Flight Fare Prediction Project Report

# 1. Introduction

The objective of this project is to predict airline ticket prices based on multiple features such as airline, source, destination, total stops, duration, and timings. Accurate flight fare prediction can enhance customer decision-making and support pricing strategies for travel platforms.

# 2. Dataset Overview

The dataset includes flight-related attributes such as airline, route, duration, stops, and timings, along with the target variable: Price. The dataset is pre-split into training and test sets for model evaluation.

# 3. Data Preprocessing

Key preprocessing steps included:  
- Parsing date and time fields into structured features (e.g., Journey day, month, hour).  
- Converting 'Duration' to total minutes.  
- Handling missing values and dropping redundant columns post-transformation.  
- Label encoding ordinal categorical features and one-hot encoding nominal features.

# 4. Exploratory Data Analysis (EDA)

EDA was performed using Seaborn and Matplotlib to analyze distributions, detect outliers, and evaluate feature relationships. Visualizations included box plots (Airline vs Price), count plots, and correlation heatmaps.

# 5. Feature Engineering

Features were extracted from timestamp columns and encoded for machine learning. Feature importance was assessed using ExtraTreesRegressor, revealing that 'Total\_Stops', 'Duration', and 'Airline' were the most influential variables.

# 6. Model Building

Several regression models were built and evaluated:  
- Linear Regression  
- Lasso and Ridge Regression  
- Decision Tree Regressor

-Random Forest Regression(with RandomizedSearchCV)

# 7. Model Comparison

|  |  |  |
| --- | --- | --- |
| Model | RMSE | R² Score |
| Linear Regression | ≈ 2700 | ≈ 0.80 |
| Ridge Regression | ≈ 2650 | ≈ 0.82 |
| Lasso Regression | ≈ 2680 | ≈ 0.81 |
| Decision Tree Regressor | ≈ 2200 | ≈ 0.88 |
| Random Forest Regressor | ≈ 2000 | ≈ 0.90 |

Random Forest performed the best, especially after hyperparameter tuning using RandomizedSearchCV.

# 8. Evaluation Metrics (KPIs)

* Root Mean Squared Error (RMSE): Evaluates prediction error magnitude. Lower is better.
* R² Score: Indicates variance explained by the model. Closer to 1 is better.
* Feature Importance: Identified key drivers of price prediction using tree-based models.
* Training vs Test Comparison: Ensures the model generalizes well and avoids overfitting.

# 9. Conclusion

The Flight Fare Prediction project demonstrates how machine learning can be applied to solve real-world regression tasks. After comparing multiple models, Random Forest Regressor emerged as the top performer with the highest accuracy and lowest error. Feature engineering and proper data transformation were crucial in model success. This approach can be extended into a production-grade system for travel agencies and fare comparison platforms.