**Zomato Restaurant Classification Using Machine Learning and Deep Learning**

Restaurants have long been an integral part of the social fabric, providing people with a place to gather, eat, and enjoy. But in today’s digital era, customers rely heavily on online platforms to discover new dining options and rate their experiences. This shift toward digital engagement creates a need for classification models that can categorize restaurants based on customer preferences, locations, prices, and more. Machine learning and deep learning offer sophisticated tools to solve such classification problems, providing businesses with actionable insights.

In this project, we used a Zomato dataset containing restaurant details to build predictive models that classify restaurants based on **Price Range**. Our goal was to explore various machine learning and deep learning models to maximize prediction accuracy and offer insights into which factors most influence restaurant classification.



**1. Problem Definition**

The primary objective of this project is to classify restaurants based on their price range, utilizing various features from the dataset. The price range is categorized into four distinct levels: premium, mid-range, budget, and value. Understanding the price classification of restaurants is crucial for both customers and businesses. Customers can make informed dining choices, while restaurant owners can tailor their services and marketing strategies to target specific demographics effectively.

In today's competitive food industry, the ability to accurately classify restaurants based on their pricing can provide significant advantages. For instance, a food delivery platform can recommend budget-friendly options to cost-conscious consumers while highlighting premium dining experiences for those seeking upscale meals.

From a technical standpoint, this problem is framed as a supervised learning classification task. The dataset comprises various features, including location, average cost for two, availability of online delivery, aggregate ratings, and more. The challenge lies in accurately predicting a restaurant's price category based on these features.

To address this problem, we implemented multiple modelling approaches, starting with traditional machine learning algorithms before advancing to deep learning techniques. This progression allowed us to explore the strengths and weaknesses of each model in capturing the nuances of restaurant pricing.

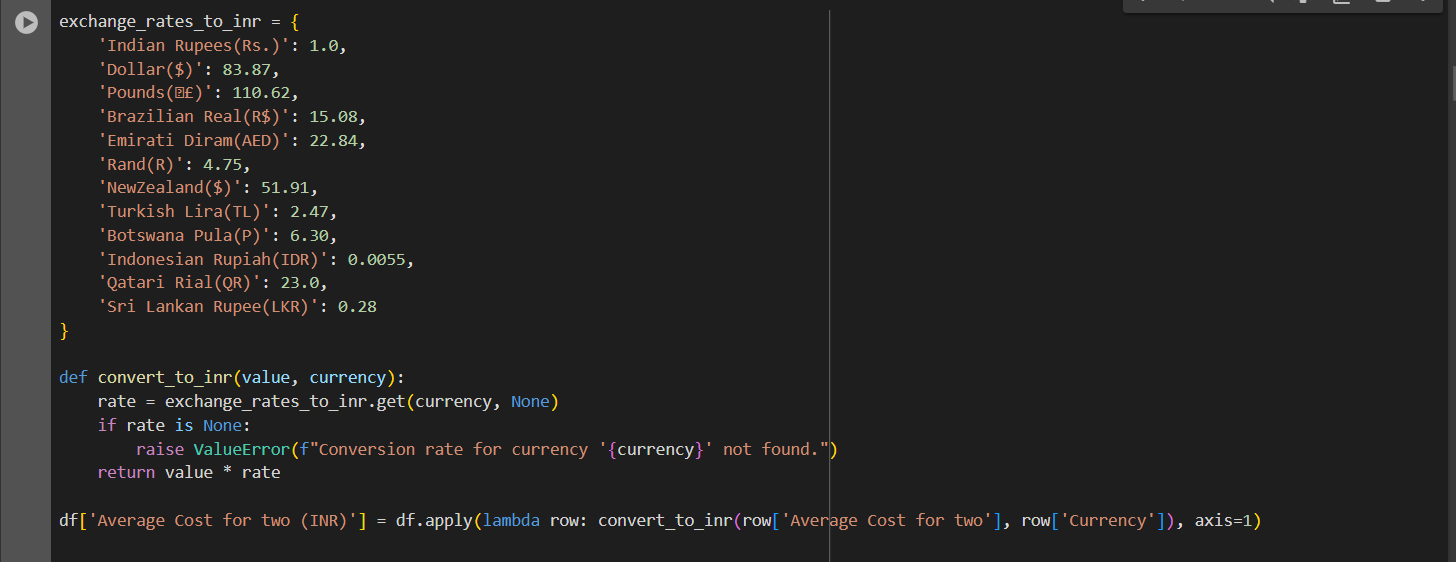
**2. Data Analysis**

The dataset used in this project contains 9,551 entries with 14 columns, each representing various features that could influence the classification of restaurants. A thorough analysis of these features is essential for understanding their significance in predicting the price range. The following table summarizes the relevant features:

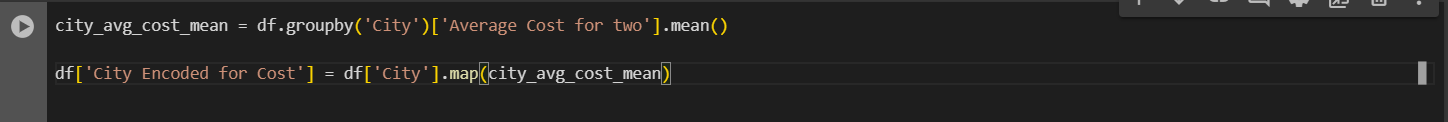
| **Column Name** | **Description** |
| --- | --- |
| **Longitude** | Geographic coordinate indicating the east-west position of the restaurant. |
| **Latitude** | Geographic coordinate indicating the north-south position of the restaurant. |
| **Cuisines** | A list of cuisines offered by the restaurant (some missing values). |
| **Average Cost for Two** | The average cost of a meal for two individuals, a crucial feature for pricing. |
| **Currency** | The currency in which the prices are listed, ensuring consistency in analysis. |
| **Has Table Booking** | Indicates whether the restaurant offers table booking (yes/no). |
| **Has Online Delivery** | Indicates if the restaurant provides online delivery services (yes/no). |
| **Is Delivering Now** | Shows whether the restaurant is currently available for delivery (yes/no). |
| **Switch to Order Menu** | Indicates if the restaurant has a switch to an order menu feature (yes/no). |
| **Price Range** | The categorical variable indicating the price classification (target variable). |
| **Aggregate Rating** | The overall rating of the restaurant, which can influence customer choices. |
| **Rating Colour** | A color-coded representation of the rating (used for visual appeal). |
| **Rating Text** | Descriptive text associated with the aggregate rating (e.g., "Excellent"). |
| **Votes** | The number of votes or reviews received by the restaurant. |

