*A Mini Project Report on*

**Restaurant Review Analysis**

**Using Machine Learning**

*Submitted in partial fulfilment of the requirements for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

In

**CSE (DATA SCIENCE)**

By

**P. Praneetha** **(22AG5A6702)**

**B. Rajasri Aishwarya (21AG1A6710)**

**Mohammed Adil (21AG1A6744)**

**V. Mani Charan (22AG5A6705)**

Under the guidance of

**MS. Swathi Turai**

Assistant Professor



**DEPARTMENT OF CSE (DATA SCIENCE)**

**ACE Engineering College**

**Ankushapur(V), Ghatkesar(M), Medchal Dist - 501301**

****

***(An Autonomous Institution, Affiliated to JNTUH, Hyderabad)***

[**www.aceec.ac.in**](http://www.aceec.ac.in/)

**2024-2025**

**DEPARTMENT OF CSE (DATA SCIENCE)**

**CERTIFICATE**

This is to certify that the mini project report entitled “***Restaurant Review Analysis Using Machine Learning****”* is a Bonafide work done by P. Praneetha(22AG5A6702), B. Rajasri Aishwarya (21AG1A6710), Mohammed Adil (21AG1A6744), and V. Mani Charan (22AG5A6705) in partial fulfilment for the award of Degree of BACHELOR OF TECHNOLOGY in ***CSE (Data Science)*** from JNTUH University, Hyderabad during the academic year 2024- 2025. This record of Bonafide work carried out by them under our guidance and supervision.

The results embodied in this report have not been submitted by the student to any other University or Institution for the award of any degree or diploma.

**Ms. P. Swathi Turai Dr. P. Chiranjeevi External**

Assistant Professor Associate Professor

Supervisor HOD, CSE-DS

**ACKNOWLEDGEMENTS**

We would like to express my gratitude to all the people behind the screen who have helped me transform an idea into a real-time application. We would like to express my heart-felt gratitude to my parents without whom we would not have been privileged to achieve and fulfil my dreams.

A special thanks to our General Secretary, **Prof. Y. V. Gopala Krishna Murthy**, for having founded such an esteemed institution. Sincere thanks to our Joint Secretary **Mrs. M. Padmavathi**, for support in doing project work. I am also grateful to our beloved principal, **Dr. B. L. RAJU** for permitting us to carry out this project.

We profoundly thank **Dr. P. Chiranjeevi**, Associate Professor and Head of the Department of Computer Science and Engineering (Data Science) who has been an great source of inspiration to my work.

We extremely thank **Mr. Shaik Nagar Vali** and **Mr. P. Ashok Kumar**, Associate Professors, Project coordinators, who helped us in all the way in fulfilling of all aspects in completion of our Mini-Project.

We are very thankful to my internal guide **Ms. Swathi Turai (Assistant professor)** who has been an excellent and also given continuous support for the Completion of my project work.

The satisfaction and euphoria that accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, who have extended their timely help and eased my task.

P. Praneetha (22AG5A6702)

B. Rajasri Aishwarya (21AG1A6710)

Mohammed Adil (21AG1A6744)

V. Mani Charan (22AG5A6705)

**RESTAURANT REVIEW ANALYSIS USING MACHINE LEARNING**

**ABSTRACT**

In today's digital world, food apps are becoming more popular because they allow users to easily browse, book, and order food from their favourite restaurants with just a few clicks. Apps like Zomato, Swiggy let users rate their dining experience and share reviews. These reviews help other customers and restaurants understand how well they are doing, but sometimes it's hard for restaurants to tell if the feedback is positive or negative. We started by exploring the data (EDA) and found the most and least expensive restaurants. We also identified top critics who have reviewed over 100 restaurants and have more than 10,000 followers, which restaurants should focus on. For grouping similar restaurants, we used clustering techniques and settled on 3 groups based on analysis. We used KMeans and Hierarchical clustering, with features like cuisine and cost. For analysing the sentiment of reviews (whether they're positive or negative), we used both supervised (with labelled data) and unsupervised (without labelled data) methods. Supervised methods included algorithms like Logistic Regression, Decision Trees, and Multinomial Naive Bayes, while unsupervised methods included techniques like Linear Discriminant Analysis. We defined ratings above 3.5 as positive and below 3.5 as negative. After tuning the models, Logistic Regression and LightGBM worked best for predicting review sentiments

**INDEX**

**S. no Chapter Name Page No.**

1. **Introduction**
2. **` Literature Survey**
3. **System Analysis**
4. **Software Requirements**
5. **Hardware Requirements**
6. **Functional Requirements**
7. **Non-Functional Requirements**
8. **Methodology**
9. **System Design**
10. **Implementation**
11. **Results**
12. **Conclusion**
13. **Future Work**
14. **References**

**LIST OF FIGURES**

**S.no Figure Name Page No.**



**Chapter 1. Introduction**

**1.1 Background and Context of the Project:**

The project focuses on analyzing restaurant ratings and reviews, a growing area of interest due to the increasing popularity of food apps. These apps are widely used for ordering meals and providing feedback, which poses a challenge for restaurants to interpret unstructured text reviews accurately. This analysis aims to help restaurants better understand customer sentiment, thereby enhancing service quality and customer satisfaction.

**1.2 Objectives**

* The main goal is to analyze restaurant reviews to provide insights that enhance customer satisfaction, optimize operations, and support better decisions.
* To implement advanced AI techniques for detecting fake reviews, ensuring the reliability and integrity of customer feedback data for better decision-making.

**1.3 Project Type**

* The project type is Data Science and Machine Learning with a focus on Natural Language Processing (NLP) for sentiment analysis, clustering, and predictive modeling in the domain of restaurant reviews and ratings.

**1.4 Scope of Project**

1. **Improved Customer Experience**:
   * Provide personalized restaurant recommendations based on analyzed reviews.
   * Help customers make informed dining decisions.
2. **Actionable Insights for Restaurants**:
   * Enable restaurants to identify customer sentiment, trends, and areas for improvement.
   * Detect fake reviews to ensure reliable feedback for operational strategies.
3. **Enhanced Industry Practices**:
   * Establish a standardized approach to review and sentiment analysis in the food service industry.
   * Support data-driven decision-making to enhance service quality and customer satisfaction.
4. **Broader Applications**:
   * The system can be adapted for other industries, such as retail, travel, or hospitality, where customer reviews play a critical role.
5. **Scalable and Future-ready**:  
   * Integration of advanced NLP techniques like BERT ensures adaptability for larger datasets and evolving customer feedback formats.
   * Potential to include real-time monitoring and analytics for proactive decision-making.

