Restaurant Rating Prediction Project Report

# 1. Introduction

The objective of this project is to build a machine learning model that can predict the ratings of restaurants based on various features such as location, cuisine, pricing, and service options. Predicting accurate restaurant ratings helps customers make informed decisions and enables businesses to understand factors influencing their ratings.

# 2. Dataset Overview

The dataset contains detailed information about restaurants, including:

- Restaurant details: Name, ID, Address, Locality  
- Geographical info: City, Country, Longitude, Latitude  
- Restaurant features: Cuisine types, Average cost for two, Availability of table booking, online delivery options, delivery status, price range  
- User interactions: Number of votes  
- Target variable: Aggregate rating (the rating score to be predicted)

# 3. Data Preprocessing

Before training the model, the dataset underwent the following preprocessing steps:

- Removal of irrelevant columns: Non-essential columns such as restaurant ID, name, and address were removed to reduce noise.  
- Handling missing values: Missing data was imputed using the mode for categorical columns and median for numerical columns to maintain data consistency.  
- Encoding categorical variables: Categorical features (like City, Cuisine, Currency) were converted into numeric form using Label Encoding so that machine learning models can process them.  
- Feature and target separation: Features were separated from the target rating variable for model training.

# 4. Model Selection and Comparison

## 4.1 Purpose of Model Selection

Selecting the appropriate machine learning model is critical to ensure high prediction accuracy and robustness when applied to new data. Different algorithms vary in complexity, assumptions, and suitability for data characteristics.

## 4.2 Models Evaluated

The following regression models were evaluated:  
- Decision Tree Regressor: A simple tree-based model that splits data based on feature thresholds.  
- Random Forest Regressor: An ensemble of decision trees that averages results to improve accuracy and reduce overfitting.  
- Linear Regression: A basic model assuming linear relationships between features and rating.  
- XGBoost Regressor: An advanced gradient boosting method known for high accuracy but requiring careful tuning.

## 4.3 Performance Metrics Used

- Mean Squared Error (MSE): Measures average squared difference between actual and predicted ratings. Lower values indicate better fit.  
- R-squared (R²) Score: Indicates proportion of variance explained by the model. Values closer to 1 imply better predictions.

## 4.4 Results Summary

|  |  |  |
| --- | --- | --- |
| Model | Mean Squared Error | R² Score |
| Decision Tree | 0.174 | 0.924 |
| Random Forest | 0.088 | 0.961 |
| Linear Regression | 1.586 | 0.303 |
| XGBoost | 0.089 | 0.961 |

## 4.5 Analysis

- Random Forest and XGBoost performed the best with the lowest error and highest R² scores.  
- Decision Tree performed moderately well but was outperformed by ensemble methods.  
- Linear Regression was unsuitable due to its assumption of linearity, resulting in poor predictions.

# 5. Final Model Choice: Random Forest Regressor

The Random Forest Regressor was selected as the final model because:  
- It balances accuracy and simplicity effectively.  
- It is robust against overfitting by averaging multiple decision trees.  
- It provides useful insights into feature importance, helping understand which factors influence ratings most.  
- It requires less hyperparameter tuning and computational resources compared to XGBoost, making it easier to deploy and maintain.

# 6. Model Training and Evaluation

The final Random Forest model was trained on the entire cleaned dataset to leverage all available data. Evaluation on a test split demonstrated strong performance with an R² score above 0.96, indicating the model explains over 96% of rating variance.

# 7. Conclusion and Future Work

This project successfully developed a reliable model to predict restaurant ratings based on multi-dimensional data. The use of Random Forest enabled accurate predictions while maintaining interpretability and efficiency.

Project URL: <https://restaurant.streamlit.app/>