**Key Features for Classification**

1. **Average Cost for Two:** This feature is particularly significant as it directly impacts the price classification of restaurants. Higher average costs typically correlate with premium dining experiences. In our analysis, we converted the Average Cost for Two values into Indian Rupees (INR) to ensure uniformity across different currencies. By standardizing these values, we mitigate the influence of outliers and enhance the model’s predictive capability, ensuring that it can accurately classify restaurants across various price ranges**.**



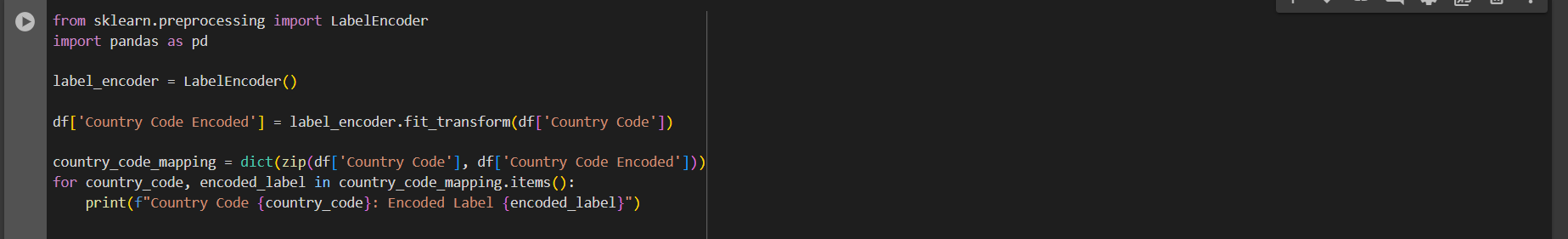
1. **Aggregate Rating**: The overall customer rating plays a critical role in influencing dining choices. Restaurants with higher ratings are often associated with better quality and service, impacting their price range.
2. **Has Online Delivery**: With the increasing popularity of food delivery services, this feature is vital. Restaurants that offer online delivery may cater to a different customer demographic, potentially affecting their pricing strategy.
3. **Has Table Booking**: This feature can indicate whether a restaurant is positioned as a more upscale option. Restaurants that provide table booking may appeal to customers seeking a more refined dining experience.
4. **Cuisines**: The types of cuisines offered can also impact pricing. Certain cuisines (e.g., fine dining or exotic options) may command higher prices compared to others (e.g., fast food).
5. **Longitude and Latitude**: Geographic location can significantly affect restaurant pricing and customer preferences. For instance, restaurants in high-demand urban areas may have higher prices due to increased rent and competition.
6. **Currency**: The currency in which prices are listed is important for ensuring consistency in pricing analysis, especially when comparing restaurants in different regions. A clear understanding of currency differences is crucial for accurate classification.
7. **Votes**: The number of votes or reviews can serve as a proxy for a restaurant's popularity and perceived quality. Higher votes often correlate with better visibility and customer trust, potentially influencing the restaurant's price range.
8. **City Encoding**: To capture the relationship between the city and the target variable, we performed target encoding for the 'City' feature. This approach involves calculating the mean average cost for two in each city and encoding it in the dataset as follows:



Target encoding is chosen here due to the high cardinality of city values. By mapping the average cost for two to each city, we aim to capture the logical relationship between city location and food prices. This encoding leverages the continuous nature of the 'Average Cost for Two' column, providing valuable insights while maintaining the categorical nature of the target column, 'Price Range'.

Understanding the significance of these features allows us to make informed decisions during the preprocessing stage and model building, ultimately leading to improved prediction accuracy.

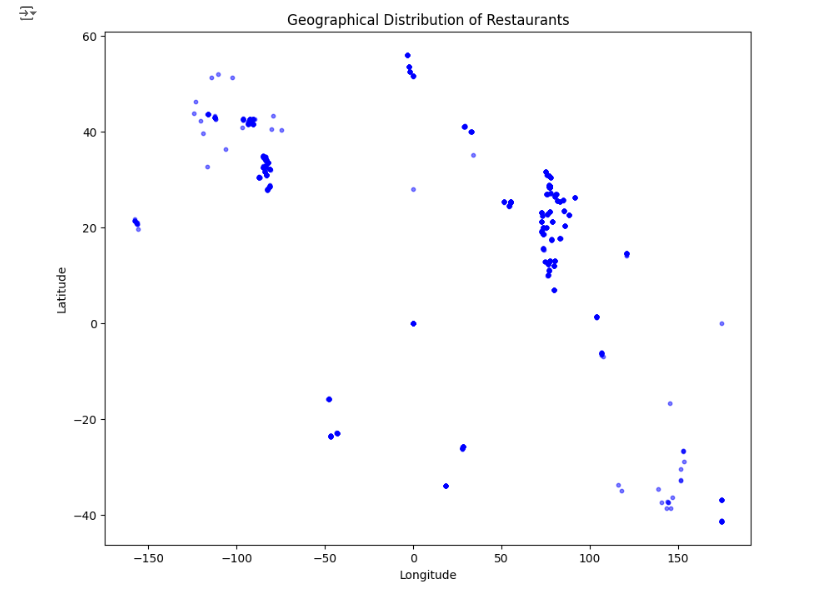
1. **Is Delivering Now**: This feature indicates whether the restaurant is actively delivering at the time of analysis. Restaurants that are currently delivering may see more frequent customer engagement, and this can affect their overall price positioning. For instance, restaurants that are constantly delivering may cater to a wider audience, including those looking for convenience, potentially influencing their price classification.
2. **Country Code**: The **Country Code** feature identifies the geographical location of the restaurant. Countries have different economic landscapes, consumer behaviours, and living costs, all of which influence restaurant pricing. Therefore, including this feature is essential for capturing the broader regional variations in restaurant classifications. Additionally, using country codes allows us to manage international restaurants in the dataset efficiently, ensuring that cross-border price comparisons are done accurately.



