Flight Fare Prediction – Project Report

1. Introduction

Air travel has become one of the most common modes of transport, and ticket prices often fluctuate depending on multiple factors such as airline, time of booking, number of stops, and travel duration. Predicting flight fares accurately can help travelers make informed choices and airlines optimize pricing strategies. This project focuses on developing a predictive model for flight fares using statistical methods and machine learning.

---

2. Objective

The objective of this project is to:

- Predict flight fares based on historical data.
- Identify the key factors influencing ticket prices.
- Compare different machine learning models and select the best-performing one.

---

3. Dataset Description

The dataset contains information on flight details and prices. Key features include:

- **Airline** – Name of the airline (e.g., Jet Airways, IndiGo, Air India).
- **Date of Journey** – Date when the flight is scheduled.
- **Source & Destination** – Cities of departure and arrival.
- **Route** – Full flight route, including stopovers.
- **Departure & Arrival Time** – Scheduled departure and arrival.
- **Duration** – Total flight duration.
- **Total Stops** – Number of stops (non-stop, 1 stop, 2 stops, etc.).

- **Price** – Target variable, representing the ticket price.

---

4. Data Preprocessing

To ensure the dataset is suitable for modeling, the following preprocessing steps were carried out:

- **Handling Missing Values** – Imputed missing records using statistical methods.
- **Duplicate Removal** – Removed duplicate records to avoid bias.
- **Feature Engineering**:
    - Extracted **day** and **month** from the journey date.
    - Derived **hour** and **minute** from departure and arrival times.
    - Converted duration into minutes.
- **Encoding Categorical Variables** – Used one-hot encoding for categorical features (e.g., Airline, Source).
- **Outlier Detection** – Identified outliers using the IQR method to handle extreme price variations.

---

5. Exploratory Data Analysis (EDA)

EDA was performed to understand relationships between variables:

- **Histograms & Boxplots** – Showed that Jet Airways had the highest price variation.
- **Pivot Tables** – Demonstrated how prices varied with airlines and number of stops.
- **Correlation Analysis** – Duration and total stops were found to be strong predictors of price.

---

6. Statistical Analysis

- **T-tests** were conducted to check whether average ticket prices differ significantly between airlines.
- **Chi-Square Tests** were used to examine the relationship between categorical variables such as **Source** and **Price categories**.
- **Stationarity Tests** (ADF test) were applied when analyzing time trends in prices.

---

7. Model Development

Multiple models were trained to predict flight fares:

1. **Linear Regression** – Baseline model, simple but less accurate.
2. **Lasso & Ridge Regression** – Regularized models to handle multicollinearity.
3. **Decision Tree Regressor** – Captured nonlinear relationships.
4. **Random Forest Regressor** – Outperformed other models due to ensemble averaging.

Model Evaluation Metrics:

- Root Mean Square Error (RMSE)
- R^2 Score

Best Model: Random Forest, with the lowest RMSE and highest R^2.

---

8. Results & Insights

- Jet Airways tickets were the most expensive with high variability.
- Non-stop flights were cheaper compared to 1-stop or 2-stop flights.
- Flight duration and number of stops were the most important predictors of price.
- Random Forest achieved the best performance compared to linear models.

---

9. Conclusion

This project demonstrated how flight fares can be predicted using machine learning models. The Random Forest model provided the most accurate predictions, and insights revealed that airline type, number of stops, and flight duration are key factors in pricing.

---

10. Future Work

- Incorporate additional features like booking time and seasonal effects.
- Use advanced models such as Gradient Boosting (XGBoost, LightGBM).
- Build a real-time fare prediction system for deployment.