Flight Price and Destination Prediction Project

🛬 Project Overview

This project focuses on predicting flight prices and destinations using various machine learning models. We analyzed a dataset containing information about different flights, including airlines, origins, destinations, travel dates, and prices. Our goal was to develop accurate prediction models that could provide valuable insights for both consumers and airlines.

🎯 Key Objectives

- Predict flight prices based on various factors
- Predict flight destinations based on other features
- Compare performance of different machine learning algorithms
- Analyze factors influencing flight prices and destinations

🛠️ Technologies Used

- Python
- Pandas for data manipulation
- Scikit-learn for machine learning algorithms
- XGBoost for gradient boosting
- Matplotlib and Seaborn for data visualization
- Missingno for visualizing missing data

📊 Dataset

The dataset includes information such as:

- Airline
- Origin and destination
- Travel date
- Departure and arrival times
- Flight duration
- Number of stops
- Price (target variable for regression)
- Destination (target variable for classification)

🔍 Methodology

1. Data Preprocessing
    - Removed unnecessary columns
    - Handled missing values
    - Performed one-hot encoding for categorical variables
2. Exploratory Data Analysis
3. Feature Engineering
4. Model Development

    - Implemented various models: Linear Regression, Decision Trees, Random Forest, XGBoost, SVM, Naive Bayes
5. Model Evaluation
    - Used metrics such as R-squared, MSE, Accuracy, ROC AUC
    - Implemented 10-fold cross-validation

📈 Results

Regression (Price Prediction):

- Best performing model: Decision Tree Regressor (R^2 = 0.7336 without normalization)
- XGBoost showed potential overfitting (Train accuracy: 0.9338, Test accuracy: 0.7302)

Classification (Destination Prediction):

- Decision Tree Classifier performed well (Test accuracy: 0.8811)
- SVM showed balanced performance between training and test sets

Classification (Price Category Prediction):

- Random Forest and XGBoost showed good performance
- Naive Bayes demonstrated balanced performance between training and test sets

🔑 Key Findings

1. Tree-based models (Decision Trees, Random Forest, XGBoost) showed good performance but potential overfitting
2. Linear models and SVM displayed more stable performance between training and test sets
3. 10-fold cross-validation provided more robust model performance estimates
4. Key factors influencing flight prices: total stops, duration, and specific airlines
5. For destination prediction, origin and airline are crucial features

🚀 Future Work

1. Feature selection and engineering to improve model performance
2. Hyperparameter tuning to address overfitting in tree-based models
3. Exploration of ensemble methods to enhance prediction accuracy
4. Investigation of external factors affecting flight prices and destinations
5. Development of a user-friendly interface for price and destination prediction