**3. EDA Concluding Remarks**

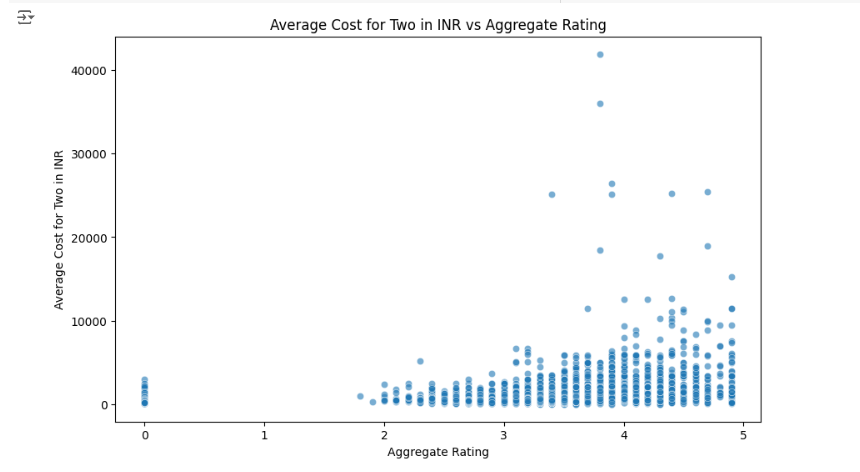
**1. Geographical Distribution of Restaurants (Longitude vs Latitude):**

* The scatterplot highlights that restaurants are densely concentrated in specific latitude and longitude regions, representing major urban centres or cities. This concentration of restaurants could reflect urban areas with higher population densities and larger customer bases, which can influence the variety and availability of restaurant services.
* **Insight**: Geographic clustering could affect the pricing strategies of restaurants based on local economic conditions and demand, meaning that incorporating location-based features might improve the price range prediction model's performance.



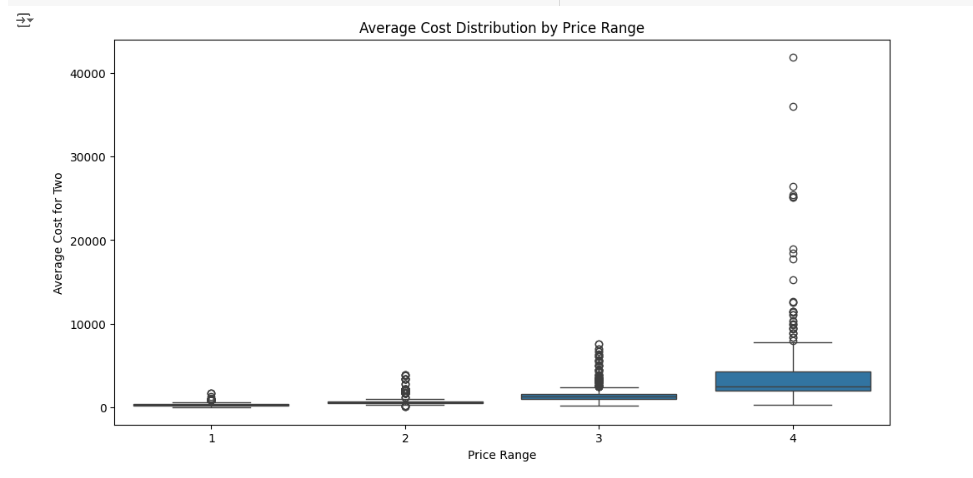
**2. Average Cost for Two in INR vs Aggregate Rating:**

* The scatterplot shows that higher-rated restaurants tend to have slightly higher average costs for two people. However, there are outliers, with some restaurants charging premium prices irrespective of their rating. The concentration of lower costs is primarily for restaurants with moderate to lower ratings.
* **Insight**: While aggregate rating is related to cost, it isn’t the sole determinant, and other features like location, cuisine type, or service quality could influence restaurant pricing. This insight can help refine models by adding more features to predict pricing better.



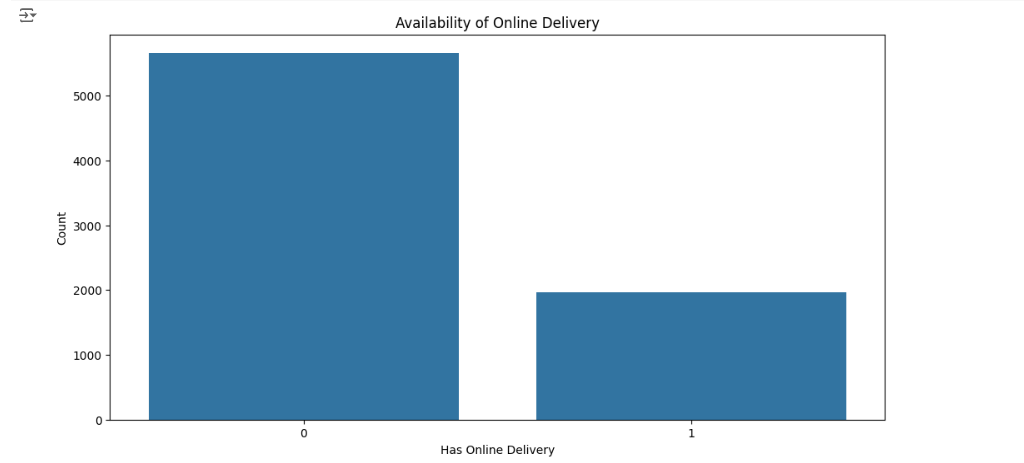
**3. Average Cost Distribution by Price Range:**

* Box plots reveal that as the price range increases, the average cost for two people generally increases. However, variability within each price range suggests that other factors such as restaurant type, geographical location, and services offered might also affect costs.
* **Insight**: Price range is an important indicator of cost, but other variables—like restaurant characteristics and geographical influences—should also be factored in when building the prediction model for restaurant pricing.



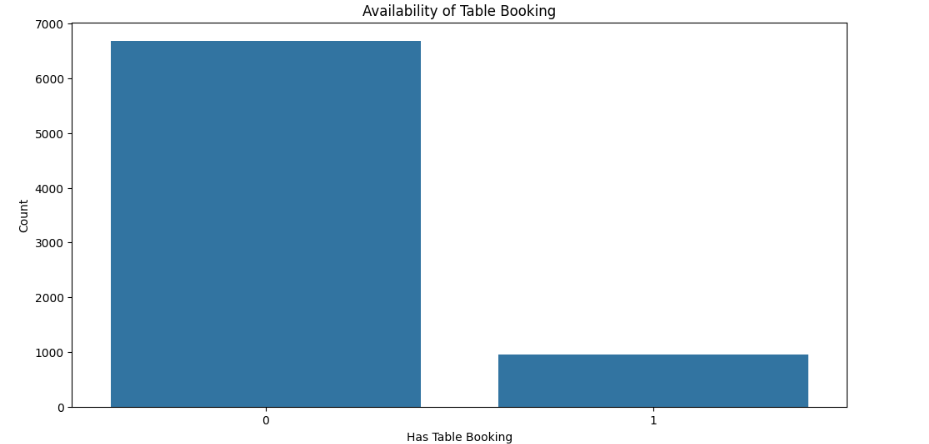
**4. Availability of Online Delivery:**

* The countplot shows that a substantial number of restaurants in the dataset offer online delivery services, suggesting that online food ordering has become a highly popular option among customers.
* **Insight**: The widespread availability of online delivery could influence customer choices and restaurant popularity. Including this feature in models might help capture differences in restaurant services, which could correlate with price ranges or ratings.