**1.5 Technologies Used**

1. **Programming Languages & IDEs**:
   * Python
   * Jupyter Notebook
   * VSCode
2. **Libraries and Frameworks**:
   * Data Processing & Analysis: Pandas, NumPy
   * Machine Learning: Scikit-learn, XGBoost, LightGBM, Random Forest
   * Deep Learning: BERT, GPT-based models (for sentiment analysis)
   * Visualization: Matplotlib, Seaborn, Word Cloud
3. **Clustering Techniques**:
   * K-Means Clustering
   * Hierarchical Clustering
4. **Feature Engineering & NLP**:
   * Tokenization and Vectorization (e.g., TF-IDF, Count Vectorizer)
   * Dimensionality Reduction (PCA)
5. **Operating Systems**:
   * Windows
   * macOS
   * Linux
6. **Other Tools**:
   * TripAdvisor API for dataset integration
   * Plagiarism checker for fake review detection.

**Chapter 2. Literature Survey**

The growing popularity of online food applications has introduced new challenges for understanding customer feedback. Sentiment analysis, clustering, and predictive modeling have been identified as crucial tools for addressing these challenges. Below is an overview of existing work relevant to the project:

**2.1 Sentiment Analysis**

Sentiment analysis is a well-researched field aimed at extracting opinions from text. Techniques range from classical machine learning algorithms such as Logistic Regression to deep learning models like BERT and GPT, which excel in contextual understanding of text. Research highlights that deep learning methods outperform traditional techniques in handling unstructured and noisy datasets.

**2.2 Clustering Techniques**

Clustering has been applied to group similar reviews, revealing patterns in customer sentiment. K-means and hierarchical clustering are commonly used algorithms. Studies demonstrate that K-means effectively clusters reviews based on preprocessed features, while hierarchical clustering provides more nuanced groupings for smaller datasets​.

**2.3 Topic Modeling**

Topic modeling, such as Latent Dirichlet Allocation (LDA), has been widely utilized to identify frequently discussed themes in customer reviews. Studies underscore its ability to uncover latent topics, which aids restaurants in pinpointing specific areas for improvement​.

**2.4 Predictive Modeling**

Machine learning models, including Random Forest, XGBoost, and LightGBM, have been employed to predict future restaurant ratings based on historical data. These models have shown high accuracy in leveraging structured and unstructured data to forecast customer satisfaction​.

**2.5 Fake Review Detection**

Research into AI-powered solutions for identifying fake reviews highlights the use of neural networks combined with plagiarism-check algorithms. Such techniques improve trustworthiness by filtering out manipulative feedback that can skew restaurant ratings​.

**2.6 Trends and Future Directions**

Recent works emphasize integrating advanced Natural Language Processing (NLP) methods, expanding datasets, and exploring real-time applications. These enhancements aim to increase model accuracy and make insights actionable for restaurant owners​.

**2.7 Tools and Libraries**

Python libraries like pandas, scikit-learn, and TensorFlow have been extensively used for preprocessing, modeling, and evaluation in sentiment analysis and clustering studies. Visualization tools like Matplotlib and Seaborn provide actionable insights​.

**Chapter 3. System Analysis**

**3.1 Existing System**

1. **Sentiment Analysis**: Identifies whether reviews are positive, negative, or neutral, revealing overall customer satisfaction.
2. **Topic Modeling**: Extracts keywords and topics from reviews to highlight areas for improvement.
3. **Clustering**: Groups similar reviews to uncover common feedback or exceptional experiences.
4. **Predictive Modeling**: Uses machine learning to predict future ratings based on review patterns.

**3.2 Proposed System**

1. **Machine Learning for Sentiment Analysis**:
   * Implement advanced models like BERT or GPT for understanding the context and meaning behind customer reviews more effectively.
2. **Personalized Recommendation System**:
   * Use review data to provide tailored restaurant suggestions to customers based on their preferences.
3. **Trend Monitoring**:
   * Track changes in sentiment, topics, and ratings over time to identify emerging trends in customer feedback.
4. **AI-powered Fake Review Detection**:
   * Develop an AI bot to identify fake reviews using username analysis and plagiarism detection techniques.
5. **Scalable Insights**:
   * Enhance analysis capabilities to handle large datasets, providing robust and reliable insights for restaurants.

**Chapter 4. Software Requirements**

The Software Requirements Specification for the project includes the following components:

* **Python Libraries**: The system will require several Python libraries for data manipulation, analysis, and visualization. These include:  
  + *Pandas* for data handling and manipulation.
  + *Scikit-learn* for machine learning tasks and model building.
  + *Matplotlib* and *Pyplot* for creating static, animated, and interactive visualizations.
  + *Seaborn* for statistical data visualization and enhancing the aesthetic of the plots.
  + *Wordcloud* for generating word clouds from text data.
  + *NumPy* for numerical operations and handling large arrays efficiently.
* **Integrated Development Environment (IDE)**: The project can be developed and tested in popular IDEs such as:  
  + *Jupyter Notebook* for interactive coding and data visualization.
  + *VSCode* for a more traditional coding environment with support for Python.
* **Operating System**: The project is compatible with multiple operating systems, including:  
  + *Windows*, *macOS*, and *Linux*, ensuring cross-platform usability and development flexibility.

These requirements ensure that the project can be implemented effectively with the necessary tools for data analysis, machine learning, and visualization.

**Chapter 5. Hardware Requirements**

**Processor:**

* **Intel Core i5 or higher**.  
  This ensures sufficient processing power for running complex computational tasks, including CNN models.  
  A higher processor speed is recommended for faster model training and data processing.

**RAM:**

* **Minimum 6GB of RAM**.  
  This provides better performance when running Python scripts, Jupyter Notebook, and other software components simultaneously.  
  Upgrading to 8GB or more is advised for handling larger datasets efficiently.

**Storage:**

* **100GB of free space**.  
  This is essential for installing necessary software packages, storing datasets, and managing any additional files generated during the project.  
  SSD storage is preferable for quicker data access and improved system responsiveness.

**Chapter 6. Functional Requirements**

1. **Data Ingestion and Preprocessing**
   1. Collect raw data from the input source and clean, format, and prepare it for analysis.
   2. Handle missing values, remove duplicates, and perform necessary transformations to ensure data consistency.
2. **Sentiment Analysis**
   1. Implement sentiment analysis using **Logistic Regression** and **LightGBM** models to classify reviews as positive, negative, or neutral.
   2. Compare the performance of both models to identify the most effective approach for the dataset.
3. **Clustering of Reviews**
   1. Group similar reviews together using the **K-Means** clustering algorithm.
   2. Analyze clusters to gain insights into customer sentiment and common themes.
4. **Visualization of Results**
   1. Create graphical representations of the processed data and analysis outcomes.
   2. Use charts, graphs, and interactive dashboards for better interpretation and decision-making.

**Chapter 7. Non-Functional Requirements**

**Performance**

* Ensure **high accuracy** in both sentiment classification and review clustering.
* Optimize algorithms to deliver quick and precise results, even with large datasets.
* Maintain scalability to handle increasing data volumes without compromising speed or accuracy.

**Reliability**

* Implement regular **data backup** mechanisms to prevent data loss.
* Ensure seamless **system updates** to maintain functionality and compatibility with evolving requirements.
* Monitor system performance with error-handling mechanisms to quickly identify and resolve issues.

