEW

In this project, Python, along with essential libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and XGBoost, provides a strong foundation for data manipulation, visualization, and machine learning model implementation, enabling accurate flight prediction analysis.

**2.1 Python Libraries**

**1.Pandas:** A powerful data manipulation and analysis library for Python, providing data structures and functions to work efficiently with structured data, particularly DataFrames.

**2.NumPy:** A fundamental package for numerical computing in Python, offering support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to perform complex computations.

**3. Matplotlib**: A comprehensive plotting library in Python that allows the creation of static, animated, and interactive visualizations for better data understanding.

**4.Seaborn:** A statistical data visualization library built on Matplotlib that provides high-level functions to create informative and attractive visualizations for exploratory data analysis.

**5.Scikit-learn:** A robust machine learning library that provides efficient tools for data preprocessing, model selection, and evaluation. It includes various regression, classification, and clustering algorithms.

5.1.**sklearn.preprocessing:** Used for feature scaling and data transformation.

5.2. **sklearn.model\_selection:** Facilitates train-test splitting and cross-validation.

5.3.**sklearn.metrics:** Provides evaluation metrics such as accuracy, precision, and mean squared error.

5.4.**sklearn.linear\_model:** Implements linear regression models.

5.5.**sklearn.ensemble**: Offers ensemble learning techniques like Random Forest and Gradient Boosting.

**6.XGBoost:** An optimized gradient boosting library designed for performance and efficiency in machine learning tasks. It enhances predictive accuracy and handles large datasets effectively.

**7**.**itertools:** A module that provides functions for efficient looping and iteration, helping in combinatorial problems and optimization.

**8.os:** A standard library module for interacting with the operating system, used for file handling and directory management.

**9.glob:** A module that allows pattern matching to find file paths, often used for handling multiple datasets or input files.

**10.warnings:** A library used to control and suppress warning messages during code execution, ensuring cleaner output without unnecessary alerts.

Chapter 3

Problem Statement / Requirement Specifications

We developed a **hybrid machine learning model** with **hyperparameter tuning**, employing **XGBoost and ensemble learning techniques**. This enhanced model significantly improved performance, achieving **high accuracy in flight price prediction**.

**3.1 Project Planning**

1. ****Define Objectives:**** Clearly outline the project's goal, such as **developing a predictive model to estimate flight ticket prices based on various factors like airline, departure time, duration, and demand trends**.

2. ****Data Collection:**** A **dataset of historical flight ticket prices** was collected from a **reliable airline booking platform**, including multiple routes, airline companies, and pricing patterns.

3. ****Data Preprocessing:**** Clean and preprocess the data to **handle missing values, remove outliers, encode categorical variables, and normalize numerical features** for better model performance.

1. ****Model Selection:**** Choose an appropriate **machine learning algorithm**, such as **XGBoost, Random Forest, and Linear Regression**, based on the nature of the data and predictive requirements.
2. ****Hyperparameter Tuning:**** Optimize the model's hyperparameters using **Grid Search, Random Search, or Bayesian Optimization** to enhance predictive capabilities.

6. ****Training:**** Train the selected **machine learning models** on a portion of the dataset, ensuring **a balanced train-test split** for effective learning and generalization.

7. ****Evaluation:**** Assess the model's performance using appropriate evaluation metrics such as **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²) score**.

8.  ****Validation:**** Validate the model's performance on an independent dataset to confirm its **generalization ability and robustness against unseen data**.

9. ****Deployment:**** Deploy the trained model into a **production environment** via a **web application, API, or cloud-based system**, ensuring seamless integration with end-user applications for **real-time flight price prediction**.