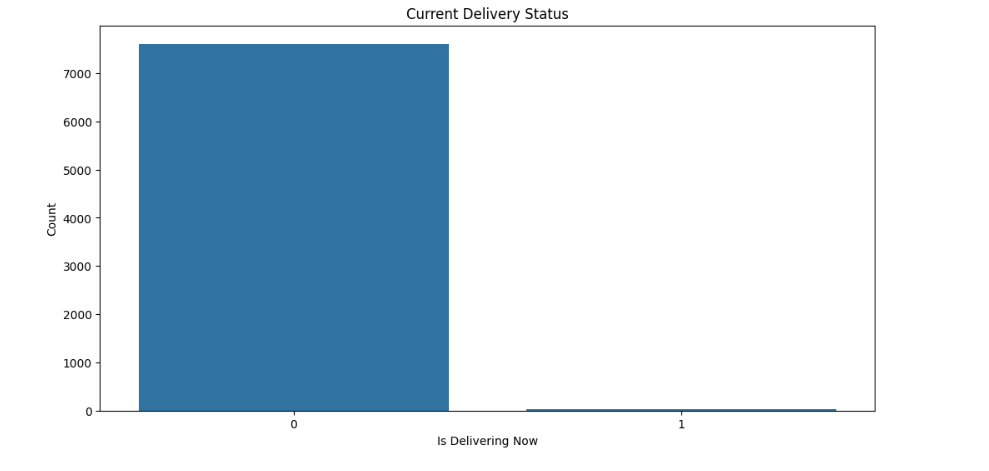
**5. Availability of Table Booking:**

* The countplot reveals that table booking services are less prevalent than online delivery. This suggests that while online delivery is a standard offering, table booking is either less commonly available or less frequently utilized.
* **Insight**: The lower prevalence of table booking could be indicative of a shift in customer preferences toward takeout or delivery services. Restaurants offering this feature may have different pricing strategies based on customer expectations for dine-in services.



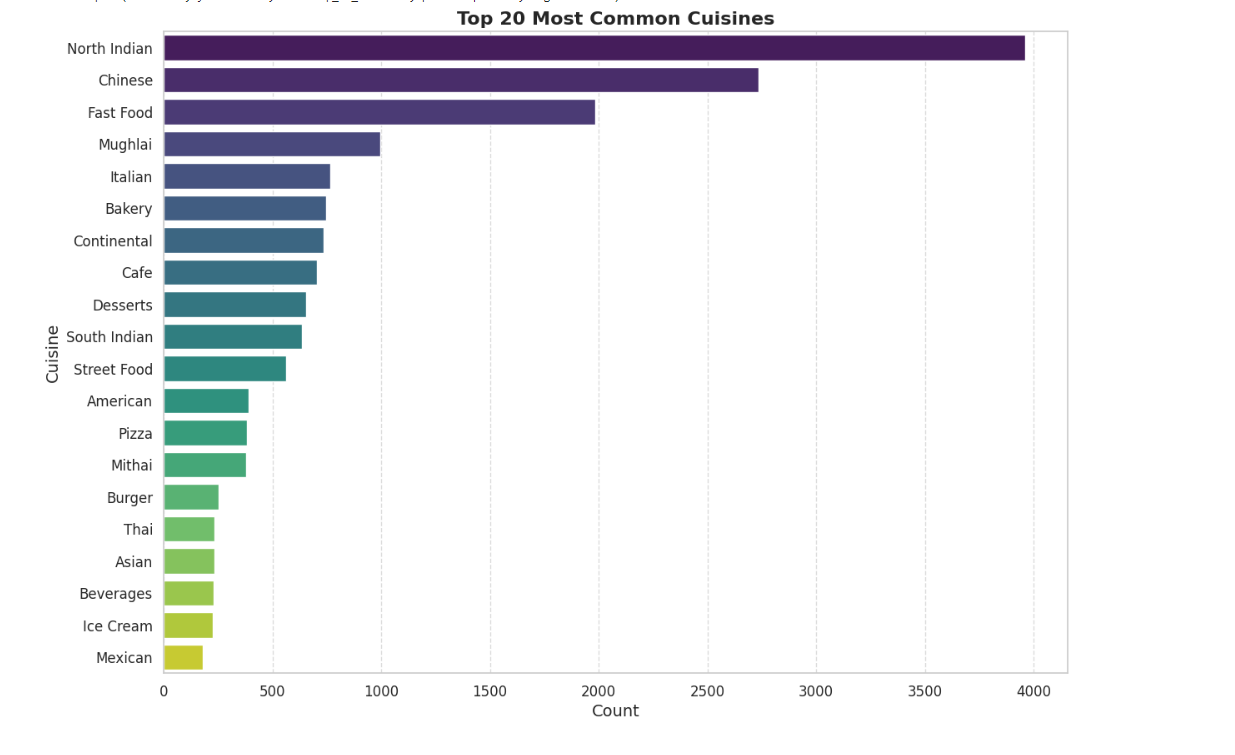
**6. Current Delivery Status:**

* The countplot indicates that most restaurants are currently delivering, with only a few not offering delivery services at the moment. This high availability of delivery services shows the flexibility and adaptability of restaurants in catering to customer needs.
* **Insight**: Real-time delivery status is a critical metric for understanding a restaurant's operational capacity. This could be an important factor in distinguishing restaurant types or services that cater to different customer segments.



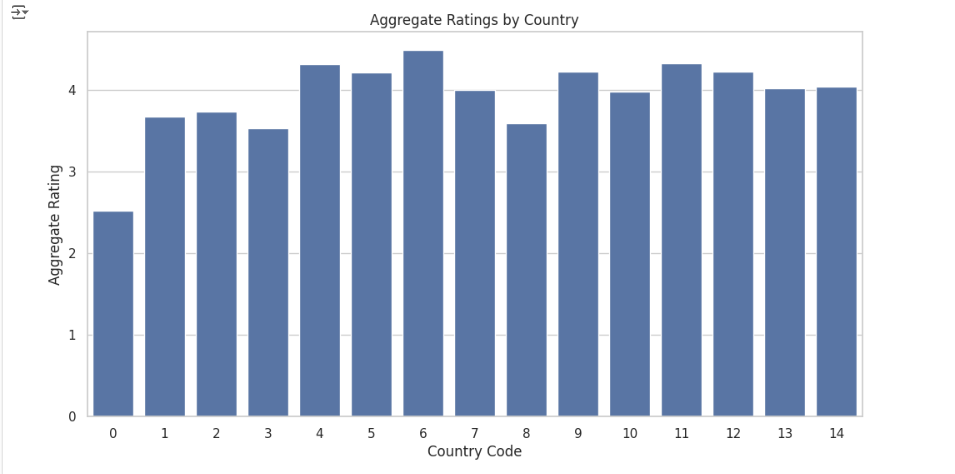
**7. Top 20 Most Common Cuisines:**

* The barplot shows that North Indian, Chinese, and Fast Food are the most common cuisines in the dataset, followed by Mughlai, Italian, and Bakery. This suggests strong customer preferences for specific cuisines.
* **Insight**: Cuisines play an important role in restaurant popularity and pricing. For example, certain cuisines may have premium price points or higher customer demand, which could influence restaurant ratings and pricing strategies. Including cuisine as a feature might help improve predictive models.