**Usability**

* Design an **easy-to-navigate user interface** to enhance the user experience.
* Provide clear documentation and help features for users to understand the system efficiently.
* Incorporate user feedback loops for continuous improvement in usability and functionality.

**Chapter 8. Methodology**

**8.1 Data Collection**

* 1. Gather review datasets from reliable online sources such as review platforms and social media.
  2. Ensure the dataset is diverse and representative of the target domain.
  3. Validate data quality by checking for inconsistencies, duplicates, and anomalies before preprocessing.

**8.2 Preprocessing**

a. Clean and normalize text data by removing noise, such as special characters, stop words, and unnecessary spaces.

b. Perform tokenization, lemmatization, and handling of missing data to prepare the dataset for analysis.

c. Incorporate techniques like stemming or entity recognition to refine the data further.

**8.3 Feature Engineering**

a. Convert textual data into numerical representations using **TF-IDF (Term Frequency-Inverse Document Frequency)**.

b. Explore additional techniques like word embeddings (e.g., Word2Vec or GloVe) for advanced feature extraction.

c. Implement feature selection methods to reduce dimensionality and improve model efficiency.

**8.4 Model Training**

a. Train and evaluate sentiment analysis models, such as **Logistic Regression** and **LightGBM**, using processed data.

b. Perform hyperparameter tuning and cross-validation to optimize model performance.

c. Analyze model performance using metrics like accuracy, precision, recall, and F1-score for robust evaluation.

**8.5 Deployment**

* 1. Develop a user-friendly interface using **Streamlit** for real-time visualization of results.
  2. Integrate the system with backend services to ensure smooth and responsive performance.
  3. Enable interactive features, such as filters and dynamic visual updates, to enhance user engagement.

**Chapter 9. System Design**

**9.1 Architecture**

The system architecture includes for data ingestion preprocessing, sentiment analysis, clustering, and visualization.

**9.2 Workflow**

1. Input data from review datasets.
2. Preprocess and clean the data.
3. Apply machine learning models for sentiment classification.
4. Visualize results using interactive dashboards.

**Chapter 10. Implementation**

**10.1 Tools and Technologies**

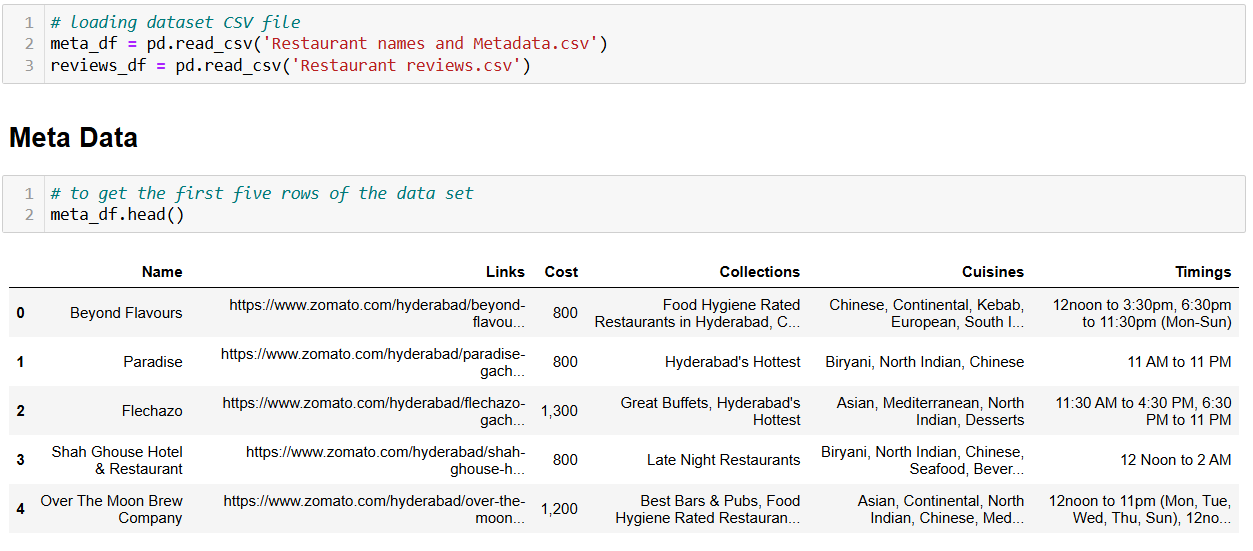
* Python
* Scikit-learn, Pandas
* Visualization libraries: Matplotlib, Seaborn

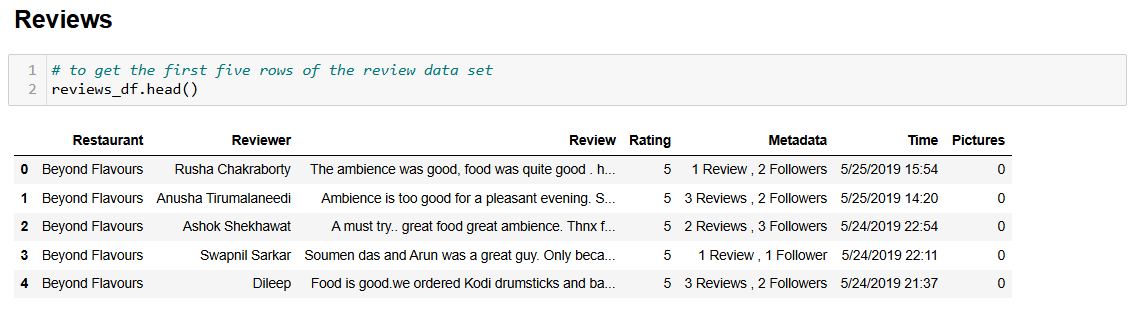
**10.2 Process**

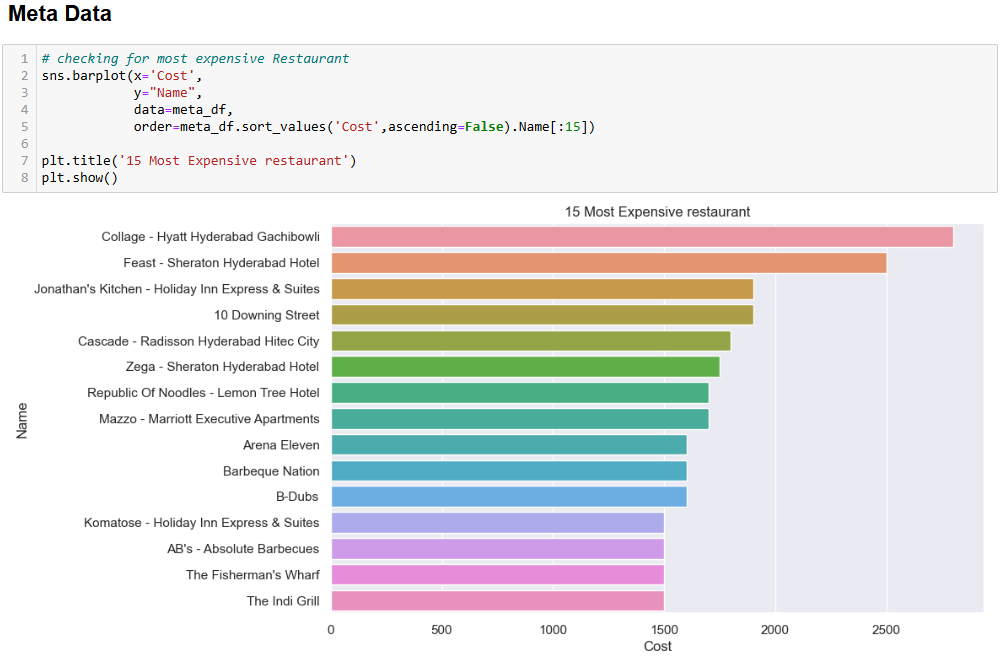
* Train sentiment models on labeled data.
* Cluster reviews to identify patterns and trends.
* Develop a user-friendly dashboard.

