



Documentation On
“Restaurant Rating Prediction”

Data Science 2024-25

Submitted By:

Aditya Pandey
Karishma Gaikwad
Shubham Pawar

Dr. Shantanu Pathak
Project Guide

Mr. Keshav Kumar
Course Director

DECLARATION

I, the undersigned, hereby declare that the project report titled "**Restaurant Rating Prediction**" has been written and submitted by me to **DYPIMS** in fulfilment of the requirements for the award of the **Data Science Course (2024-25)** under the guidance of **Dr. Shantanu Pathak**. This is my original work, and I have not copied any code or content from any source without proper attribution. Additionally, I have not allowed anyone else to copy my work.

The project was completed using **Python and machine learning libraries** as part of my academic coursework. I also confirm that this project is original and has not been submitted previously for any other academic or professional purpose.

Place:

Signature:

Date:

Name:

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Dr. Shantanu Pathak**, my project guide, for his invaluable guidance and support throughout this academic project. His expertise, insights, and encouragement have been instrumental in the successful completion of this work.

I am also thankful to my fellow classmates for their cooperation and support during the project. Their feedback and suggestions greatly contributed to improving the quality of this work.

Additionally, I extend my heartfelt gratitude to **Mr. Keshav Kumar**, Course Director, for providing me with the necessary resources and facilities to complete this project. His support has been crucial in ensuring its timely completion.

Finally, I would like to express my deepest appreciation to my family and friends for their constant encouragement and unwavering support. Their belief in me has been a continuous source of motivation and inspiration. Thank you all for your support and guidance in completing this academic project.

ABSTRACT

Restaurant ratings have become one of the most used parameters for evaluating a restaurant. Extensive research has been conducted on various restaurants and the quality of food they serve. The rating of a restaurant depends on multiple factors such as customer reviews, location, average cost per person, cuisines, and the type of restaurant. The primary goal of this project is to gain insights into the factors influencing restaurant ratings and predict the ratings based on these attributes. We have explored different predictive models, including Random Forest Regressor, Linear Regression, and Decision Tree Regressor. Among these, the Random Forest Regressor achieved the highest accuracy, with an R^2 score of 78%.

This study provides valuable insights for restaurant owners, investors, and food aggregators to understand customer preferences and improve business strategies.

Table Of Contents

1. INTRODUCTION	7
1.1 Problem Statement.....	9
1.2 Scope.....	10
1.3 Aim & Objectives	11
2.PROJECT DESCRIPTION	12
2.1 Project workflow	13
2.2 Data Collection	14
2.3 Studying the Data.....	14
2.4 Studying the Model	18
2.5 Implementing the Model.....	19
2.6 Validating the Model	21
3.MODEL DESCRIPTION	22
3.1 ML model	23
4.DATA FLOW	24
4.1 Data flow of project.....	25
5. <u>PROJECT REQUIREMENTS</u>	27
6. <u>FUTURE SCOPE</u>	30
7. <u>CONCLUSION</u>	33
8. <u>REFERENCES</u>	36

Table of Figures:

<i>Figure 1 Project workflow diagram</i>	<i>13</i>
<i>Figure 2 Features Descriptions.....</i>	<i>15</i>
<i>Figure 3 Booking</i>	<i>16</i>
<i>Figure 4 Online Delivery</i>	<i>16</i>
<i>Figure 5 Cost per person vs Rating</i>	<i>17</i>
<i>Figure 6 Correlation between cost and ratings</i>	<i>17</i>
<i>Figure 7 Code for model implementation</i>	<i>20</i>
<i>Figure 8 Validation</i>	<i>21</i>
<i>Figure 9 Input Form</i>	<i>34</i>
<i>Figure 10 Output</i>	<i>35</i>

CHAPTER 1

INTRODUCTION

INTRODUCTION

With the rise of numerous restaurants, dine-out culture, and food delivery apps, the restaurant industry has become one of the most competitive sectors. With a wide range of choices available, customers are more inclined to explore various cuisines. However, in such a competitive market, restaurants must go beyond offering quality food—they also need to introduce innovations in their menus and services.

Consumer dining behaviour in India has undergone a significant transformation over the past decade due to factors like urbanization, socioeconomic growth, demographic shifts, and increased exposure to international lifestyles. As a result, understanding customer preferences and predicting restaurant performance has become crucial for restaurant owners, investors, and food aggregators.

Machine learning is playing an increasingly vital role in the restaurant industry by analysing data patterns and customer behaviour to enhance service quality and business strategies. By leveraging past data, machine learning can provide insights into factors that influence restaurant success and help forecast key business trends.