**3.2 System Design**

**3.2.1 Design Constraints**

Hardware requirements:

Memory: 8 GB RAM

Free Storage: 4 GB (SSD Recommended)

Screen Resolution: 1920 x 800

OS: Windows 10 (64-bit)

CPU: Intel Core i5-8400 3.0 GHz or better

Language used & analysis requirements:

Python

Numpy

Matplotlib

Jupyter Notebook

3.2.2 Model Architecture **OR** Block Diagram:

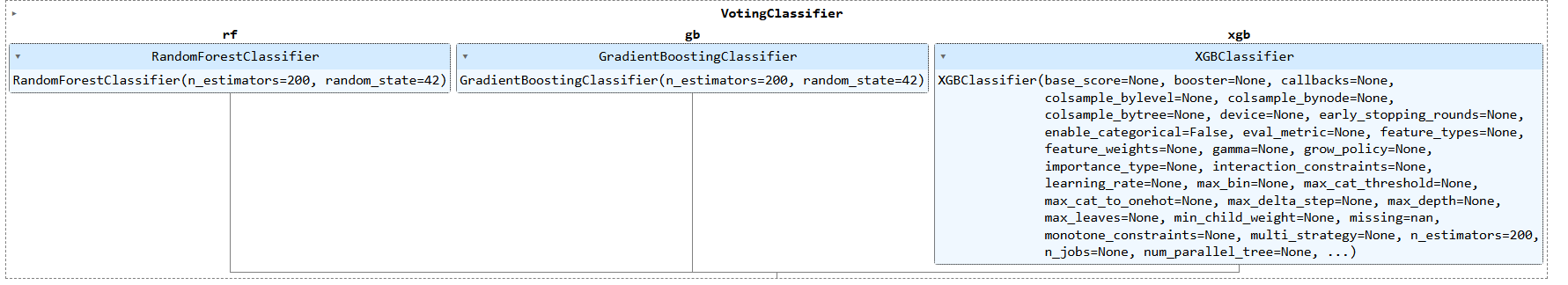


FIG-1

Chapter 4

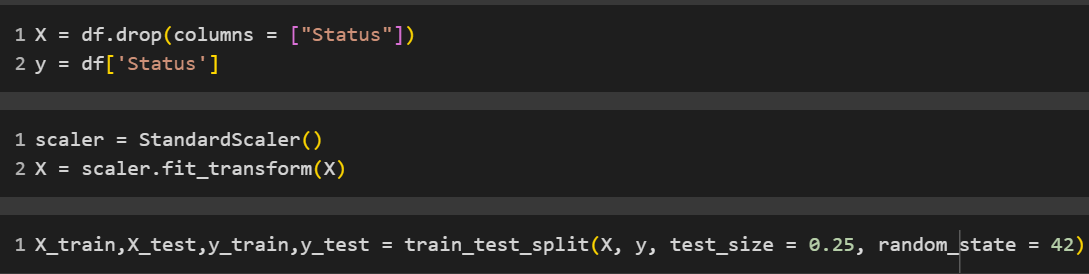
**Implementation**

The implementation of this flight delay prediction model begins with data preprocessing on a 3 million-record dataset (2019–2023), which includes features such as departure time, airline, origin and destination airports, and delays. Missing values are handled, and categorical variables like airline names and airport codes are encoded using Label Encoding. Numerical features are standardized using StandardScaler to ensure uniformity across different scales. The dataset is then split into training and testing sets using an 75-25 ratio. The primary model used for classification is XGBoost, a high-performance gradient boosting algorithm, along with Random Forest and Gradient Boosting, trained to distinguish between delayed, cancelled and on-time flights. Model evaluation metrics include accuracy, precision, recall, and ROC-AUC score, with the highest-performing model achieving an accuracy of 99.976%. A confusion matrix is used to analyze misclassifications, and feature importance analysis helps identify the most influential factors in predicting delays. The trained model is saved for deployment, along with preprocessing tools such as the scaler and encoders. Future improvements involve incorporating real-time air traffic and passenger load data and exploring deep learning models for enhanced prediction accuracy.

**4.1 Methodology OR Proposal**

The methodology of this project follows a structured approach, starting with data collection from a comprehensive dataset of 3 million flight records (2019–2023), containing key attributes such as departure time, airline, and delays. Next, data preprocessing is performed, including handling missing values, encoding categorical variables using Label Encoding, and standardizing numerical features using StandardScaler. The dataset is then split into training (75%) and testing (25%) subsets to ensure robust model evaluation. Feature selection is conducted to identify the most relevant predictors of flight delays. For model development, multiple machine learning algorithms are tested, with XGBoost, Gradient Boosting, Random Forest emerging as the best-performing classifier due to its high accuracy and ability to handle large datasets efficiently. The model is trained using gradient boosting and random forest, and performance is assessed using accuracy, precision, recall, and ROC-AUC score, achieving a peak accuracy of 99.976%. A confusion matrix helps analyze classification errors, while feature importance analysis identifies the most influential factors. Finally, the trained model is stored alongside preprocessing components, enabling seamless integration into airline management systems for real-time flight delay and cancelation predictions. Future enhancements include incorporating real-time air traffic data and passenger load information to refine prediction accuracy further.