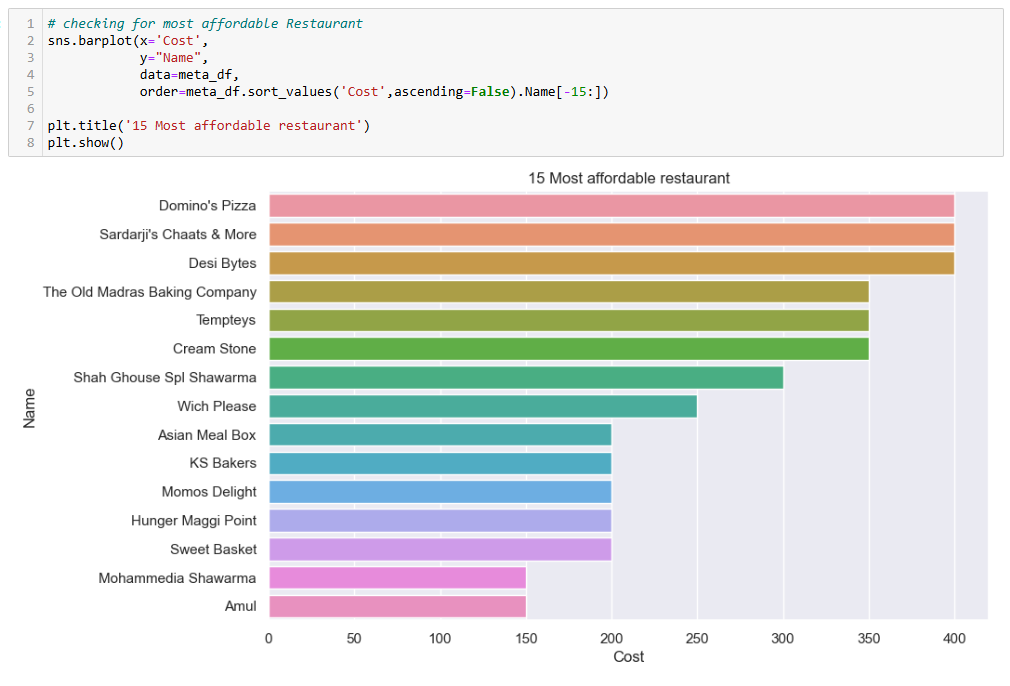
**Chapter 11. RESULTS**

**11.1.1 CSV Data Set of Restaurant Reviews and Ratings**

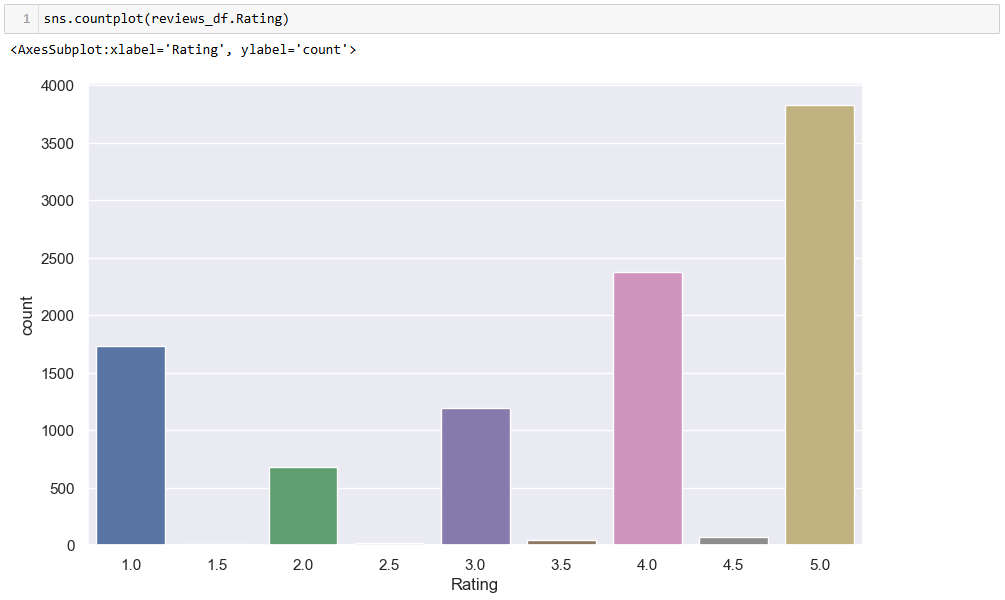
****

****

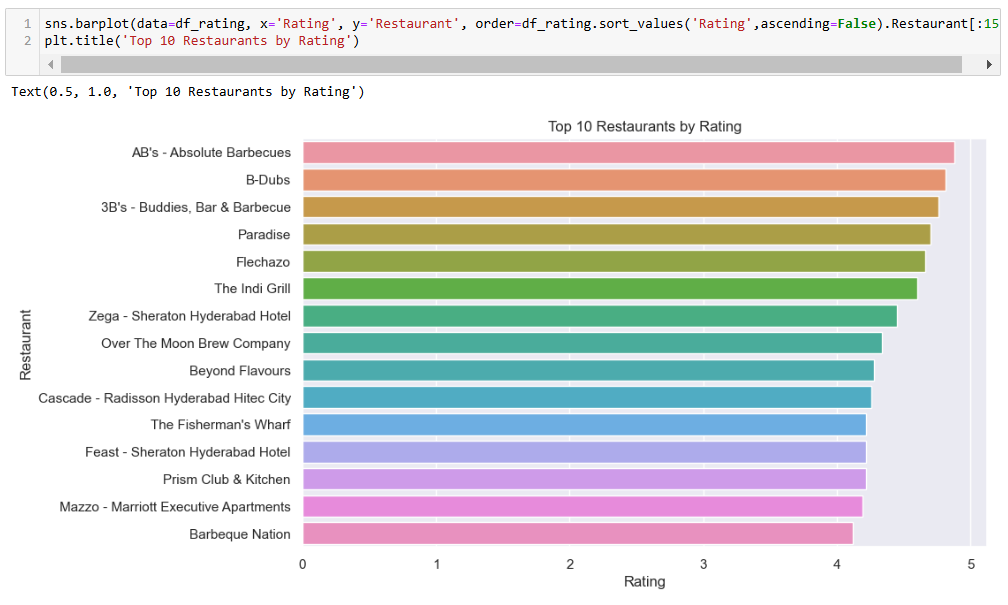