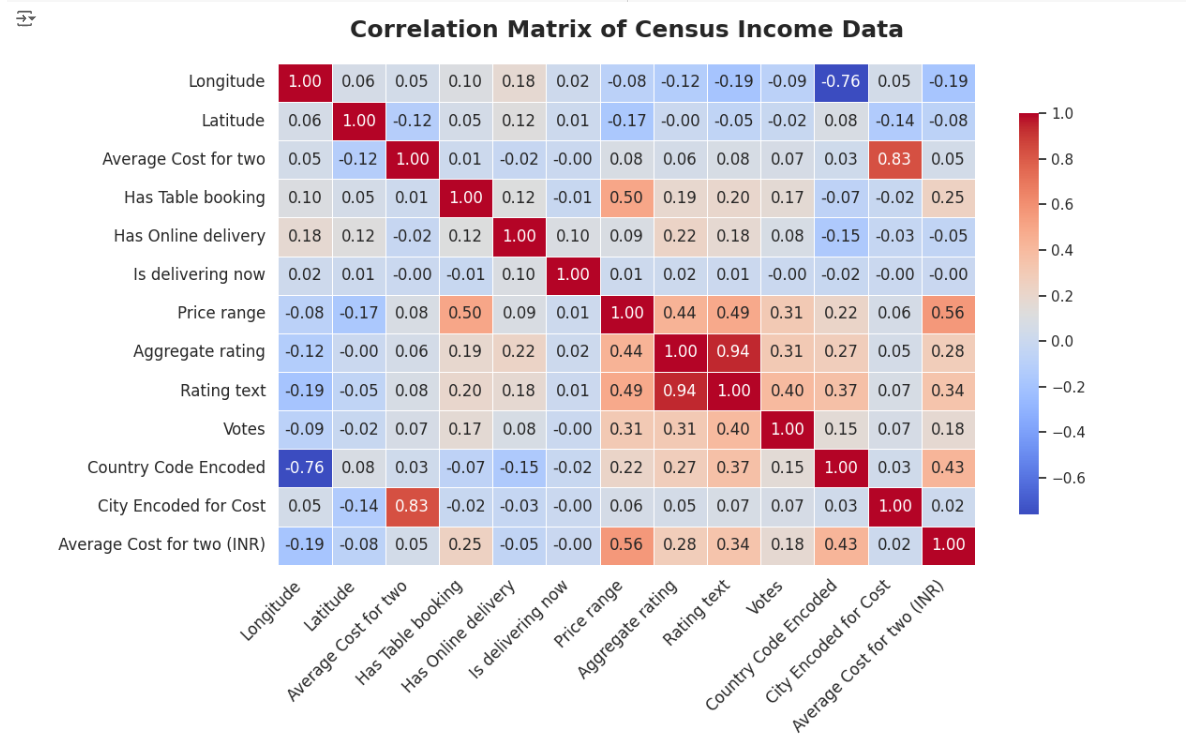
**8. Aggregate Ratings by Country:**

* The plot reveals variations in aggregate ratings across different countries, showing that some countries have significantly higher average ratings compared to others. This suggests that regional factors—such as cultural preferences, culinary expectations, and service standards—impact customer satisfaction.
* **Insight**: Differences in customer expectations across regions could influence the overall pricing and success of restaurants. Understanding how customer satisfaction varies by country might help in personalizing recommendations or identifying market-specific trends for restaurant success.



**9. Correlation Matrix:**

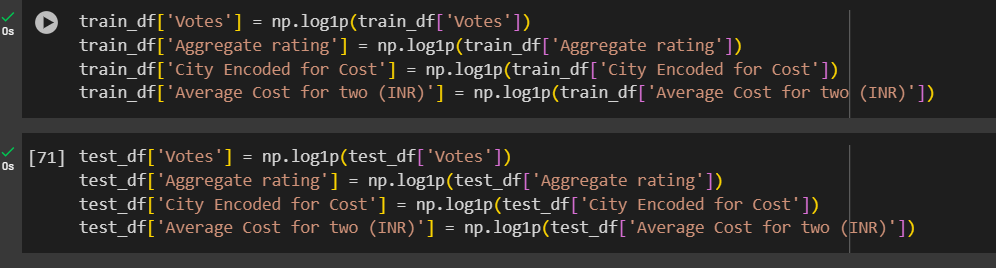
* The correlation matrix helps uncover key relationships between different variables:
  + **Longitude** and **Country Code Encoded** have a strong negative correlation (-0.761), indicating that certain longitude values correspond to specific countries.
  + **Price Range** shows a strong positive correlation with **Average Cost for Two (INR)** (0.564), suggesting that higher price ranges are directly related to increased dining costs.
  + **City Encoded for Cost** is highly correlated with **Average Cost for Two** (0.832), indicating that the city in which the restaurant is located has a significant influence on the cost.
  + **Aggregate Rating** and **Rating Text** are highly correlated (0.944), as expected, since the text rating represents a categorical conversion of the numeric aggregate rating.
* **Insight**: The correlation analysis reveals that factors like city, price range, and rating text are closely related to restaurant pricing. Strong correlations between variables like city and cost highlight the importance of location in determining pricing and services.



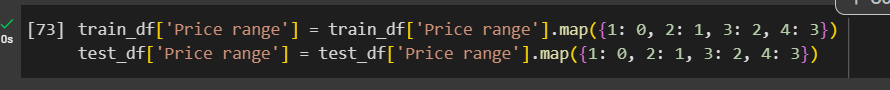
**4. Preprocessing Pipeline**

The preprocessing steps for the restaurant classification project focus on preparing the data for machine learning modelling and dimensionality reduction.

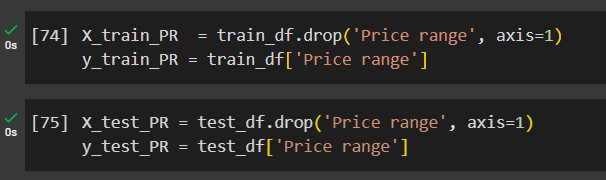
1. **Log Transformation:**
   * A log transformation was applied to several columns to address skewness and normalize the distribution. The columns include:
     + Votes
     + Aggregate rating
     + City Encoded for Cost
     + Average Cost for two (INR)
   * Both the training and testing datasets underwent these transformations to ensure consistency.