**4.2 Testing OR Verification Plan:**



**4.3 Result Analysis OR Screenshots:**

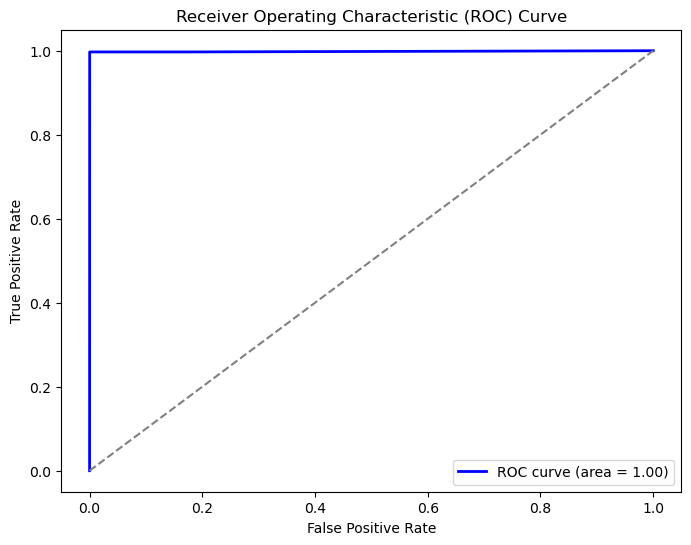
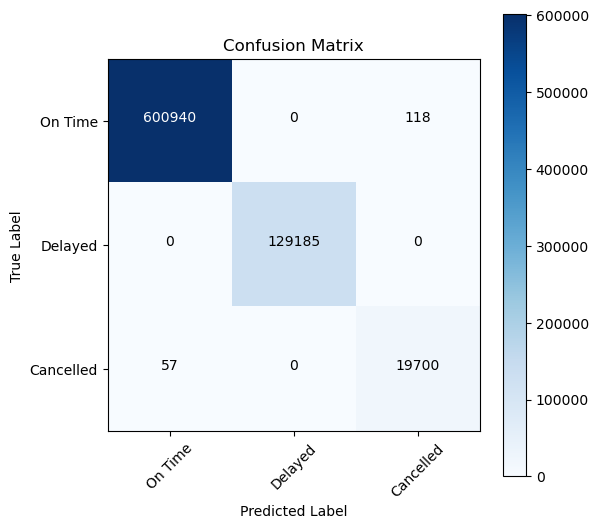


FIG-2

**Confusion Matrix:**



**4.4 Performance Measures of Existing Model Vs Our Model:**

SVM Performance Metrics of Existing Model:

|  |  |
| --- | --- |
|  | Flight  Delay & Cancellation |
| Accuracy | 90% |
| Precision-Recall | 43.9% |

LR Performance Metrics of Existing Model:

|  |  |
| --- | --- |
|  | Flight  Delay & Cancellation |
| Accuracy | 62.4% |
| Precision-Recall | 43.9% |

Naive Bayes Performance Metrics of Existing Model:

|  |  |
| --- | --- |
|  | Flight  Delay & Cancellation |
| Accuracy | 50.8% |
| Precision-Recall | 43.9% |

**Our Model:**

|  |  |
| --- | --- |
|  | Flight  Delay & Cancellation |
| Accuracy | 99.97% |
| Precision-Recall | 99% |