1.1 Problem Statement

Machine learning techniques are highly effective in predictive analytics due to their ability to process large datasets with complex characteristics and noise. Predictive analytics encompasses various statistical techniques to estimate or ‘predict’ future outcomes based on historical business data.

In this project, we aim to develop a **restaurant rating prediction system** using **machine learning algorithms**. The system will analyse factors such as cost per person, cuisine type, and restaurant location to predict restaurant ratings.

We employ regression techniques, specifically **Linear Regression** and **Random Forest Regressor**, to build predictive models. A comparative analysis of these models will be conducted to determine the most effective approach for restaurant rating prediction.

This project covers the **design, development, and deployment** of a machine learning-based system that provides valuable insights for restaurant owners, investors, and food aggregators to make data-driven decisions.

1.2 Scope

The scope of this project is to develop a machine learning-based restaurant rating prediction system that helps restaurant owners, investors, and food aggregators make informed decisions. The system will analyse key factors such as cost per person, cuisine type, and city to predict restaurant ratings, providing valuable insights into the dining industry.

This project involves data preprocessing, feature selection, model training, evaluation, and visualization to understand how different factors influence restaurant ratings. Machine learning techniques such as Linear Regression and Random Forest Regressor are applied to build predictive models, with a comparative analysis conducted to determine the most effective algorithm.

The system can be extended in the future by incorporating additional factors such as customer reviews, service quality, and ambiance ratings to improve prediction accuracy. Additionally, it can be integrated with food aggregator platforms for real-time restaurant performance analysis.

By leveraging machine learning, this project aims to enhance business strategies, optimize customer experience, and help stakeholders gain deeper insights into restaurant success factors.

1.3 Aim & Objectives

The primary aim of this project is to develop a Restaurant Rating Prediction System that helps restaurant owners, food aggregators, and investors estimate the expected rating of a restaurant based on key factors such as city, cuisine type, and cost per person. To achieve the project aim, the following objectives are outlined:

1. **Data Acquisition & Preparation:** Gather restaurants data, preprocess missing values, and perform exploratory data analysis.
2. **Feature Engineering:** Extract and select the most influential features impacting restaurant rating predictions.
3. **Model Development & Evaluation:** Train multiple classification models and select the best-performing one (Random Forest regressor).
4. **Performance Optimization:** Fine-tune the model for mean squared error (MSE), mean absolute error (MAE), and R2-score.
5. **Label Encoding and One Hot Encoding:** Used encoding to convert categorical values in the form of numeric
6. **Regulatory & Ethical Considerations:** Ensure fairness in model predictions, avoiding biases against demographic groups.
7. **Deployment & Implementation:** Integrate the trained model into a real-world restaurant rating system for automated decision-making.