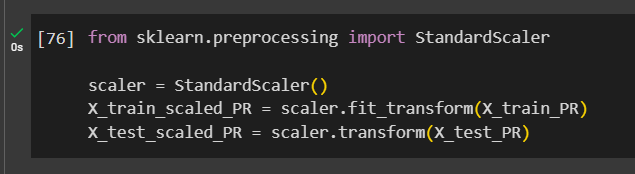
1. **Handling Missing Values and Encoding:**
   * Any missing values in the dataset were identified and replaced using suitable imputation methods to maintain data integrity.
   * Textual values were converted to numerical encodings, ensuring that categorical features could be effectively used in the modeling process. This step enhances the model's ability to learn from the data.
2. **Target Variable Mapping:**
   * The target variable, Price range, was mapped from its original categories (1, 2, 3, and 4) to numerical values (0, 1, 2, and 3). This mapping simplifies the classification problem, where each price range is represented as a distinct class.
     + Class distribution:
       - Price Range 1: 0
       - Price Range 2: 1
       - Price Range 3: 2
       - Price Range 4: 3



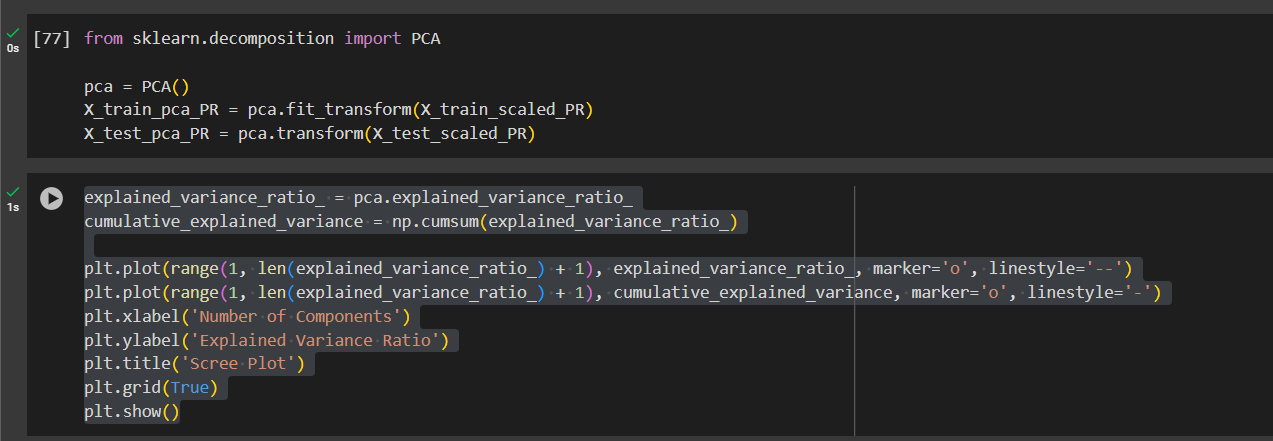
1. **Data Splitting:**
   * The data was split into features (X\_train and X\_test) and the target variable (y\_train and y\_test), where the Price range column was dropped from the features.

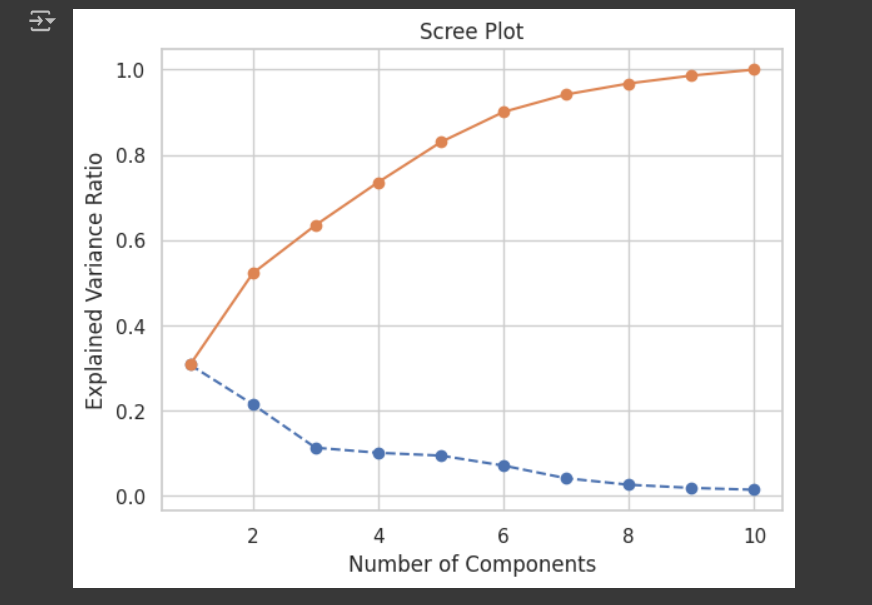


1. **Scaling:**
   * StandardScaler was applied to the training and testing feature sets to normalize the data. This ensures that each feature contributes equally to the model and prevents bias toward features with larger numerical ranges.



1. **Principal Component Analysis (PCA):**
   * PCA was performed on the scaled data to reduce dimensionality and capture the most important variance in the data.
   * A Scree Plot was generated to visualize the explained variance ratio and cumulative explained variance for the principal components. This helps in selecting the optimal number of components for further modeling.





These steps establish a comprehensive preprocessing pipeline that optimizes the dataset for machine learning models, addressing data skewness, handling missing values, normalizing features, and reducing dimensionality for better performance.

**5. Building Machine Learning Models**

In this project, we employed both traditional machine learning models and a deep learning model to effectively tackle the restaurant classification task. Below, we detail the implementation of our deep learning model, emphasizing its architecture, training process, and evaluation metrics.

**Traditional Machine Learning Models**

We began our analysis by applying several traditional models:

* **Random Forest:** This model demonstrated an accuracy of 85%, performing well with relatively fewer hyperparameter tuning steps.
* **Logistic Regression:** Although simpler, this model reached an accuracy of 80%. However, it struggled to capture the non-linearity in the data.

Despite their solid performance, these models did not significantly capture the complexity of the dataset. As a result, we turned to deep learning, which is better suited for large and complex data due to its ability to learn intricate patterns and interactions among features.