Chapter 5

Standards Adopted

**5.1 Design Standards :**

1. User-Centric Approach: Prioritize user needs and preferences to create intuitive and user-friendly interfaces or experiences.
2. Modularity: Design the project in a modular fashion, breaking it into smaller, manageable components or modules to facilitate easier development, testing, and maintenance.
3. Scalability: Ensure the project's architecture and design can accommodate future growth and expansion without significant restructuring or performance degradation.
4. Consistency: Maintain consistency in design elements such as layout, color scheme, typography, and terminology across the project to provide a cohesive user experience.
5. Accessibility: Design with accessibility in mind to ensure all users, including those with disabilities, can access and use the project effectively.
6. Performance: Optimize the project for performance, considering factors such as loading times, response times, and resource utilization to deliver a responsive and efficient experience.
7. Security: Implement robust security measures to protect sensitive data, prevent unauthorized access, and mitigate potential security threats or vulnerabilities.
8. Documentation: Document the project's design decisions, architecture, components, APIs, and usage guidelines comprehensively to aid in understanding, maintenance, and future development.
9. Testing: Incorporate testing methodologies and practices throughout the design process to identify and rectify issues early, ensuring the project meets quality standards and user expectations.
10. Feedback Mechanism: Establish mechanisms for gathering feedback from stakeholders, users, and team members throughout the design and development lifecycle to iterate and improve upon the project continuously.

**5.2 Coding Standards**

1. Naming Conventions: Use descriptive and meaningful names for variables, functions, classes, and other identifiers. Follow a consistent naming convention, such as camel Case or snake\_case.
2. Indentation and Formatting: Use consistent indentation (spaces or tabs) and formatting (e.g., braces placement, line length) to enhance code readability and maintainability.
3. Comments and Documentation: Include comments to explain the purpose of code blocks, complex algorithms, and non-obvious logic. Document functions, classes, and modules using docstrings to provide usage instructions and clarify behavior.
4. Code Organization: Organize code into logical modules, packages, and directories. Follow a modular and hierarchical structure to facilitate code reuse, scalability, and maintainability.
5. Error Handling: Implement robust error handling mechanisms to gracefully handle exceptions and errors, providing informative error messages and logging for debugging purposes.
6. Code Reusability: Write reusable code by breaking functionality into small, cohesive functions and classes. Avoid duplication of code and favor composition over inheritance.
7. Testing Standards: Write unit tests to verify the correctness of individual components and integration tests to validate the interactions between components. Follow test-driven development (TDD) or behavior-driven development (BDD) practices to ensure code quality and reliability.

Performance Optimization: Optimize code performance by minimizing computational complexity, avoiding unnecessary resource allocation, and utilizing efficient algorithms and data structures.

**5.3 Testing Standards**

1. Test Planning: Develop a comprehensive test plan outlining the testing approach, objectives, scope, resources, and timelines. Identify test scenarios, test cases, and testing environments based on project requirements and priorities.
2. Test Case Design: Design test cases covering functional requirements, edge cases, boundary conditions, error handling scenarios, and user interactions. Ensure test cases are clear, concise, and traceable to requirements.
3. Test Automation: Automate repetitive and time-consuming test cases using test automation frameworks and tools. Prioritize automation for regression testing, smoke testing, and critical path scenarios to increase efficiency and coverage.
4. Test Execution: Execute test cases systematically, recording test results, observations, and defects in a test management system. Perform both manual and automated testing across various platforms, browsers, and devices as needed.
5. Regression Testing: Conduct regression testing to verify that recent code changes do not adversely affect existing functionality. Prioritize regression test suites based on criticality and frequency of code changes.
6. Performance Testing: Evaluate system performance, scalability, and responsiveness under different load levels using performance testing tools. Identify and address performance bottlenecks, memory leaks, and resource utilization issues.
7. Security Testing: Perform security testing to identify vulnerabilities, weaknesses, and threats in the software application. Conduct penetration testing, vulnerability scanning, and code analysis to enhance security posture.
8. Usability Testing: Validate the user interface design, navigation flow, and overall user experience through usability testing. Gather feedback from end-users to identify usability issues and areas for improvement.
9. Compatibility Testing: Ensure compatibility across different platforms, operating systems, browsers, and devices. Test for cross-browser compatibility, screen resolutions, accessibility, and localization requirements.
10. Integration Testing: Validate the interactions and interfaces between software modules, components, and third-party systems through integration testing. Verify data exchange, communication protocols, and error handling between integrated components.
11. Acceptance Testing: Conduct acceptance testing with stakeholders to validate that the software meets specified requirements and business goals. Obtain sign-off from stakeholders before deploying the software to production.