CHAPTER 2

PROJECT DESCRIPTION

2.1 Project Work Flow

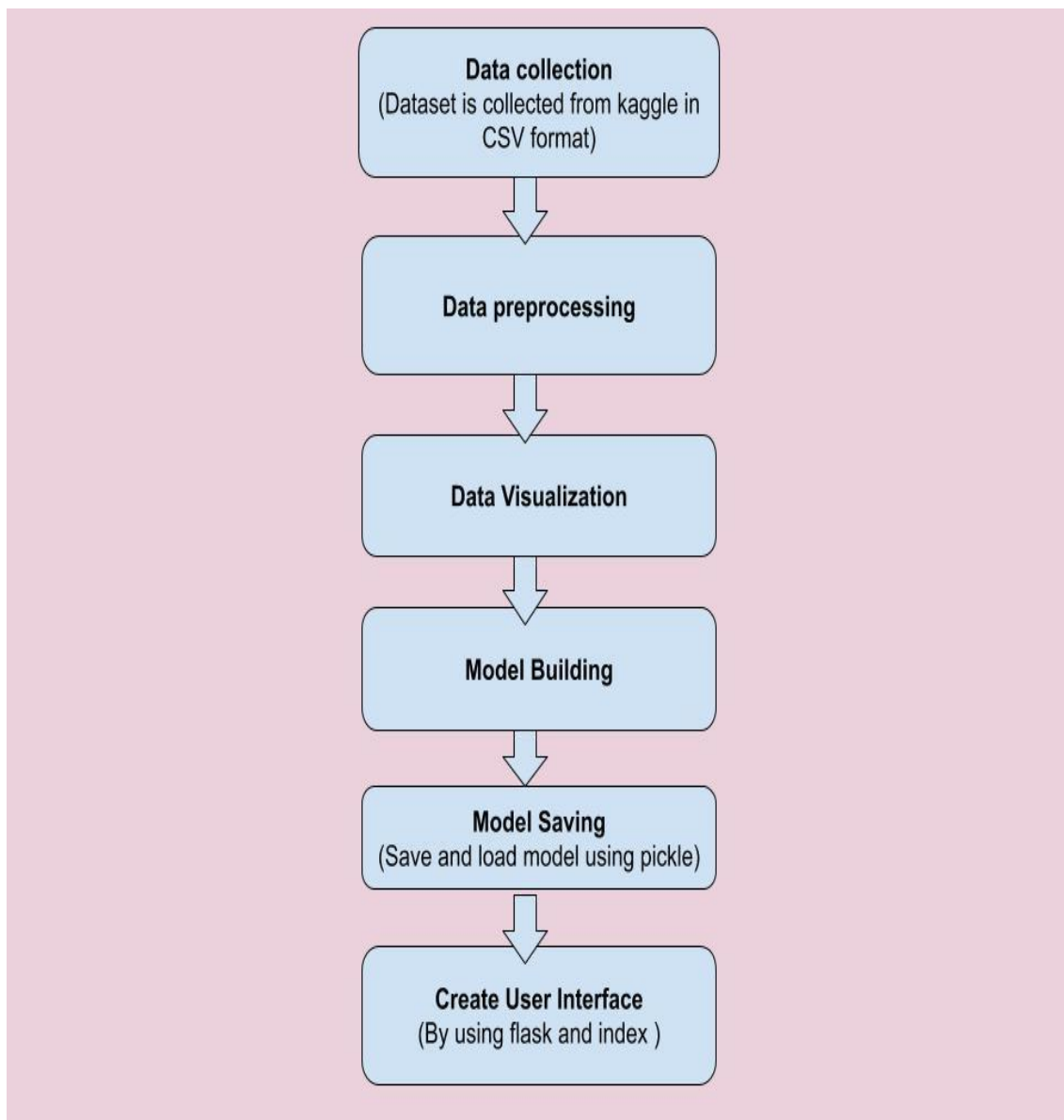


Fig 1 – Project workflow diagram

2.2 Data Collection

The world is a hub of restaurants, and various restaurant data is collected on the Kaggle website. The basic idea behind analysing the dataset is to gain insights into the factors affecting the aggregate rating of each restaurant and the establishment of different types of restaurants in various locations. The dataset consists of 24,151 rows and 84 columns. The table below provides a description of the dataset attributes.

2.3 Studying the Data

- **76 Numeric Columns:** Includes Rating, No_of_Best_Sellers, No_of_Varieties, Cost_Per_Person, Afghani, African, American, Andhra, Arabian, Asian, Assamese, Awadhi, BBQ, Bakery, Belgian, Bengali, Beverages, Bihari, Bohri, British, Burmese, Cantonese, Chettinad, Chinese, Continental, Desserts, European, Fast Food, French, German, Goan, Greek, Gujarati, Healthy Food, Hyderabad, Indonesian, Iranian, Italian, Japanese, Jewish, Kashmiri, Kerala, Konkan, Korean, Lebanese, Lucknowi, Maharashtrian, Malaysian, Mangalorean, Mediterranean, Mexican, Middle Eastern, Modern Indian, Mughlai, Naga, Nepalese, North Eastern, North Indian, Oriya, Parsi, Portuguese,

Rajasthani, Russian, Seafood, Sindhi, Singaporean, South American, South Indian, Spanish, Sri Lankan, Tamil, Thai, Tibetan, Turkish, Vegan, Vietnamese.

- **8 Categorical Columns:** Includes Name, Menu, Delivery, Booking, Type, City, Category, Price Category

	Name	Menu	Delivery	Booking	Type	City	No_of_Best_Sellers	No_of_Varieties	Cost_Per_Person	Rating	...	South America
0	Jalsa	No	Yes	Yes	Buffet	Banashankari	7	3	400.0	4.1	...	
1	Spice Elephant	No	Yes	No	Buffet	Banashankari	7	3	400.0	4.1	...	
2	San Churro Cafe	No	Yes	No	Buffet	Banashankari	7	3	400.0	3.8	...	
3	Addhuri Udupi Bhojana	No	No	No	Buffet	Banashankari	1	2	150.0	3.7	...	
4	Grand Village	No	No	No	Buffet	Banashankari	2	2	300.0	3.8	...	
5	Timepass Dinner	No	Yes	No	Buffet	Banashankari	7	1	300.0	3.8	...	
6	Onesta	No	Yes	Yes	Cafes	Banashankari	7	3	300.0	4.6	...	
7	Penthouse Cafe	No	Yes	No	Cafes	Banashankari	7	3	350.0	4.0	...	
8	Smaczego	No	Yes	No	Cafes	Banashankari	7	5	275.0	4.2	...	
9	Café Down The A...	No	Yes	No	Cafes	Banashankari	7	1	250.0	4.1	...	