****

****

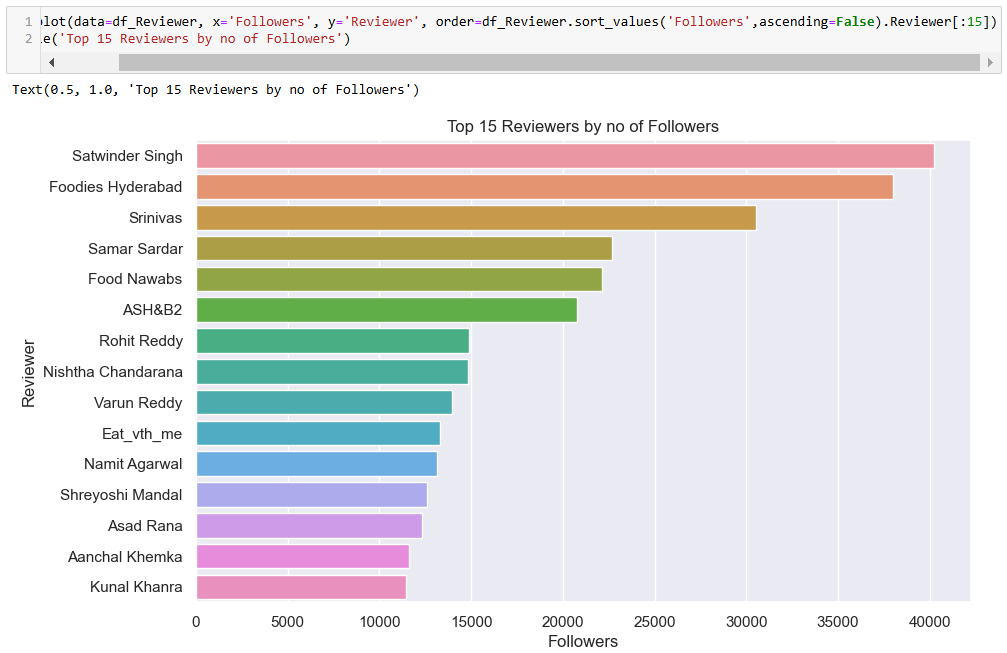
**11.1.2 RATING ANALYSIS**

****

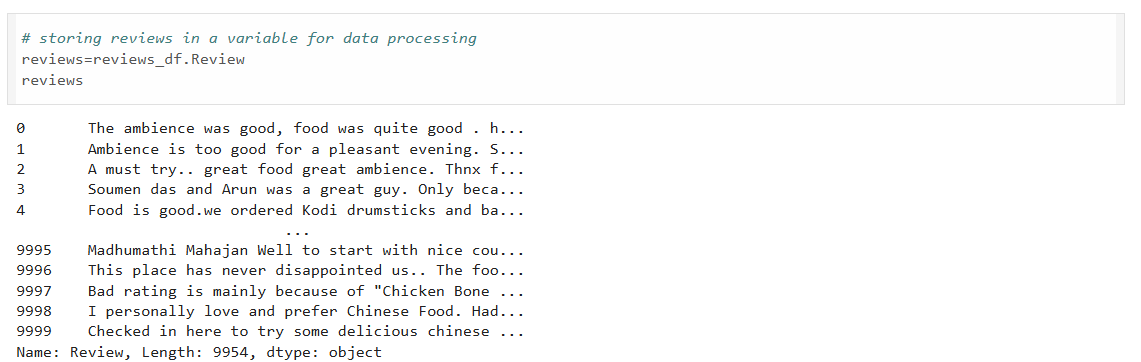
**11.1.3 RESTAURANT WITH BEST REVIEWS**



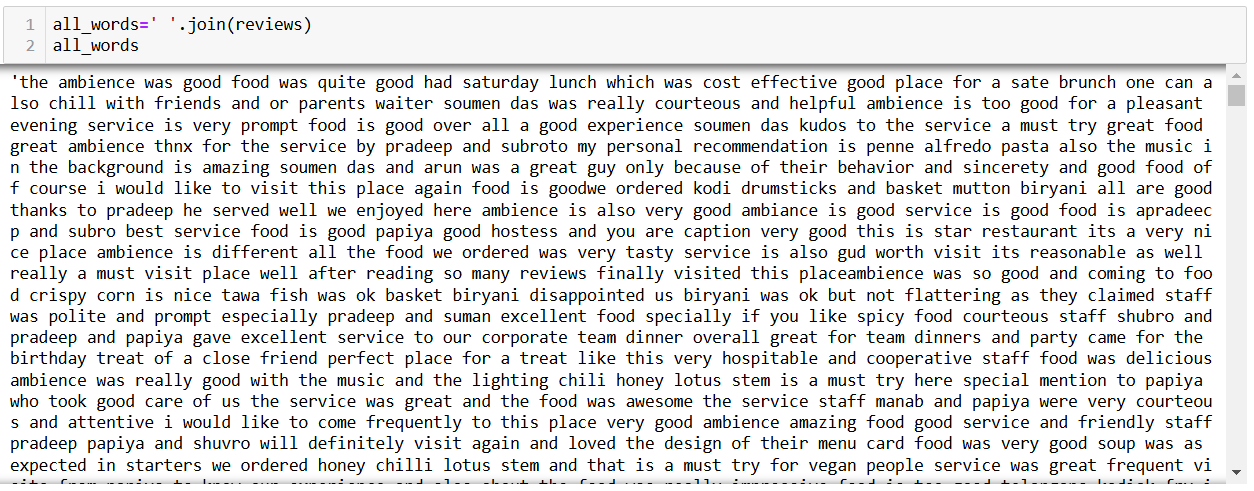
**11.1.4 ANALYSISNG THE REVIEWS OF INFLUNCERS**



