****

**University of Bahrain**

**College of Information Technology**

**Department of Computer Science**

**ITCS494**

Machine Learning Project

Abdullah Aktham Khaleel 202004678

*Musherah Moqbel Ali Alzoubah 202002276*

Jaber A.Hussain Ebrahim Buhusain 202010303

Submitted to

Dr. Faisal Al-Khateeb

# 1. Introduction

# The aim of this project is to develop machine learning models to predict flight prices using a dataset obtained from Kaggle. Accurate prediction of flight prices can significantly benefit both travelers and airlines. For travelers, it provides the ability to make informed decisions about when to purchase tickets, potentially saving money and ensuring better travel planning. For airlines, understanding and predicting price fluctuations can help optimize pricing strategies, maximize revenue, and manage demand more effectively.

# Flight pricing is influenced by a multitude of factors, including the airline, departure and arrival cities, flight duration, number of stops, and the time remaining until the flight. The complexity of these factors and their interactions make flight price prediction a challenging task that is well-suited for machine learning techniques.

# This report presents the results of three machine learning models—Naive Bayes, Random Forest, and XGBoost—trained and evaluated on the flight price dataset. Each model has unique characteristics and strengths, making them valuable for comparison:

# Naive Bayes: A probabilistic model based on Bayes' theorem, typically used for classification. Despite its simplicity and assumptions of feature independence, it is applied here to see how it performs in a regression context for predicting flight prices.

# Random Forest: An ensemble learning method that constructs multiple decision trees and outputs the mean prediction of individual trees. It is known for its ability to handle a large number of features and its robustness against overfitting, making it a strong candidate for flight price prediction.

# XGBoost: An optimized gradient boosting algorithm that builds an ensemble of decision trees sequentially, with each tree correcting the errors of its predecessor. XGBoost is renowned for its performance and speed, often winning machine learning competitions due to its ability to model complex relationships in the data.

# To ensure the robustness of the models, we also perform K-fold cross-validation, which involves splitting the dataset into multiple folds and training the model multiple times, each time using a different fold as the test set. This technique helps in assessing the generalization ability of the models and mitigates the risk of overfitting.

# The results from these models are compared using Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual flight prices. A lower MSE indicates better model performance. This comprehensive analysis aims to identify the most effective model for flight price prediction and provide insights into the factors influencing flight prices.

# By leveraging machine learning, this project aims to enhance the accuracy of flight price predictions, thereby benefiting both consumers and airlines through better decision-making and optimized pricing strategies

# .2. Dataset Description

The dataset used in this project consists of various features that influence flight prices. The features include:

- airline: The airline operating the flight  
- flight: Flight identifier  
- source\_city: City from which the flight departs  
- departure\_time: Time of departure  
- stops: Number of stops during the flight  
- arrival\_time: Time of arrival  
- destination\_city: City of arrival  
- class: Travel class (e.g., Economy, Business)  
- duration: Duration of the flight  
- days\_left: Number of days left until the flight  
- price: The price of the flight ticket (target variable)

# 3. Model Descriptions

In this project, three machine learning models were used:

- Naive Bayes: A probabilistic model based on Bayes' theorem, typically used for classification but applied here for regression.  
- Random Forest: An ensemble learning method that constructs multiple decision trees and outputs the mean prediction of individual trees, reducing overfitting and improving accuracy.  
- XGBoost: An optimized gradient boosting algorithm that builds an ensemble of decision trees sequentially, with each tree correcting the errors of its predecessor.

