PROJECT REPORT

ZOMATO RESTAURANT CLUSTERING

USING UNSUPERVISED MACHINE LEARNING

SUBMITTED BY:
HARSHITA

1.Problem Statement

There is a lot of data about prices, customer ratings, reviews, and food preferences in the restaurant business. But it's hard to find meaningful patterns and customer groups by looking at this data by hand. The goal of this project is to put restaurants into meaningful groups based on their traits without using pre-defined labels. The goal of the project is to use unsupervised machine learning methods to find restaurants that are similar, find hidden patterns, and give insights that can help businesses make better decisions, like targeted marketing, pricing optimization, and strategies for getting customers to interact with them.

2.Objectives

- Perform comprehensive exploratory data analysis on Zomato restaurant data to understand distributions and relationships.
- Preprocess and clean data including handling missing values, duplicates, and inconsistent formats.
- Engineer meaningful features from raw data including cost categories, log transformations, and value-for-money metrics.
- Apply natural language processing techniques to extract insights from customer reviews using TF-IDF vectorization.
- Implement and compare multiple clustering algorithms: K-Means, DBSCAN, and Agglomerative Clustering.
- Evaluate clustering performance using Silhouette Score and select the optimal model.
- Interpret cluster characteristics using centroid analysis to identify restaurant segments.
- Prepare the final model for deployment in real-world applications.

3.System Architecture

The project follows a structured machine learning pipeline architecture:

3.1 Data Layer

- **Restaurant Metadata Dataset:** Contains restaurant name, cost for two, cuisines, aggregate ratings, collections, and operating timings.
- **Restaurant Reviews Dataset:** Contains customer reviews, review text, ratings, and timestamps.

- Both datasets are merged on restaurant identifiers to create a unified dataset for analysis.

3.2 Processing Layer

- **Data Cleaning Module:** Handles missing values using mean/median imputation, removes duplicates, standardizes formats.
- **Feature Engineering Module:** Creates derived features including cost categories, log-transformed review counts, and value-for-money ratios.
- **NLP Pipeline:** Processes review text through lowercasing, punctuation removal, stopword filtering, tokenization, lemmatization, and TF-IDF vectorization.
- **Encoding Module:** Applies Label Encoding for ordinal features and One-Hot Encoding for categorical features like cuisines.

3.3 Model Layer

- **Scaling:** StandardScaler normalizes features to have zero mean and unit variance.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) reduces feature space while retaining variance.
- **Clustering Algorithms:** K-Means, DBSCAN, and Agglomerative Clustering models trained and evaluated.
- **Evaluation:** Silhouette Score computation for cluster quality assessment.

3.4 Output Layer

- **Cluster Labels:** Each restaurant assigned to a cluster group.
- **Cluster Profiles:** Centroid analysis revealing characteristics of each segment.
- **Visualization:** PCA-based 2D/3D plots showing cluster separation.
- **Model Serialization:** Saved model files using joblib for deployment.

4.Key Components

4.1 Data Preprocessing Pipeline

- **Missing Value Handler:** Implements statistical imputation strategies based on data distribution and feature type.
- **Duplicate Remover:** Identifies and removes exact and near-duplicate records based on key identifiers.
- **Format Standardizer:** Ensures consistent data types, date formats, and categorical value representations.

4.2 Feature Engineering Engine

- **Cost Categorization:** Bins continuous cost values into categorical ranges (Budget, Mid-range, Premium).
- **Skewness Reduction:** Applies log transformation to highly skewed features like review counts.
- **Ratio Features:** Creates value-for-money metrics by combining rating and cost information.
- **Temporal Features:** Extracts time-based patterns from operating hours and review timestamps.

4.3 NLP Processing Module

- **Text Cleaner:** Removes URLs, special characters, digits, and extra whitespace from reviews.
- **Tokenizer:** Breaks text into individual words and phrases for analysis.
- **Lemmatizer:** Reduces words to their base forms for standardization.
- **TF-IDF Vectorizer:** Converts text to numerical vectors capturing word importance.

4.4 Clustering Models

- **K-Means:** Partition-based clustering with configurable K value determined by Elbow Method.

- **DBSCAN:** Density-based clustering with epsilon and minimum points parameters for noise detection.
- **Agglomerative:** Hierarchical clustering with various linkage methods for tree-based grouping.