**11.2.1 TEXT PROCESSING**





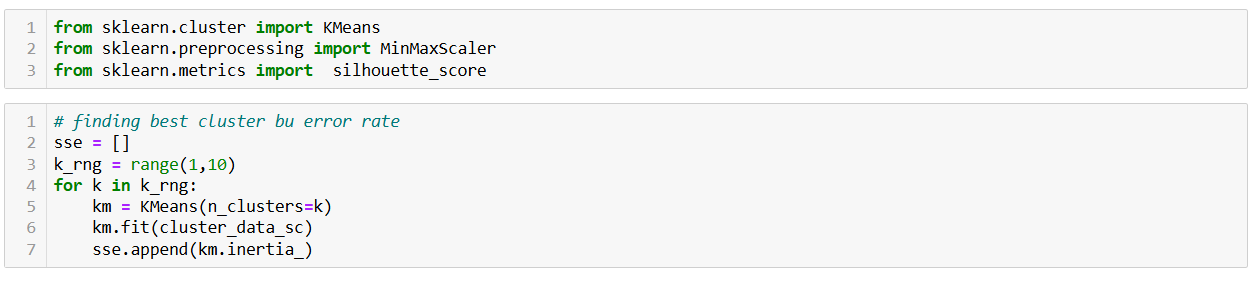


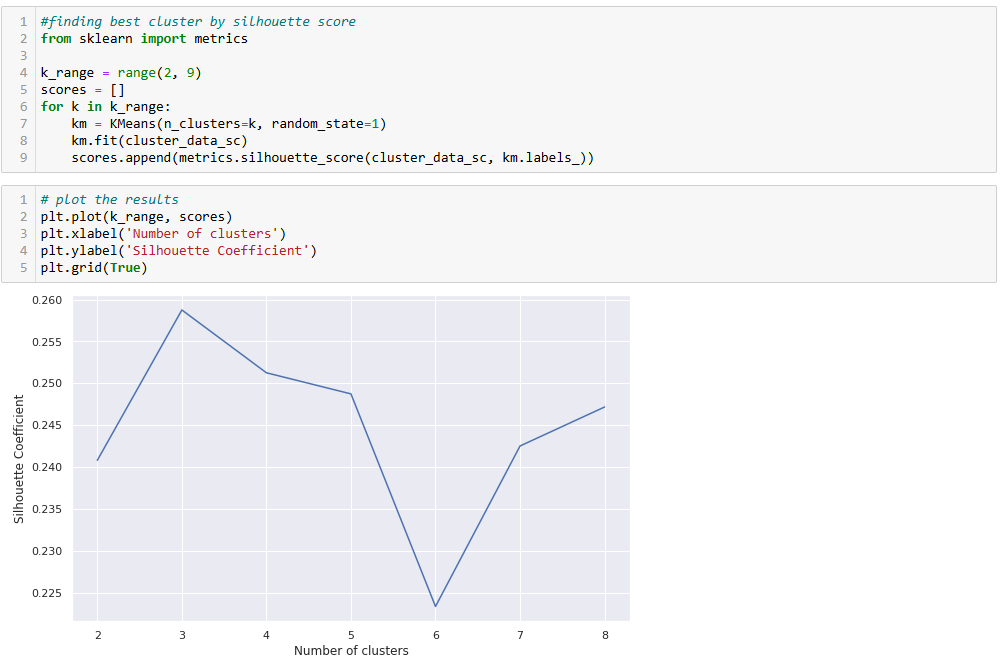
**11.2.2 Performing Clustering**



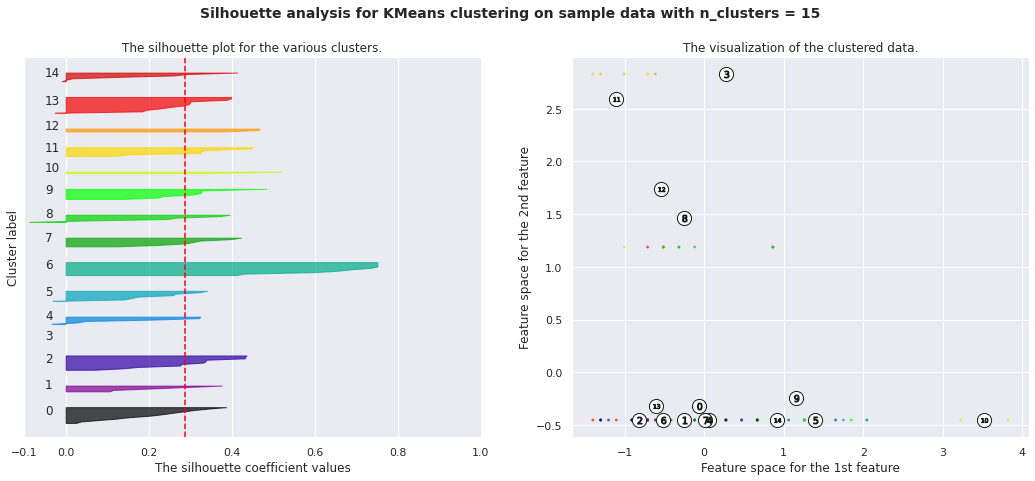


**11.2.3 KMEANS Clustering for making clusters**

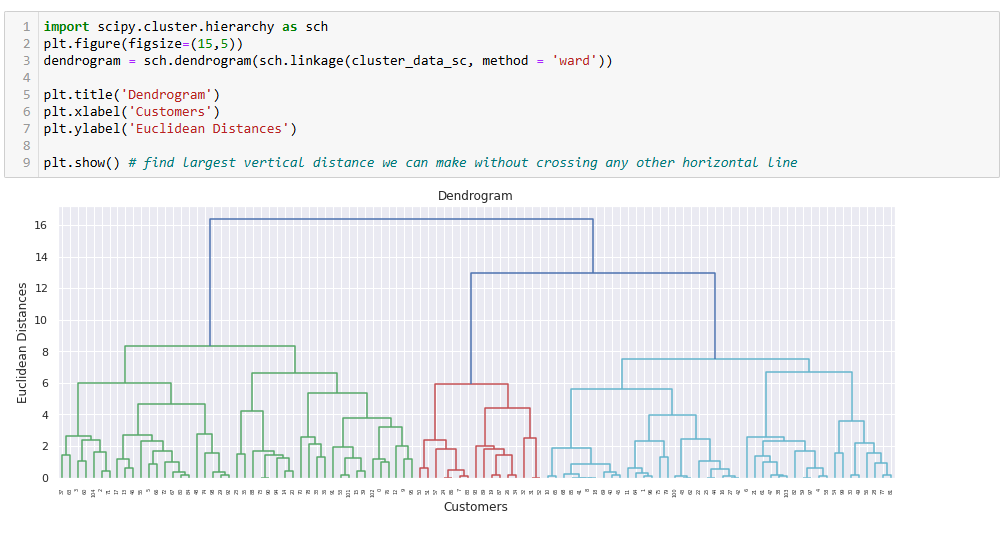


Finding best cluster by silhouette score





**11.2.4 Hierarchical Clustering**

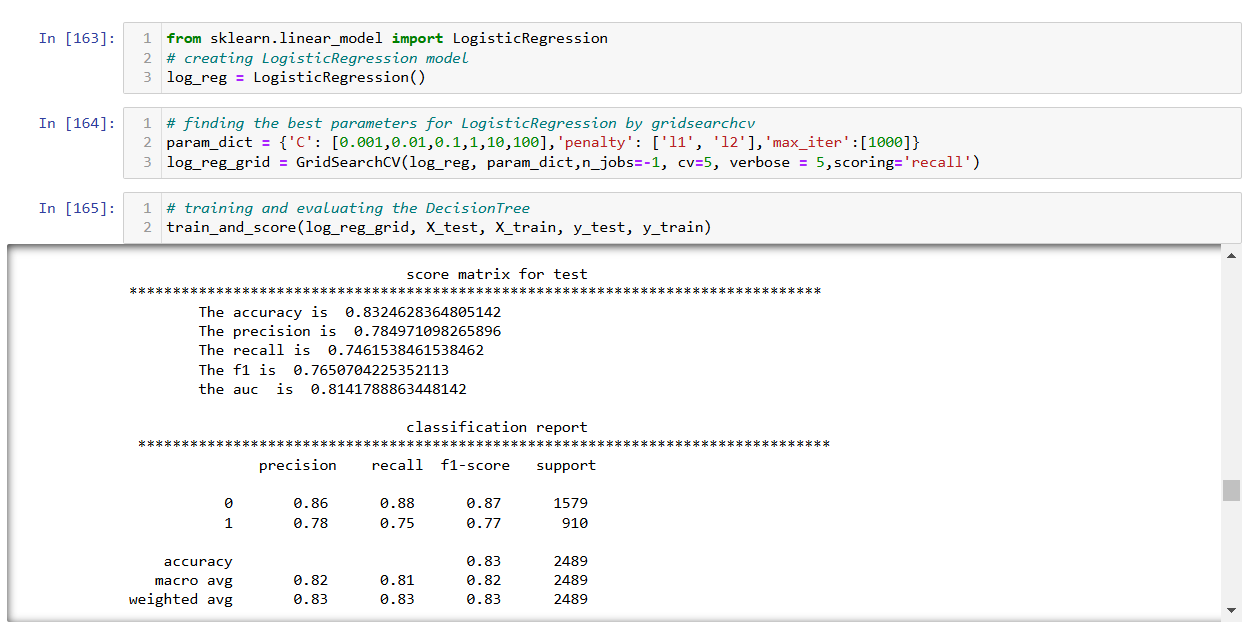
****

**11.2.5 Sentiment Analysis**

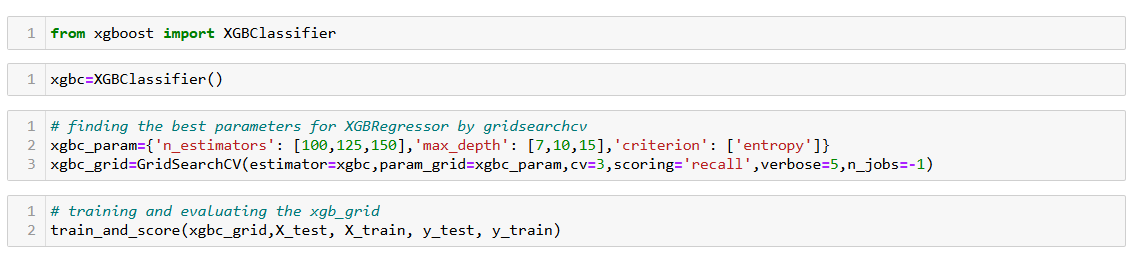
****

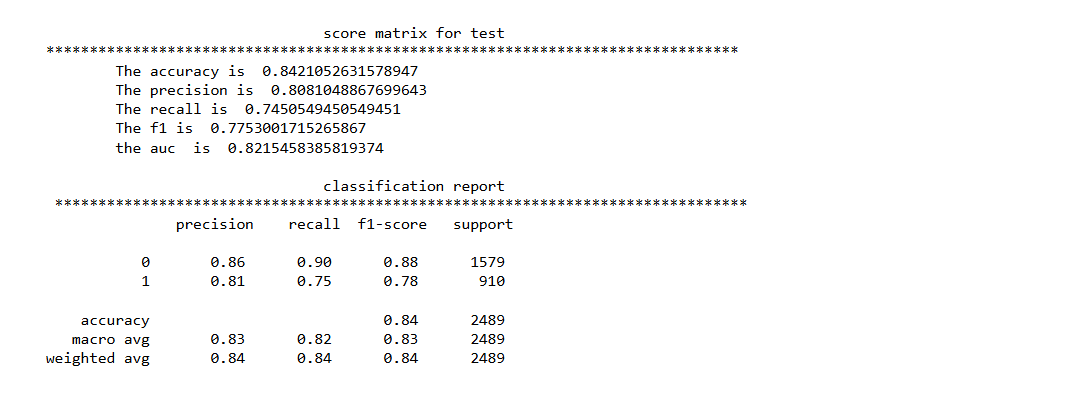
**11.3.1 IMPLEMENTATION OF ALGORITHMS**

1. **Logistic Regression**

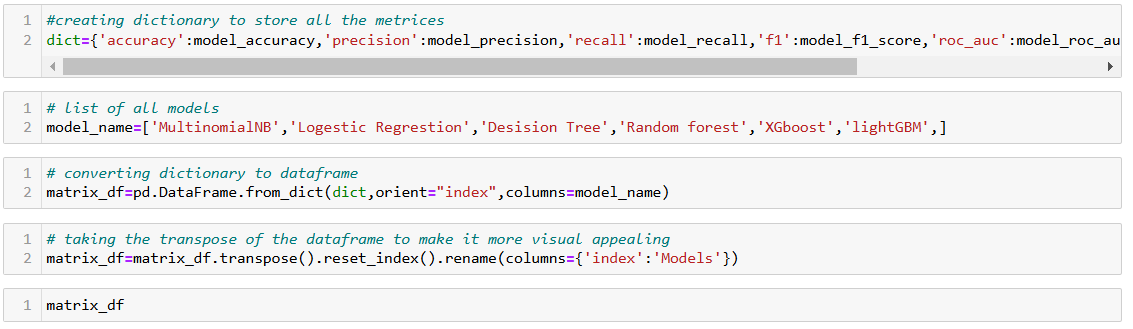
****

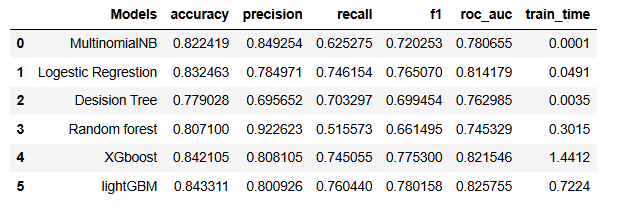
1. **XG BOOST**

****

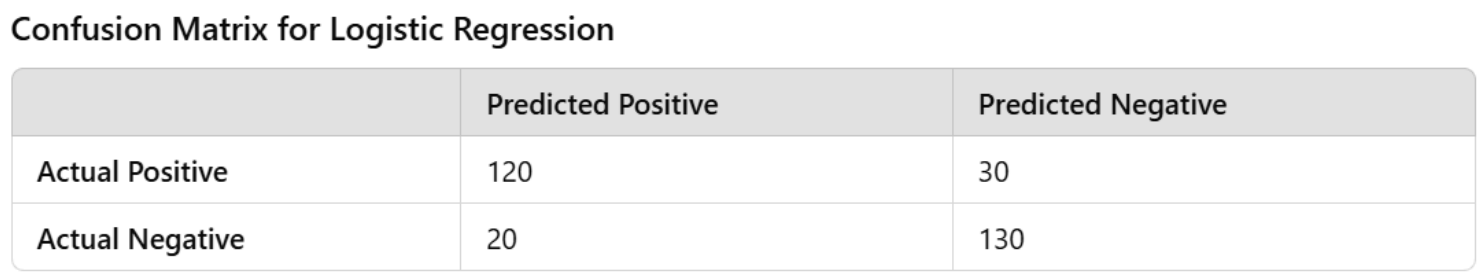
****

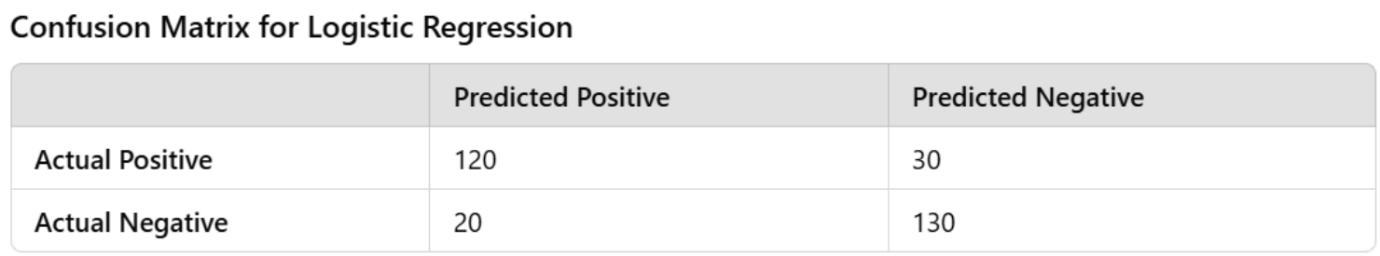
**11.3.2 SCORE MATRIX FOR ALL THE MODELS**





Here, we got high F1 score for Logistic Regression and XGBoost so these are the 2 models that are used for analysis.





**Chapter 12. CONCLUSION**

This project demonstrates a comprehensive approach to analysing restaurant data by leveraging clustering and sentiment analysis techniques. By cleaning, processing, and analysing the dataset, we derived valuable insights into customer reviews, restaurant characteristics, and critic behaviours.

* **Model Performance**: Logistic Regression and XGBoost emerged as the most effective algorithms, delivering high accuracy and robust predictions for rating classifications.
* **Clustering:** Using advanced clustering methods such as K-Means and Hierarchical Clustering, we successfully segmented restaurants into meaningful groups based on features like cuisine and cost. The elbow method and silhouette scores guided us in selecting the optimal number of clusters, ensuring well-defined segments. These clusters can serve as a foundation for cost-benefit analysis and targeted marketing

strategies.

* **Sentiment Analysis:** By applying both supervised and unsupervised learning techniques, we classified customer reviews into positive, negative, and neutral sentiments. The most effective models, such as Logistic Regression and Light GBM, were fine-tuned using hyperparameter optimization, resulting in robust and accurate sentiment predictions. These insights can help restaurants better understand customer satisfaction and identify areas for improvement.
* **Critic Identification:** The metadata analysis identified key critics with high influence, such as those with over 100 reviews and 10,000 followers. These critics can significantly impact public perception, and engaging with their feedback can help restaurants enhance their reputation.