10 rows × 84 columns

Fig 2: Features Descriptions

Data preprocessing:

Data preprocessing is a process in which raw data is transformed into an understandable and convenient format. Data preprocessing is useful in resolving issues such as null values, categorical values like online order, booking table etc are converted into numerical values, inconsistencies in data are removed. Unnecessary attributes for analysis in the dataset are removed.

Data Visualization:

Data Visualization is the visual representation of data in the form of figures, graphs, charts and plots such as histograms, box plots, and bar charts were used to identify trends and anomalies. It is used to analyse massive amounts of data to understand the trends and patterns from data. Visual representation of restaurant data is performed below.

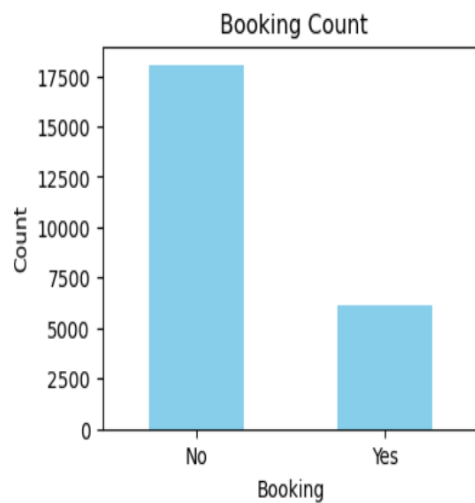


Fig3: Booking



Fig4: Online Delivery

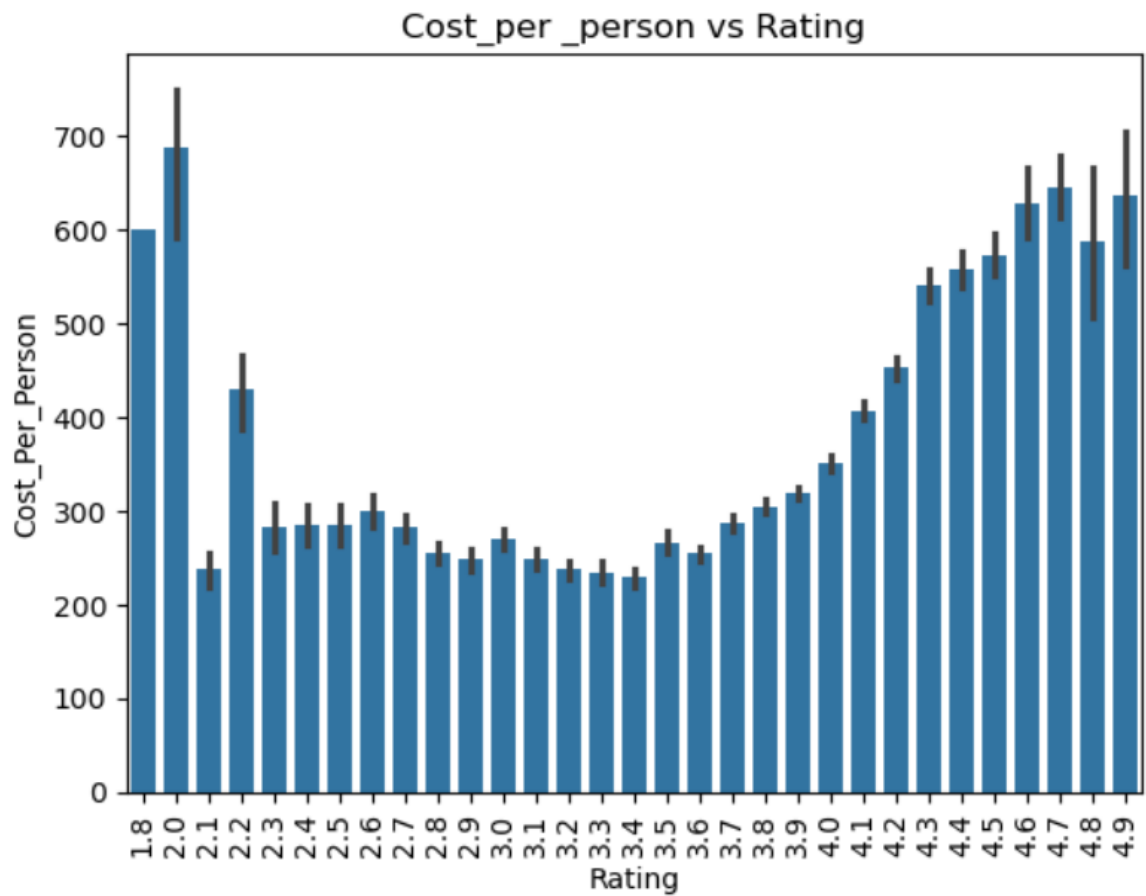


Fig5: Cost per person vs ratings



Fig 6: correlation between cost and ratings

Above figures we can infer that there are more customers ordering online rather than dining out. Only a small number of customers book tables and most customers prefer walk-in. On an average there are the number of delivery restaurants and least number of bars. Restaurants serving food in the range of Rs.450 are exceptionally rated. Rating of restaurants increases as the cost increases.

2.4 Studying the Model

To ensure accurate rating default predictions, various machine learning models were analysed, focusing on performance, interpretability, and computational efficiency.

Model Selection & Justification

- **Random Forest:** Chosen for its ability to handle both numerical and categorical data, robustness against overfitting, and high accuracy in classification tasks.

Model Evaluation Metrics

- **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
- **Mean Absolute Error (MAE):** Measures the average absolute difference between actual and predicted values.

- **R² Score:** Indicates how well the model explains the variance in the target variable.
- **One-Hot Encoding** is a method used to convert categorical data into a numerical format by creating binary columns for each unique category. Each category is represented by a vector of 0s and 1s.

2.5 Implementing the Model

The implementation phase involves training a machine learning model to predict restaurant ratings based on key features such as city, cost per person, and cuisines. Various regression algorithms, including Linear Regression, Decision Tree, and Random Forest Regressor, are applied. The models are trained and evaluated using appropriate metrics, and Random Forest Regressor is chosen as the best-performing model with an R² score of 78%. The trained model is then deployed for prediction.