5.Data Flow

1. Raw datasets are loaded from CSV files into pandas DataFrames.
2. Initial data quality assessment identifies missing values, duplicates, and data types.
3. Datasets are merged on restaurant ID to combine metadata and reviews.
4. Missing values are imputed using appropriate strategies (mean for numerical, mode for categorical).
5. Duplicate records are identified and removed based on primary keys.
6. Feature engineering creates new variables: cost categories, log review counts, value-for-money ratios.
7. Customer reviews undergo NLP preprocessing: cleaning, tokenization, lemmatization.
8. TF-IDF vectorization converts processed text into numerical features.
9. Categorical variables are encoded using Label Encoding and One-Hot Encoding.
10. All numerical features are scaled using StandardScaler for normalization.
11. PCA is applied for dimensionality reduction and visualization preparation.
12. Multiple clustering algorithms are trained on the processed feature set.
13. Elbow Method determines optimal K value for K-Means clustering.
14. Silhouette Scores are computed for all models to evaluate cluster quality.
15. Best performing model (K-Means) is selected based on evaluation metrics.
16. Cluster centroids are analyzed to interpret the characteristics of each group.
17. Final model is serialized using joblib for deployment readiness.
18. Model is tested on unseen data to validate generalization capability.

6.Technologies Used

Layer/Component	**Technologies/Libraries**
Data Manipulation	Python 3.8+, Pandas, NumPy
Visualization	Matplotlib, Seaborn, Plotly
Machine Learning	Scikit-learn (sklearn)
Natural Language Processing	NLTK, Scikit-learn TfidfVectorizer
Preprocessing	StandardScaler, LabelEncoder, OneHotEncoder
Clustering Algorithms	KMeans, DBSCAN, AgglomerativeClustering
Dimensionality Reduction	PCA (Principal Component Analysis)
Model Evaluation	Silhouette Score, Elbow Method
Model Persistence	Joblib

7.Results and Evaluation

Metric/Quality	**Outcome**
Best Performing Model	K-Means Clustering
Optimal Number of Clusters	Determined by Elbow Method
Silhouette Score (K-Means)	Highest among all models
Cluster Interpretability	Clear distinction between Premium, Mid-range, and Budget segments
Feature Importance	Cost, Rating, Review Count, Cuisines identified as key discriminators
Model Generalization	Successfully tested on unseen restaurant data
Deployment Readiness	Model saved and validated for production use

8.Usage Workflow

1. Load raw Zomato restaurant metadata and reviews datasets.
2. Run initial data quality checks and exploratory analysis.
3. Execute data cleaning pipeline to handle missing values and duplicates.
4. Apply feature engineering to create derived variables.
5. Process customer reviews through NLP pipeline.
6. Encode categorical variables and scale numerical features.
7. Apply PCA for dimensionality reduction if needed.
8. Train multiple clustering models (K-Means, DBSCAN, Agglomerative).
9. Use Elbow Method to determine optimal K for K-Means.
10. Evaluate all models using Silhouette Score.
11. Select best model based on performance metrics.
12. Analyze cluster centroids to interpret segment characteristics.
13. Visualize clusters using PCA-reduced dimensions.
14. Save final model using joblib for deployment.
15. Test model on new restaurant data for validation.
16. Generate cluster assignment reports for business use.

9.Conclusion

This project successfully demonstrates a comprehensive end-to-end application of Unsupervised Machine Learning for Zomato restaurant clustering. By integrating data preprocessing, feature engineering, natural language processing, and multiple clustering algorithms, the solution effectively discovers hidden patterns in complex restaurant data and groups similar establishments together.

The K-Means clustering model emerged as the optimal choice, providing stable clusters with high Silhouette Scores and clear interpretability. Through centroid analysis, distinct restaurant segments were identified based on pricing, ratings, review popularity, and cuisine preferences, offering valuable insights for business intelligence and strategic decision-making.

The project architecture is modular, scalable, and deployment-ready, with the final model serialized for integration into production systems. The clustering solution can support various business applications including targeted marketing campaigns, personalized recommendation systems, competitive analysis, and strategic planning initiatives.

While the current implementation delivers strong results, several opportunities exist for enhancement including real-time data integration, advanced NLP techniques using transformer models, geographical and temporal feature incorporation, and development of interactive visualization dashboards. These improvements would further increase the solution's value and applicability in real-world scenarios.

Overall, this project validates the effectiveness of unsupervised learning techniques in extracting actionable insights from large-scale restaurant data, making it suitable for both academic evaluation and practical business applications in the food service and restaurant technology industry.