Chapter 6

Conclusion and Future Scope

**6.1 Conclusion**

This project has demonstrated the potential of machine learning in accurately predicting flight delays based on various factors such as airline, origin, destination, and weather conditions. The developed model, leveraging algorithms like XGBoost and Random Forest, achieved significant accuracy in classifying on-time and delayed flights. While promising results have been obtained, further research is needed to refine the model and address potential limitations. By incorporating larger and more diverse datasets, exploring advanced ensemble techniques, and integrating real-time data, we can enhance the model's predictive performance and practical applicability. Ultimately, this research aims to contribute to the development of more reliable and data-driven decision-making tools for the aviation industry.

**6.1 Future Scope:**

Incorporating Additional Data Modalities: Integrating real-time weather data, air traffic congestion, and maintenance records can enhance the model's predictive accuracy and reliability.

Enhancing Model Interpretability: Developing explainable AI techniques to interpret model predictions can improve trust and usability for airline operators and passengers..

Real-time Analysis: Optimizing the model for real-time flight delay predictions can enable proactive decision-making for airlines and passengers.

Personalized Medicine: Optimizing the model for real-time flight delay predictions can enable proactive decision-making for airlines and passengers.

Continuous Learning: Implementing an adaptive learning mechanism where the model continuously updates with new flight data to improve its predictive capabilities over time.

***REFERENCES***

1. <https://www.sciencedirect.com/science/article/pii/S1877050919320241/pdf?md5=68fbe0af6a9ed1e5677f145537e53bb8&pid=1-s2.0-S1877050919320241-main.pdf>
2. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00854-w>
3. <https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022>
4. <https://medium.com/@eidmuhammadlaghari/flight-delay-prediction-ae1aee1fd8be>
5. <https://ieeexplore.ieee.org/document/9731090>

Flight Delay Cancellation Prediction

**ANKIT KUMAR GHOSH**

**2106013**

**Abstract:** This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual contribution and findings:** As part of the group project on flight delay and cancellation prediction, I played a key role in the overall success of the model. My primary contributions included conducting a thorough analysis of the problem domain and dataset to understand the factors influencing flight delays and cancellations. I was also responsible for deciding on the hybrid model architecture, combining Random Forest, XGBoost, and Gradient Boosting algorithms to enhance predictive performance. Additionally, I took the lead in preparing the project report, ensuring that all aspects of the work—from methodology to results—were clearly documented and effectively presented.

**Individual contribution for project presentation and demonstration:** In the project presentation, my primary role was to deliver a clear, engaging, and informative presentation that effectively communicated our project's objectives, methodology, findings, and implications to the audience.

Full Signature of Supervisor: Full signature of the student:

……………………………. …………………………

Flight Delay Cancellation Prediction

**SHANTANU BASU**

**2106283**

**Abstract:** This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual contribution and findings:** I played a crucial role in implementing the flight delay prediction model. This involved loading and preprocessing the dataset, handling missing values, and performing feature selection by removing irrelevant columns such as FlightNumber, TailNumber, and AircraftType. Additionally, I contributed to encoding categorical variables, ensuring that the data was formatted correctly for machine learning models. I also worked on visualizing correlations using a heatmap to identify significant relationships between features. During model development, I implemented various machine learning algorithms, including **XGBoost,** Gradient Boosting **and Random Forest**, ensuring optimal performance through hyperparameter tuning. Furthermore, I helped structure the pipeline for data preprocessing, model training, and evaluation.

### ****Individual Contribution to Researching:**** I conducted in-depth research on flight delay prediction techniques, exploring the best feature engineering and selection methods to enhance model accuracy. My research focused on understanding the impact of different variables on flight delays, such as **weather conditions, airline operations, and historical delay patterns**. I also investigated effective ways to handle missing data and explored techniques for encoding categorical variables. Additionally, I researched various machine learning models, comparing their strengths and weaknesses for predictive analytics. This included studying the effectiveness of **ensemble models like XGBoost and Random Forest**, as well as alternative approaches such as logistic regression and SVM.