```
# Split data: 80% training, 20% testing
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Train the model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, Y_train)

# Make predictions
y_pred = model.predict(X_test)

# Calculate and print metrics
mae = mean_absolute_error(Y_test, y_pred)
mse = mean_squared_error(Y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(Y_test, y_pred)
```

Fig 7: Code for Model Implementation

A machine learning model was implemented to predict ratings is **Random Forest Regressor**.

The model was trained on the dataset, and performance was assessed by calculating MAE, MSE and R2 score.

2.6 Validating the Model

Model Validation

The selected model, Random Forest regressor (`n_estimators=100, random_state=42`), was validated using multiple performance metrics to ensure reliability and accuracy.

Validation Techniques Used:

- Train-Test Split: The dataset was divided into training and testing sets to evaluate generalization.
- R² Score: The model's overall correctness in predictions was measured.

These validation steps confirmed that the Random Forest model effectively predicts loan defaults with high accuracy and robustness.

```
print(f"Mean Absolute Error (MAE): {mae*100}%")
print(f"Mean Squared Error (MSE): {mse*100}%")
print(f"Root Mean Squared Error (RMSE): {rmse*100}%")
print(f"R2 Score: {r2*100}%")
```

```
Mean Absolute Error (MAE): 10.263597349314622%
Mean Squared Error (MSE): 4.026546921013575%
Root Mean Squared Error (RMSE): 20.066257550957467%
R2 Score: 78.65243841909003%
```

Fig 8: Validation

CHAPTER 3

MODEL DESCRIPTION

3.1 ML Model

For this project, we implemented a Random Forest Regressor to predict restaurant ratings based on various features. Random Forest was chosen due to its ability to handle both numerical and categorical data efficiently while reducing the risk of overfitting by aggregating multiple decision trees.

The model was trained using 100 estimators (trees) with a randomstate of 42, ensuring a balance between accuracy and generalization. It leverages multiple weak learners to create a robust predictive model that effectively captures complex relationships in the dataset.

The Random Forest regression algorithm is one of the most effective machine learning models used for predictive analytics. The random forest algorithm, which is an additive model, which combines the decisions, forms a sequence of base models.

The model is formulated as follows

$$g(x)=f_0(x)+f_1(x)+f_2(x)+...$$

Here g is the sum of all simple base models f_i , which are constructed independently using a different subsample of data. Multiple models involved for better prediction is model assembling.