# 4. Methodology

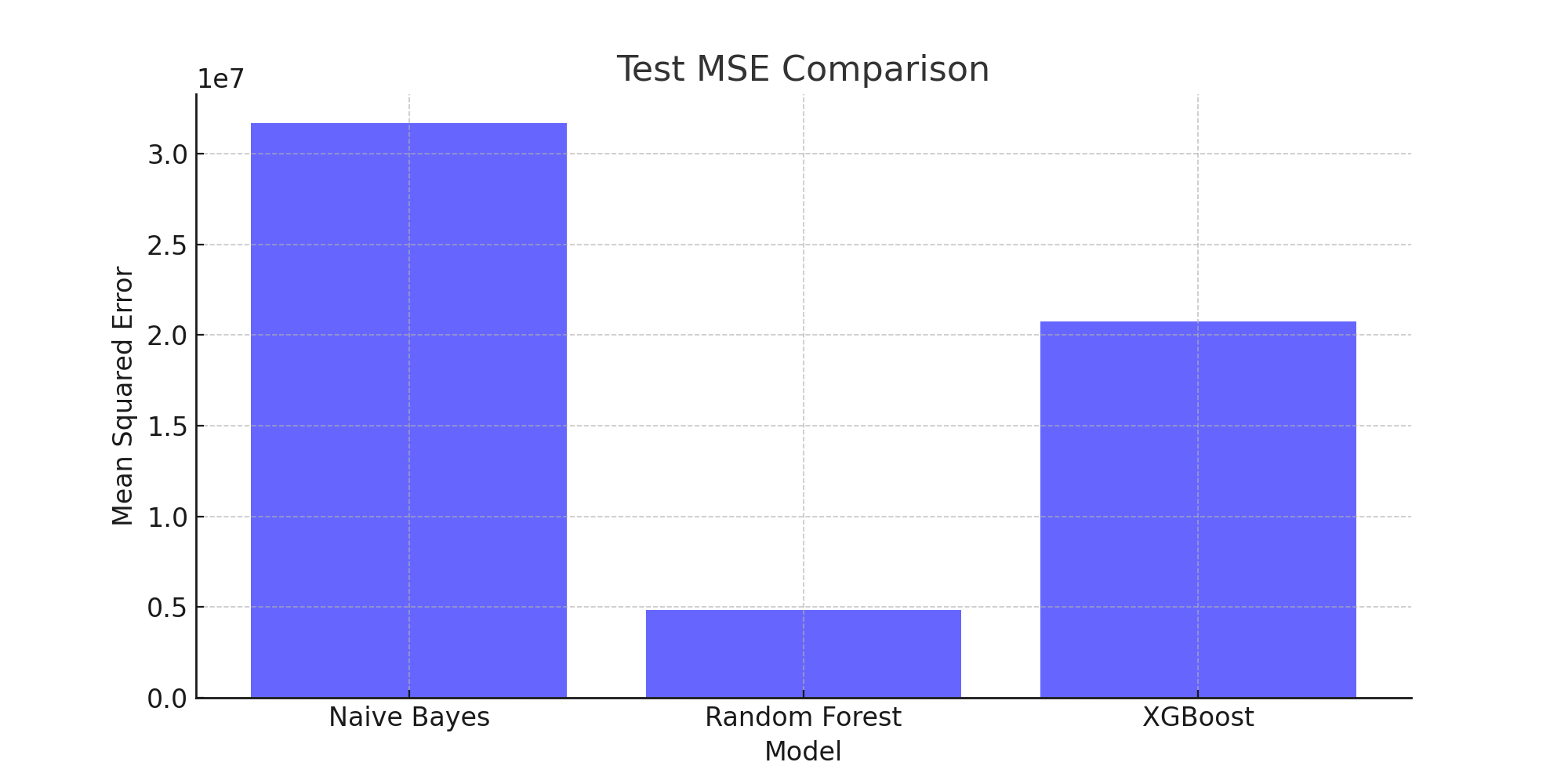
1. Data Preprocessing:  
 - Categorical features were encoded using LabelEncoder.  
 - Rows with missing values were dropped.  
 - The dataset was split into training (70%) and testing (30%) sets, with the last 15 records set aside for future prediction.  
2. Model Training and Evaluation:  
 - Three models—Naive Bayes, Random Forest, and XGBoost—were trained on the training set and evaluated on the testing set using Mean Squared Error (MSE).  
 - K-fold cross-validation with 5 folds was performed to assess the robustness of the models.

# 5. Results and Discussion

|  |  |  |
| --- | --- | --- |
| Model | Test MSE | K-Fold CV MSE |
| Naive Bayes | 31697913.10 | 32089536.04 |
| Random Forest | 4827492.45 | 5190785.07 |
| XGBoost | 20749006.23 | 20572976.29 |

The following table summarizes the MSE results for each model:

Test MSE Comparison:



# 6. Conclusion

In this analysis, three machine learning models—Naive Bayes, Random Forest, and XGBoost—were trained to predict flight prices. The Random Forest model outperformed the other models, achieving the lowest Mean Squared Error on both the test set and K-fold cross-validation. These results suggest that Random Forest is the most suitable model for this dataset, providing accurate predictions of flight prices.