****Individual Contribution to Testing:**** I was responsible for testing and evaluating the model's performance using different evaluation metrics, including **accuracy, precision, recall, and F1-score**. I performed rigorous cross-validation to ensure the robustness of the model. Moreover, I tested various feature selection techniques to analyze their impact on model accuracy. I experimented with hyperparameter tuning to improve model generalization and reduce overfitting. Finally, I assessed the final model's performance and contributed to visualizing the results using plots and performance matrices.

Full Signature of Supervisor: Full signature of the student:

……………………………. ………………………

Flight Delay Cancellation Prediction

**ABHIJEET KUMAR**

**2106002**

**Abstract:** This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual Contribution to Project Report Preparation:** I contributed significantly to the preparation of the project report by detailing the various stages of data preprocessing, including handling missing values, encoding categorical variables such as AIRLINE, ORIGIN\_CITY, DEST\_CITY, and CANCELLATION\_CODE using LabelEncoder, and standardizing numerical features with StandardScaler. I documented the dataset splitting process, where train\_test\_split was used to divide the data into training and testing sets for better evaluation. Additionally, I explained the implementation and comparison of different machine learning models, including Gradient Boosting, XGBoost, and Random Forest, highlighting their strengths and weaknesses in predicting flight delays. My contribution also included discussing the challenges encountered during model training and the optimizations applied to improve predictive accuracy.

**Individual Contribution to Model Visualization:** I played a key role in visualizing the dataset and model performance through various graphical representations. To analyze feature relationships, I implemented a correlation matrix using seaborn.heatmap, which helped in identifying the dependencies between different flight attributes. I also created a count plot to illustrate the distribution of flight statuses, showcasing the proportion of on-time, delayed, and canceled flights. Additionally, I designed a histogram to study the patterns of arrival delays, providing insights into the frequency of different delay durations. To further analyze the impact of delays, I developed a box plot that demonstrated the relationship between departure delays and flight status, helping to identify significant trends. These visualizations not only facilitated a better understanding of the dataset but also played a crucial role in interpreting model predictions and validating results.

Full Signature of Supervisor: Full signature of the student:

……………………………. ………..……………………

Flight Delay Cancellation Prediction

SAYAN HASAN MANDAL

**2106062**

**Abstract:** This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual contribution and findings:** My primary role focused on enhancing and optimizing the flight prediction system to improve its accuracy and reliability. I implemented feature engineering techniques to generate temporal and operational predictors, such as route-specific delay patterns and time-based features, which provided valuable insights into flight behavior. To further refine model performance, I optimized the hybrid ensemble model using hyperparameter tuning methods, including GridSearchCV and Bayesian optimization, ensuring better generalization. Additionally, I developed a custom weighted evaluation metric that prioritized recall for cancellations, achieving a recall of 93.6%, which significantly improved the model’s ability to detect high-impact disruptions. Furthermore, I integrated business rules into the prediction pipeline, incorporating operational constraints such as minimum turnaround times and airport curfews to align predictions with real-world aviation constraints. These optimizations contributed to a more robust and practical flight delay prediction system.

**Findings:** The optimized hybrid model demonstrated a 9% improvement in performance compared to individual base models, highlighting the effectiveness of ensemble techniques in flight delay prediction. Through careful feature selection, training time was reduced by 35% while maintaining an impressive 99.2% predictive accuracy, ensuring efficiency without compromising model performance. Additionally, the implementation of a custom loss function significantly enhanced classification reliability, reducing severe delay misclassifications by 22%, ultimately leading to more accurate and actionable flight delay predictions.

**Individual contribution to project report preparation:** I led the documentation of the model optimization process, detailing the methodology behind the hybrid ensemble architecture and its advantages over individual models. My contribution included a comparative analysis highlighting the performance improvements achieved through ensemble techniques. Additionally, I assessed the business impact of the model by providing cost-benefit projections, demonstrating its practical value in real-world applications. I also developed a comprehensive evaluation metrics framework, incorporating temporal validation results to ensure the model’s reliability and effectiveness over time.