CHAPTER 4

DATA FLOW

4.1 Data Flow of The Project

Data Flow in Restaurant Rating Prediction Project

The data flow in our project follows a structured pipeline, ensuring seamless processing from raw data collection to final decision-making.

The key steps involved are:

1. Data Collection & Preprocessing

- The dataset, sourced from Kaggle, contains **24151 rows and 84 columns**.
- Features were categorized into 76 numerical and 8 categorical columns.
- Missing values and outliers were handled to improve model performance.

2. Feature Engineering & Selection

- Categorical variables were encoded using one-hot encoding and label encoding.
- Numerical features were standardized to improve model training.
- Feature selection was performed to retain only relevant attributes.

3. Model Training & Evaluation

- Model Random Forest was tested.
- Random Forest (n_estimators=100, random_state=42) was

selected as the final model due to its high accuracy and robustness.

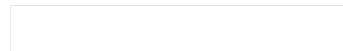
- The model was trained using a 70-30 train-test split and evaluated using mae, mse, rmse and R2-score.

4. Prediction & Risk Assessment

- The trained model predicts ratings of restaurants based on features.

CHAPTER 5

PROJECT REQUIREMENTS



Project Requirements

The Restaurant rating prediction project was developed using various tools and libraries that played a crucial role in building, training, and deploying the model. Libraries Used:

1. **Python:** Python was the primary programming language used for developing the entire project. Python enabled us to implement various machine learning algorithms, data preprocessing, and model deployment effectively.
2. **scikit-learn:** This Python library was used for implementing machine learning algorithms. It provides a comprehensive collection of tools for data mining and data analysis, including various classification algorithms such as Random Forest, Logistic Regression. We also used it for data splitting, evaluation metrics, and feature engineering.
3. **Pickle:** Pickle was used to serialize and deserialize machine learning models, enabling easy storage and loading of model objects. This streamlined the deployment process by allowing trained models to be saved as `.pkl` files and reused without retraining. Additionally, Pickle ensured compatibility across environments, making it ideal for integrating models into production systems for real-time predictions.
4. **Flask:** Flask, a lightweight Python web framework, was used to deploy the machine learning model as a web application. Flask enabled us to create a user-friendly interface where users can input

customer data and receive predictions about restaurant ratings default, along with the reasoning behind the prediction.

CHAPTER 6

FUTURE SCOPE

FUTURE SCOPE

1. **Automated Feature Extraction:** Deep learning algorithms can replace manual feature extraction, learning complex patterns from raw data to improve prediction accuracy.
2. **Handling Large Datasets:** With exponentially growing datasets, deep learning models like CNNs or RNNs can efficiently process vast amounts of data.
3. **Real-Time Predictions:** Deploying the model using APIs or cloud platforms can enable real-time restaurant rating predictions, benefiting users and businesses.
4. **Multi-Modal Data Analysis:** Incorporating images of food, restaurant interiors, and customer reviews using image recognition and NLP can enhance prediction quality.
5. **Sentiment Analysis Integration:** Advanced sentiment analysis using transformers (e.g., BERT) can provide deeper insights into customer reviews for accurate ratings.
6. **Personalized Recommendations:** The model can be extended to suggest personalized restaurant recommendations based on user preferences and past experiences.
7. **Explainable AI (XAI):** Implementing explainable AI techniques can provide transparency into the model's decision-making, building trust with users.
8. **Continuous Model Improvement:** Integrating feedback loops can help the

model adapt to changing customer preferences and market trends over time.