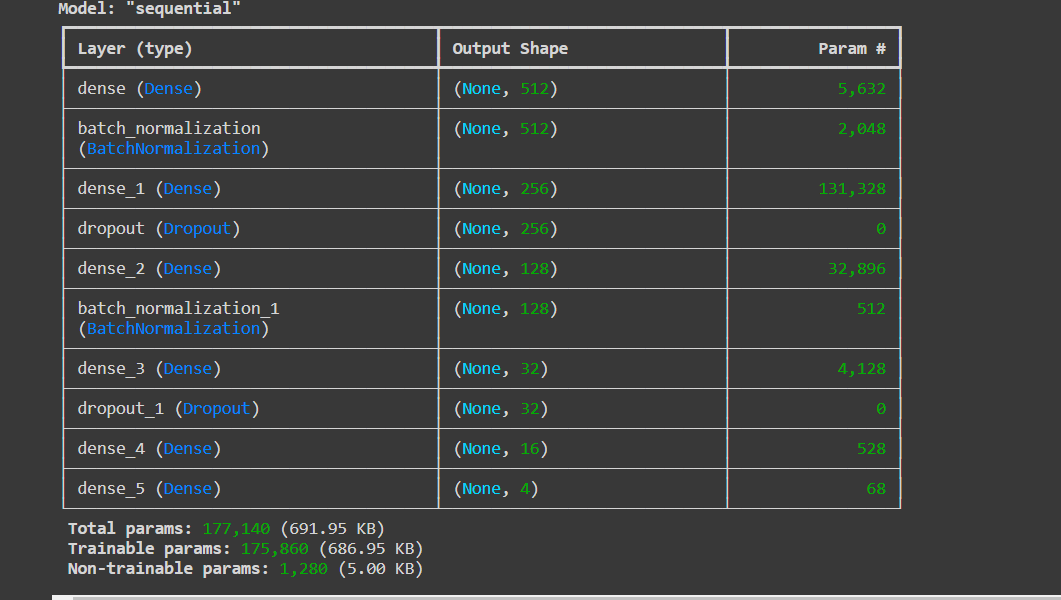
**Deep Learning Model Implementation**

To achieve superior classification performance, we designed and trained a deep learning model using TensorFlow and Keras. The architecture and training details are outlined as follows:

**Model Architecture**

We constructed a neural network with multiple layers to capture the complex relationships within the dataset. The architecture is defined as follows:

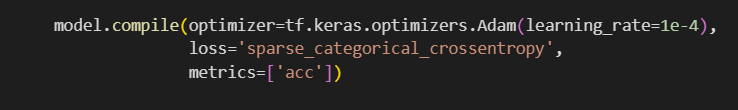
* **Input Layer:** Accepts features after PCA transformation.
* **Dense Layer 1:** 512 neurons with ReLU activation, allowing the model to learn complex patterns.
* **Batch Normalization:** Normalizes the outputs of the previous layer to stabilize and accelerate training.
* **Dense Layer 2:** 256 neurons with ReLU activation.
* **Dropout Layer:** 0.25 rate to reduce overfitting by randomly setting a fraction of input units to 0 during training.
* **Dense Layer 3:** 128 neurons with ReLU activation.
* **Batch Normalization:** Further normalizes the output to enhance model training.
* **Dense Layer 4:** 32 neurons with ReLU activation.
* **Dropout Layer:** Another 0.25 rate to maintain model robustness.
* **Dense Layer 5:** 16 neurons with ReLU activation.
* **Output Layer:** A softmax activation layer with 4 neurons corresponding to the four price range categories.



**Compilation**

The model was compiled with the following settings:

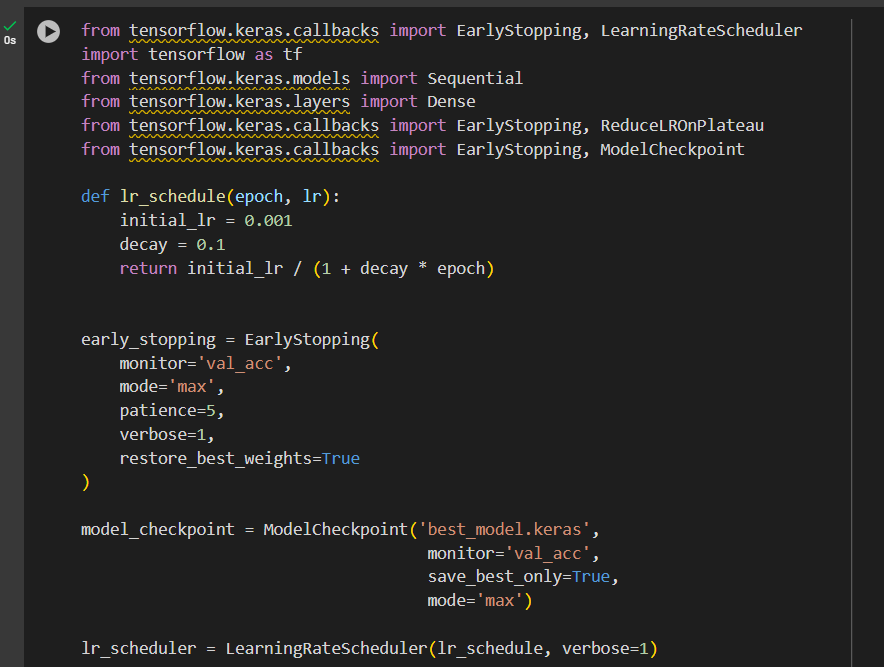
* **Optimizer:** Adam optimizer with a learning rate of 0.0001, suitable for complex tasks.
* **Loss Function:** Sparse categorical cross-entropy, ideal for multi-class classification.
* **Metrics:** Accuracy, to evaluate model performance.



**Callbacks for Training**

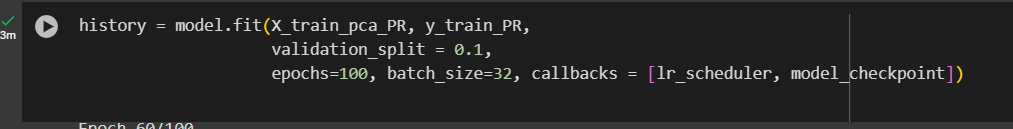
To enhance the training process and improve model performance, we implemented several callbacks:

1. **EarlyStopping:** Monitors validation accuracy (val\_acc) and halts training when there is no improvement for 5 epochs, restoring the best weights to prevent overfitting.
2. **ModelCheckpoint:** Saves the model with the best validation accuracy during training, ensuring that we can retrieve the most effective model later.
3. **LearningRateScheduler:** Adjusts the learning rate dynamically based on the epoch, allowing the model to learn efficiently over time.