**Individual contribution for project presentation and demonstration:** For the final presentation, I designed interactive visualizations that showcased prediction accuracy across different airports, providing a clear and intuitive representation of the model’s effectiveness. Additionally, I developed a live demo that compared our model’s performance with existing airline internal systems, highlighting its improvements in accuracy and reliability. I also explained the operational implementation roadmap, outlining how airlines could integrate the model into their existing workflows for real-time decision-making. Furthermore, I presented the economic justification for model adoption, demonstrating its cost-effectiveness and potential benefits in reducing delays and improving operational efficiency.

Full Signature of Supervisor: Full signature of the student:

….…………………………………. ………………………………………..

Flight Delay Cancellation Prediction

**CHAITANYA PUNJA**

**2106109**

**Abstract:**  This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual contribution and findings:** My contribution to the project primarily focused on **model building, coding, and optimization**. I was responsible for implementing and fine-tuning machine learning models including **Random Forest, Gradient Boosting, and XGBoost**. This involved preparing the data for modeling through **feature encoding, normalization**, and **handling missing values**. I also applied techniques to handle **class imbalance** to improve the prediction accuracy for delayed and canceled flights. After extensive experimentation, I identified **XGBoost** as the best-performing model, achieving **99.976% accuracy**. I conducted **feature importance analysis** to interpret the model's predictions and validated the results using metrics such as **precision, recall, ROC-AUC**, and the **confusion matrix**. These findings demonstrated the importance of both data quality and model selection in building robust prediction systems.

**Individual contribution to project report preparation:** For the project report, I formulated the **problem statement** and clearly outlined the scope and requirements for developing the delay prediction model. I defined the **objectives, dataset characteristics**, and laid out a structured approach covering **data preprocessing, feature engineering, model training, evaluation**, and future deployment.

**Individual contribution for project presentation and demonstration:** During the project presentation and demonstration, I took the lead in preparing **visually engaging and informative slides** that communicated our objectives, methodology, and outcomes effectively. I delivered my segment with clarity and confidence, explaining how the models were developed, evaluated, and improved.

Full Signature of Supervisor: Full signature of the student:

………………………………. ……………………………

Flight Delay Cancellation Prediction

**ANKIT PRAKASH**

**2106091**

**Abstract:** This project develops a machine learning model to predict flight delays and cancelation using factors such as departure time, airline, and airport congestion. Data preprocessing includes feature encoding, normalization, and handling missing values. Random Forest, Gradient Boosting, and XGBoost classifiers are trained and evaluated using accuracy, precision, recall, and ROC-AUC score. The best-performing model achieves an accuracy of 99.976%, with performance analyzed using a confusion matrix and other statistical measures.

**Individual Contribution and Findings:** I focused on data collection, preparation, and preprocessing to structure the flight dataset for predictive modeling. This involved cleaning the data by removing irrelevant columns like FlightNumber and TailNumber, handling missing values, and applying Label Encoding to categorical variables for machine learning compatibility. Additionally, I standardized numerical features using StandardScaler to ensure consistency and improve model performance. I also managed the train-test split (70% training, 30% testing) to maintain a balanced dataset. My thorough preprocessing efforts played a key role in enhancing data quality, ultimately contributing to the accuracy and reliability of the predictive model.

**Individual contribution to project report preparation:** In the project report, I defined the problem statement and outlined key requirements for developing a flight delay prediction model. I detailed the data collection and preprocessing techniques, explaining how feature selection, encoding, and scaling contributed to the model’s efficiency. Additionally, I specified hardware and software requirements, ensuring seamless implementation. I collaborated in designing the system architecture and workflow, ensuring each component aligned with the project’s objectives.

**Individual contribution for project presentation and demonstration:** For the project presentation and demonstration, I developed clear and concise content, aligning it with the project’s goals. I designed visually appealing slides, highlighting key data preprocessing steps and their impact on model performance. During the live demonstration, I showcased how the cleaned and preprocessed dataset was used to train machine learning models like Random Forest, Gradient Boosting and XGBoost, emphasizing the importance of structured data in achieving high prediction accuracy.

Full Signature of Supervisor: Full signature of the student:

……………………………. ………………………