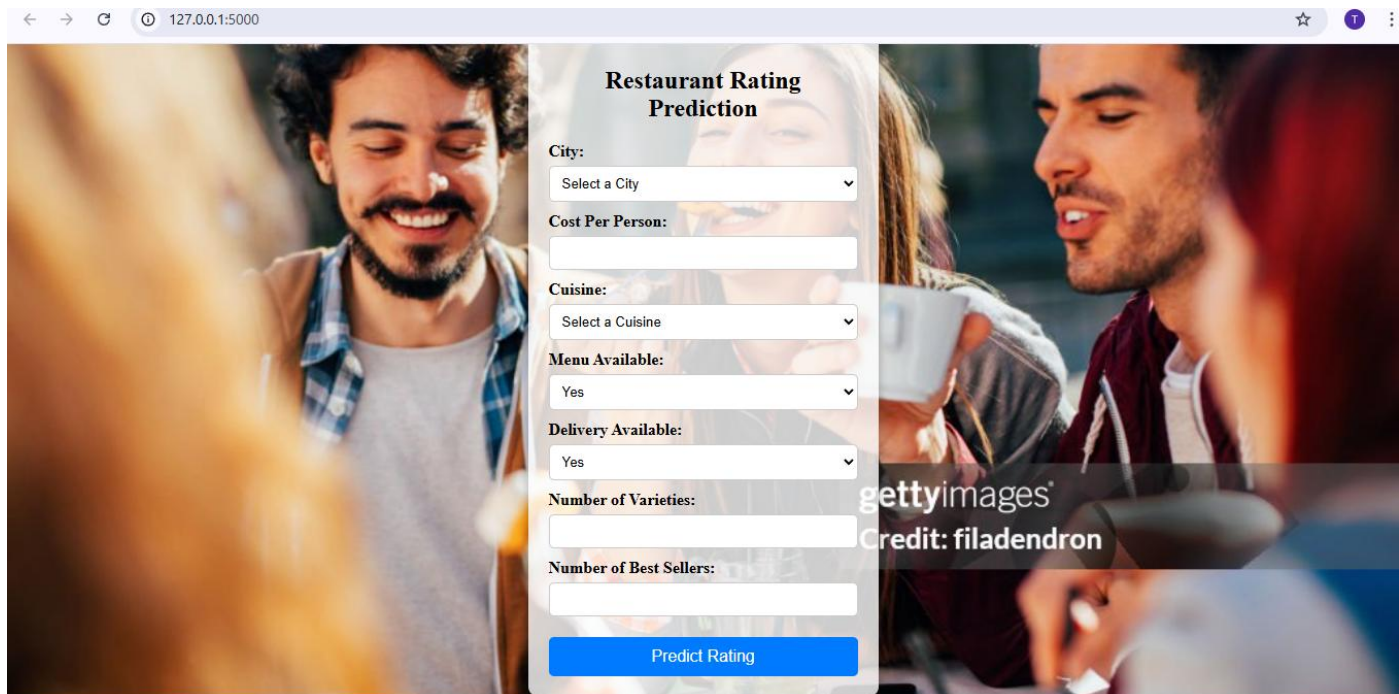
9. **Cross-Platform Integration:** Seamless integration with food delivery apps, review platforms, and restaurant management systems can provide actionable insights.

CHAPTER 7

CONCLUSION

CONCLUSION

ML techniques are gradually used in the Restaurant industry to predict the customer pattern and improve business. This paper has presented a Regression algorithm. The behaviours and methodology of them was discussed and models were created. Comparison of performance metrics of those models is investigated. Based on the results of regression techniques it is inferred that Random Forest Regression has the highest performance in terms of regression score and lowest error rate.

