

Machine Learning Project

**Submitted To:**

**Dr. Faisal Al-Khateeb**

**Prepared By:**

**FATIMA ABDULLA HUBAIL 202108753**

**ALYA SHAWQI AMEEN 202107493**

**Table of Contents**

Part1: Description of the dataset

Part2: Used Models

2.1

2.2

2.3

Part3: Comparison

**Part1: Description of the dataset**

In this project we processed to use the clean dataset among the three datasets that were provided. The "Clean Dataset" contains 300,153 entries with 12 columns, each providing comprehensive details about airline travel data. The clean dataset shows exceptional quality and suitability for analysis, as its fully completed columns ensure data completeness, eliminating the risk of missing values and enabling robust analysis. Pre-processed and ready-to-use, the dataset appears to have undergone cleaning to remove duplicates, inconsistencies, and unnecessary information, saving valuable time and effort. It offers rich information, encompassing key features such as pricing trends, travel schedules, and flight details, making it ideal for projects related to travel or pricing analysis. Furthermore, its consistent formatting and thorough data organization enhance analytical ease, ensuring trustworthy findings and simplifying data modeling and visualization.

**Part2: Used Models**

Three different methods—a Random Forest, XGBoost, and an Artificial Neural Network (ANN)—were assessed using a dataset of airline expenses in this comparative study of machine learning models. Mean Squared Error (MSE) and R^2 score metrics were used to evaluate each model's performance once it was developed in Python.

* **2.1 XGBoost:** XGBoost is an excellent choice for analyzing “Clean Dataset” due to its efficiency, performance, and ability to handle both numerical and categorical data. It excels at processing tabular data, leveraging encoded features (e.g., one-hot, target, and hashing encoding) to capture complex relationships and interactions. Its gradient-boosted decision trees optimize predictive accuracy while avoiding overfitting through techniques like tree pruning and regularization. XGBoost’s feature importance insights are valuable for understanding key drivers in price prediction, and its scalability ensures it can handle larger datasets as needed. Additionally, it natively handles missing values and is well-suited for regression tasks, making it a powerful tool for extracting insights and delivering precise results from structured data.
* **2.2 Random Forest:** Because of its durability, ease of use, and outstanding performance on structured data, Random Forest is a great machine learning model for studying `Clean Dataset`. The dataset contains encoded variables like airline, `source\_city`, and `departure\_time`, is a good fit for it because it efficiently manages both category and numerical features. Even with noisy or unbalanced data, Random Forest minimizes overfitting and yields dependable results by constructing several decision trees and combining their predictions.
* **2.3 ANN:** Artificial Neural Networks (ANNs) are a powerful choice for analyzing your clean dataset due to their ability to capture non-linear relationships and handle complex, high-dimensional data, including encoded categorical and numerical features. They excel in regression tasks like price prediction by learning intricate feature interactions without assuming a specific data distribution. ANNs are robust to noise with techniques like dropout regularization and can be tailored to your dataset's needs through flexible architecture design. While they require more data and tuning than tree-based models, ANNs’ adaptability and precision make them well-suited for extracting insights and making accurate predictions from diverse datasets.
* **Part3: Comparative Analysis**
* **3.1 XGBoost:**
  + Best R2: With an R2 value of 0.2748, XGBoost exceeds the other models in describing 27.48% of the variance in the target variable.
  + Lowest Error (MAE and RMSE): XGBoost makes lower average and squared prediction errors, as seen by its lowest MAE (15,384.30) and RMSE (19,283.99).
  + K-Fold Verification: Consistent performance across folds is reinforced by its K-fold R2 (0.2779).
* **3.2 Random Forest:**
  + Poor R2 Score: Random Forest performs worse than predicting the mean, with a negative R2 of -0.5036. This indicates that Random Forest is unable to model the variance in the target variable.
  + Maximum Errors: The highest RMSE (27,768.22) and MAE (17,721.30) point to significant prediction errors.
* **3.3 ANN:**
  + Intermediate MAE and RMSE: ANN outperforms Random Forest but falls short of XGBoost by a margin with an MAE of 16,203.61 and an RMSE of 19,937.10.
  + Negative R2 Score: The R2 of -0.5036 indicates that ANN is likewise unable to adequately capture the variance in the data.
  + Consistent K-Fold MAE: K-Fold validation demonstrates a stable performance with a low standard deviation (±247.76) and an average MAE of 16,506.62.

In conclusion, **XGBoost** emerges as the best-performing model, delivering the lowest errors (MAE: 15,384.30, RMSE: 19,283.99) and the highest R² score (0.2748), indicating its ability to effectively capture the variance in the target variable. **ANN** shows moderate performance with errors slightly higher than XGBoost (MAE: 16,203.61, RMSE: 19,937.10) and a negative R² score (-0.5036), suggesting room for improvement in generalizing the data. **Random Forest**, however, performs the worst, with the highest errors (MAE: 17,721.30, RMSE: 27,768.22) and a negative R² score (-0.5036), indicating significant inefficiencies in modeling the dataset.