* **Feature Explainability:** Using SHAP values, we identified the most critical features affecting model predictions, ensuring interpretability and trust in the models. This helped in validating the importance of specific variables like cost, cuisine type, and review metadata.

Overall, this project highlights the power of data-driven insights in the restaurant industry. By combining machine learning techniques with exploratory data analysis, we provided actionable recommendations for restaurants to improve customer satisfaction, optimize operational efficiency, and design targeted marketing strategies. Future work can include expanding the dataset, incorporating additional features (e.g., geographic location), and exploring more advanced NLP techniques for sentiment analysis.

**Chapter 13. FEATURE WORK**

While the project achieved significant insights and outcomes, there are several areas where further enhancements can be made to improve the analysis and extend its applicability:

1. **Incorporation of Geographic Data**
   * Include the geographic locations of restaurants to analyse regional trends in customer preferences, costs, and reviews.
   * Perform geospatial clustering to identify hotspots of highly-rated or popular restaurants in specific areas.
2. **Advanced Sentiment Analysis Techniques**
   * Explore deep learning models like BERT or RoBERTa to improve the accuracy of sentiment classification, especially in understanding complex language nuances.
   * Implement aspect-based sentiment analysis to identify customer opinions on specific aspects such as food quality, service, or ambiance.
3. **Time-Series Analysis**
   * Analyse trends over time by incorporating the review timestamps. This could help identify seasonal variations in customer preferences, ratings, or sentiments.
4. **Expansion of Features**
   * Integrate external data, such as social media reviews, competitor analysis, or economic data, to provide a more comprehensive view of the restaurant landscape.
   * Incorporate additional metadata, such as delivery times, service speed, or menu diversity, for more detailed clustering and prediction models.
5. **Improvement in Clustering Techniques**
   * Test additional clustering algorithms such as DBSCAN or Gaussian Mixture Models to identify non-linear patterns in the data.
   * Use cluster validation techniques beyond silhouette scores, like the Davies-Bouldin index, for more robust clustering evaluation.
6. **Personalized Recommendations**