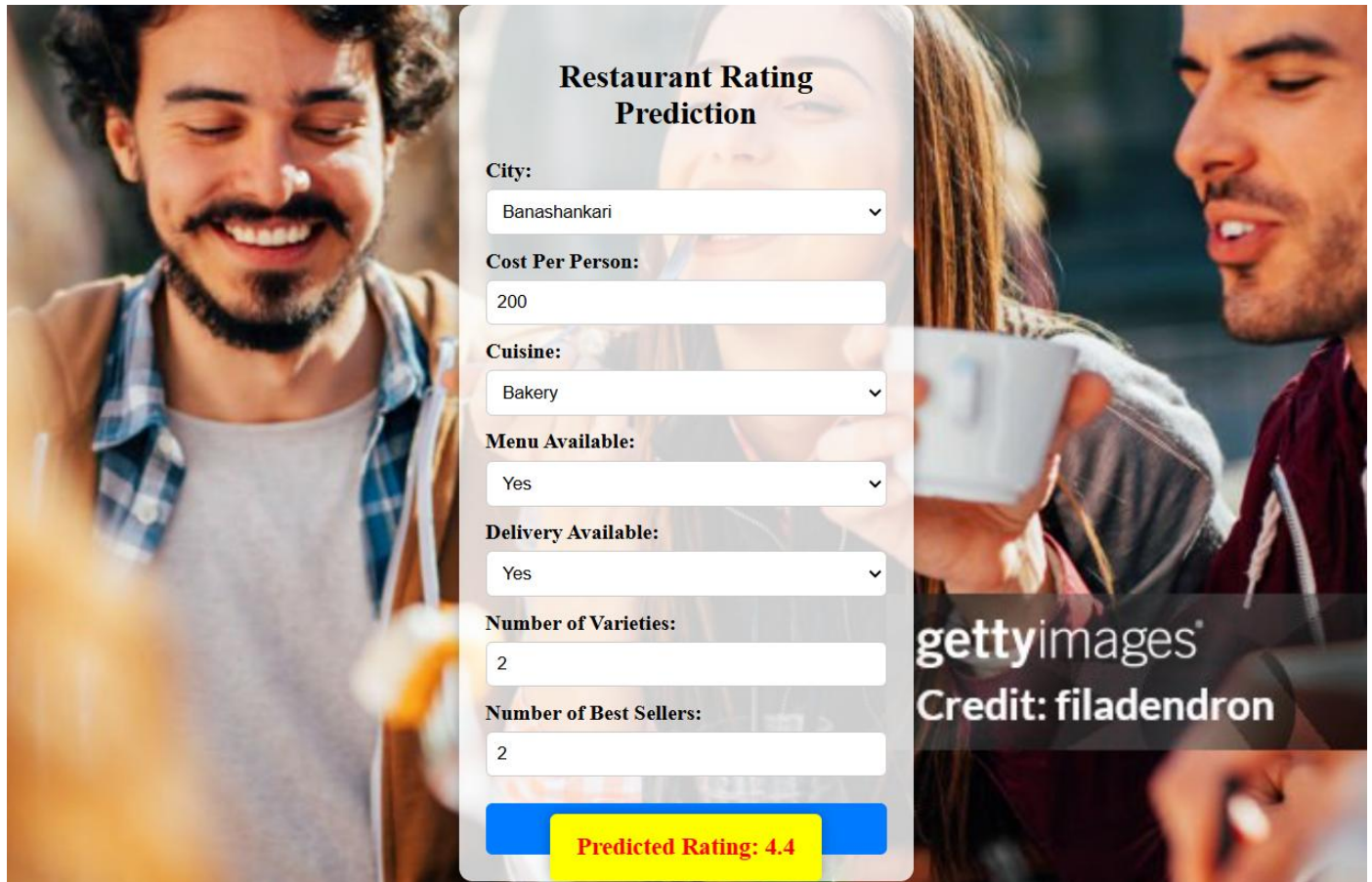


The screenshot displays a web browser window with the address bar showing '127.0.0.1:5000'. The main content is a web form titled 'Restaurant Rating Prediction'. The form contains the following fields and options:

- City:** A dropdown menu with the text 'Select a City'.
- Cost Per Person:** A text input field.
- Cuisine:** A dropdown menu with the text 'Select a Cuisine'.
- Menu Available:** A dropdown menu with the text 'Yes'.
- Delivery Available:** A dropdown menu with the text 'Yes'.
- Number of Varieties:** A text input field.
- Number of Best Sellers:** A text input field.
- Predict Rating:** A blue button.

The background of the form is a blurred image of people in a restaurant setting. A watermark 'gettyimages' and 'Credit: filadendron' are visible on the right side of the image.

Fig 9: Input Form



The image shows a web form titled "Restaurant Rating Prediction" overlaid on a background photograph of a smiling man and a woman looking at a smartphone. The form contains several input fields and dropdown menus. At the bottom, a yellow button displays the "Predicted Rating: 4.4".

Restaurant Rating Prediction

City:
Banashankari

Cost Per Person:
200

Cuisine:
Bakery

Menu Available:
Yes

Delivery Available:
Yes

Number of Varieties:
2

Number of Best Sellers:
2

Predicted Rating: 4.4

gettyimages[®]
Credit: filadendron

Fig 10: Output

REFERENCES

Here are some references you can add to the report:

1. Books/Research Papers:

- J. Brownlee, *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End*, Machine Learning Mastery, 2016.
- Mohan S Acharya, Asfia Armaan, Aneeta S Antony “A Comparison of Regression Models for Prediction of Graduate Admissions” Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019)

2. Online Documentation and Articles:

- "Random Forest Classifier - Scikit-learn Documentation."
<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- "Flask Documentation." <https://flask.palletsprojects.com/>
- Langchain Documentation. <https://langchain.readthedocs.io/>

3. Websites and Blogs:

- "Introduction to Random Forest in Machine Learning."
<https://towardsdatascience.com/random-forest-in-machine-learning-c>"AWS EC2 Instance Overview."

<https://aws.amazon.com/ec2/>

4. Software/Tools:

- Scikit-learn documentation for model building and evaluation. <https://scikit-learn.org/>

5. Additional References:

- "Groq API Documentation" <https://groq.com/docs/>