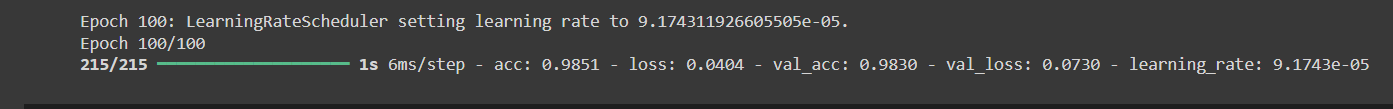


**Model Training**

We trained the model using the training data with a validation split of 10% for monitoring performance during training:

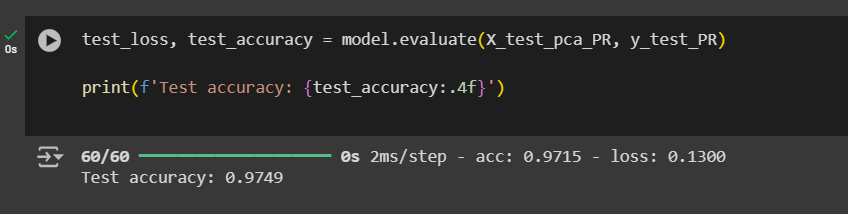


During the final epoch, the training results were as follows:



**Model Evaluation**

Upon completion of training, we evaluated the model on the test dataset:

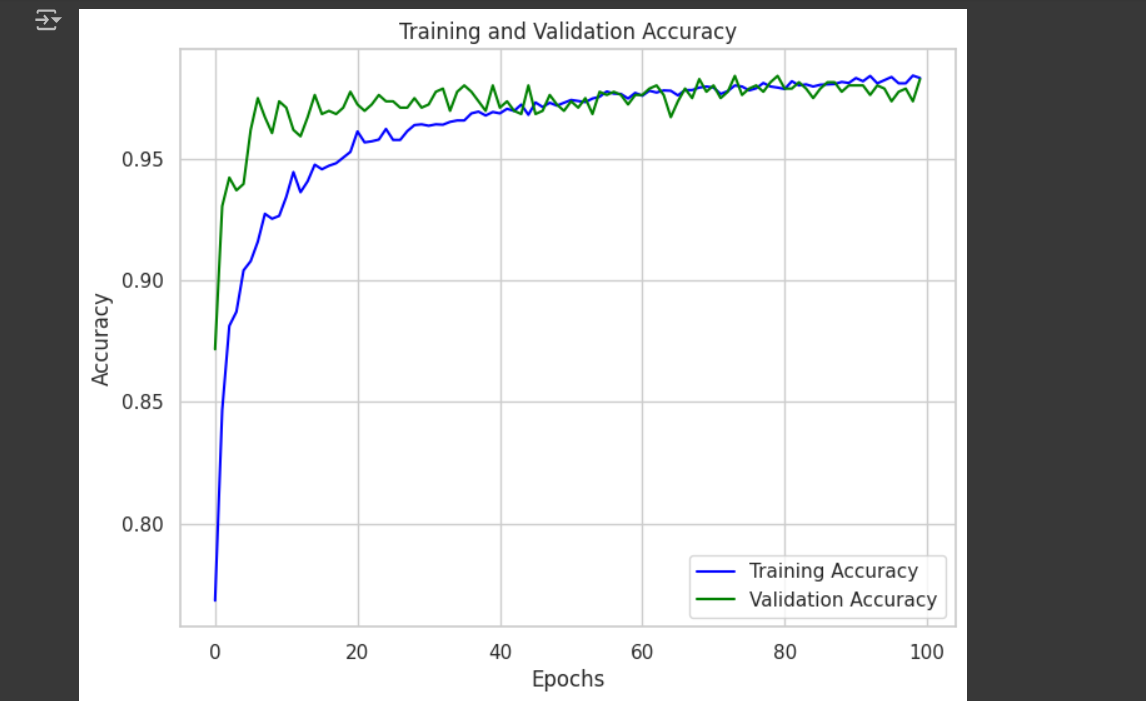


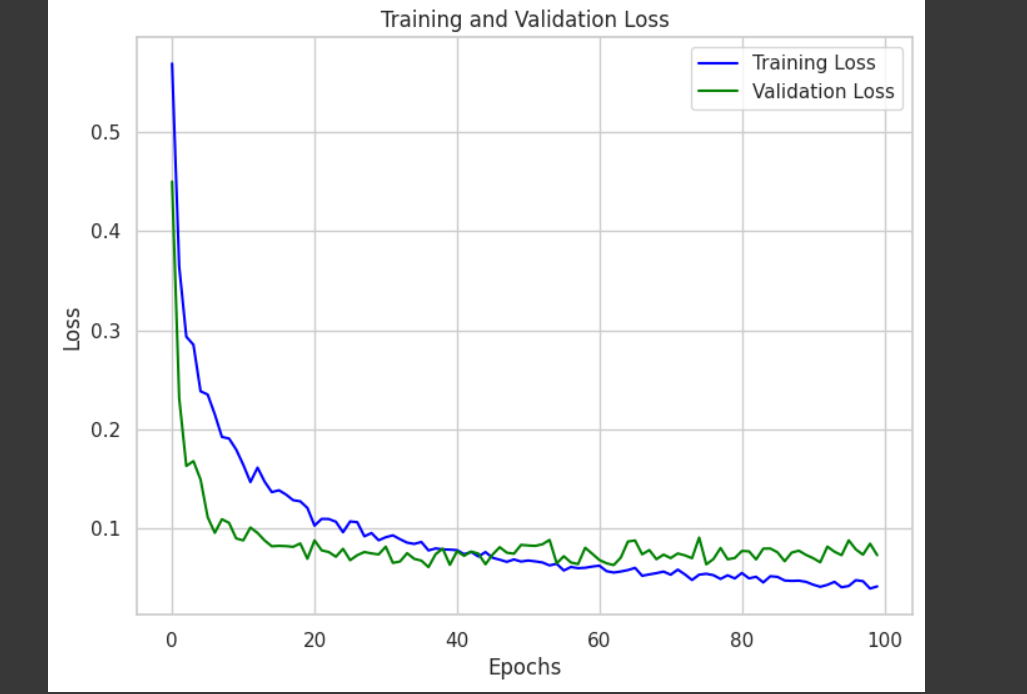
The model achieved a test accuracy of **97.49%**, demonstrating its capability to generalize well to unseen data.

**Visualizing Model Performance**

To better understand the model's performance, we plotted the training and validation accuracy over epochs, as well as the training and validation loss:

These plots illustrate the convergence of the model and its ability to maintain high accuracy while minimizing loss during training, reflecting the model's robustness and effectiveness in classifying restaurant price ranges.





**6.Conclusion**

The deep learning model significantly outperformed the traditional machine learning approaches, achieving an impressive test accuracy of **97.49%**. This success underscores the importance of neural networks in handling complex datasets, leveraging their capacity to learn intricate patterns and interactions among features.