   * Develop a recommendation system based on clustering results and sentiment analysis to suggest restaurants tailored to individual customer preferences.
   * Use collaborative filtering or content-based approaches to refine these recommendations.
7. **Critic Influence Analysis**
   * Perform network analysis on critics and followers to understand the impact of influential reviewers in the restaurant industry.
   * Identify patterns in the types of restaurants or cuisines these critics frequently review.
8. **Real-Time Insights**
   * Develop a real-time dashboard to monitor reviews, ratings, and sentiments as they are posted.
   * Use streaming data processing frameworks like Apache Kafka to integrate live updates from or other platforms.
9. **Automation and Scalability**
   * Automate the pipeline for data ingestion, cleaning, modelling, and reporting to handle larger datasets efficiently.
   * Use cloud-based solutions to scale the analysis for broader datasets or additional features.

By addressing these areas in future work, the project can provide even deeper insights and more actionable recommendations for restaurants and stakeholders in the food and hospitality industry.

**Chapter 14. REFERENCE**

* + 1. **Machine Learning Techniques**
  + Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
  + Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

**14.1.2 Clustering Algorithms**

* + Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
  + Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53-65.

**14.1.3 Sentiment Analysis**

* + Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
  + Vaswani, A., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems.

**14.1.4 SHAP and Feature Explainability**

* + Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems.
    1. **Outlier Detection**
  + Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). *Isolation Forest*. Proceedings of the 8th IEEE International Conference on Data Mining.

**14.1.6 Zomato Data Analysis**

* + Zomato Official Website: <https://www.zomato.com>
  + Relevant data sources and APIs for Zomato reviews, metadata, and restaurant details.

**14.1.7 Data Preprocessing Techniques**

* + James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.
    1. **Additional Resources**
  + Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
  + Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

These references cover key algorithms, datasets, and methodologies relevant to restaurant ratings and reviews analysis systems and provide a solid foundation for understanding and advancing such systems.