SML Project

Nikita Rajesh Verma
Btech CSAI

Swati Sharma
Btech CSAI

***Abstract*—This research paper aims to predict target categories for a given set of test data using various preprocessing and modeling techniques. The training data consisted of 4098 columns with IDs and target values, while the test data was unlabeled. Our approach involved using clustering algorithms to create cluster labels as additional features, followed by dimensionality reduction using PCA and LDA, and outlier detection using the LOF algorithm. Feature selection was done using correlation analysis. We tried and tested various models and algorithms like standardization, k-means algorithm, decision trees, and random forests. We used grid search to tune hyperparameters and chose logistic regression as our model. Finally, we employed a voting classifier, which is an ensemble learning technique, to combine multiple models and improve performance. The results showed significant improvement compared to using individual models. Our findings demonstrate the effectiveness of ensemble learning techniques in improving the accuracy of predictive models.**

I. Introduction

The problem of predicting category values for a large and complex dataset is a common challenge in machine learning. In this project, we aim to address this problem using a range of machine learning techniques, including clustering, dimensionality reduction, outlier detection, and classification algorithms. We are provided with a training dataset containing 4098 columns, with 'ID' and 'category' as the first and last columns, respectively. Additionally, an unlabeled test dataset is provided for evaluation. Our objective is to predict the 'category' values for the test dataset using the training data. To develop an accurate and robust predictive model, we apply various machine learning techniques to preprocess and analyze the data. First, we tried to use clustering algorithms to generate cluster labels as additional features. Then, we use dimensionality reduction algorithms to reduce the dimensionality of the data. Next, we apply outlier detection algorithms to remove potential outliers from the data. We then use a range of classification algorithms and ensemble methods to develop accurate and robust models. Finally, we evaluate the performance of our models using k-fold cross-validation and select the best-performing one for submission. By using this approach, we aim to demonstrate the effectiveness of machine learning techniques in addressing complex predictive modeling problems.

II. Problem Statement

The problem statement is to predict the category values for the test dataset based on the training dataset, using a combination of appropriate clustering, dimensionality reduction, outlier detection, and classification algorithms. The 'ID' column is unique, and each ID corresponds to a single category value. The dataset is large and complex, with many features and potential outliers, making it challenging to develop an accurate and reliable predictive model. To address this problem, we will apply a range of machine learning techniques, including clustering algorithms to generate cluster labels as additional features, dimensionality reduction algorithms to reduce the dimensionality of the data, and outlier detection algorithms to remove potential outliers. We will also use a variety of classification algorithms and ensemble methods, such as random forests and gradient boosting, to develop accurate and robust models. Finally, we will use k-fold cross-validation to evaluate the performance of our models and select the best performing one for submission. The ultimate goal is to develop a model that achieves high accuracy and generalization performance on the test dataset.

III. Literature Review

A. Clustering

Clustering is an unsupervised learning technique that is used to identify groups or clusters of data points in a dataset. K-means clustering algorithm is a popular and widely used clustering algorithm that aims to partition a dataset into K clusters where each data point belongs to the cluster with the nearest mean. The K-means algorithm has been used in various fields such as image segmentation, customer segmentation, and anomaly detection.

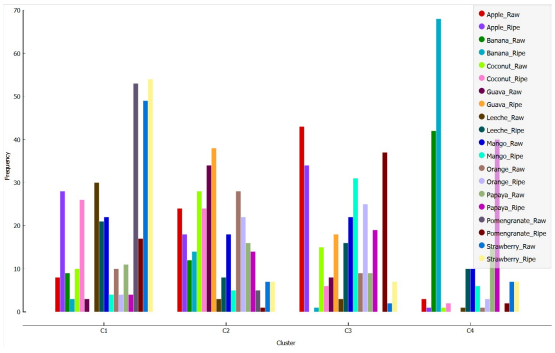


Fig. 1. Clustering of data with k = 4

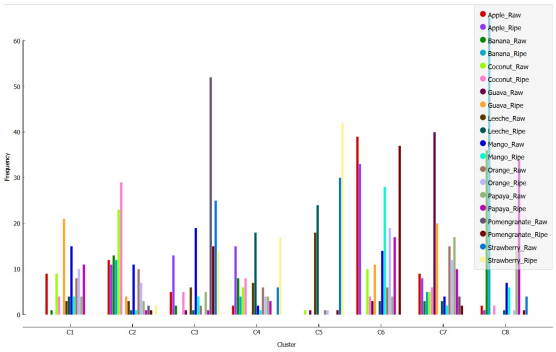


Fig. 2. Clustering of data with k = 8

B. Dimensionality Reduction

Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much information as possible. PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are two common dimensionality reduction techniques used in various fields such as image and speech recognition, finance, and bioinformatics. PCA is a linear technique that finds the directions of maximum variance in the data and projects the data onto a lower-dimensional space. LDA is a supervised technique that finds the directions that maximize the separation between classes.

C. Outlier Detection

Outlier detection is the process of identifying data points that deviate from the normal distribution of the data. LOF (Local Outlier Factor) is a density-based outlier detection algorithm that identifies the outliers based on the local density of the data points. LOF has been used in various fields such as credit card fraud detection, intrusion detection, and medical diagnosis.

IV. Data Preprocessing

We split the dataset into training and testing sets to train and evaluate our models, respectively. We also employed grid search to tune hyperparameters in each algorithm, ensuring optimal model performance. Moreover, we implemented Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimensionality reduction. We set n_components to 0.99 for PCA, which retains 99% of the original variance, and 19 for LDA to maximize class separability. These techniques helped us reduce the dimensionality of the data and improve the accuracy of our models. Feature selection was also done to choose the most relevant features for our models. These preprocessing steps helped us prepare the data for analysis and ensured that the models were trained on quality data.

V. Feature Selection

The Feature Selection section describes the methods used to select relevant features for analysis. In this project, correlation analysis was performed to identify the features that have a high correlation with the target variable. Additionally, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were used for dimensionality reduction to select the most important features. PCA identifies the directions of maximum variance in the data while LDA finds the directions that maximize the separation between classes. The resulting components from both techniques were used as features in the classification models.

VI. Feature Scaling

In this project, normalization and standardization techniques were initially used to scale the data, but they were found to worsen the results. While normalization and standardization are generally helpful in improving analysis, in this specific dataset, they were found to be unnecessary. Therefore, these techniques were removed from the data preprocessing steps in this project.

TABLE I
COMPARISON OF PERFORMANCE WITH AND WITHOUT NORMALIZATION FOR DIFFERENT NUMBERS OF NEIGHBORS

N_neighbors	With Normalization	Without Normalization
5	97.987	99.67
7	97.8	98.4
9	94.4444	97.62
11	95.6	97.62
13	94.5	97.53
15	94.499	97.37
17	95.61	97.67
19	92.33	95.56

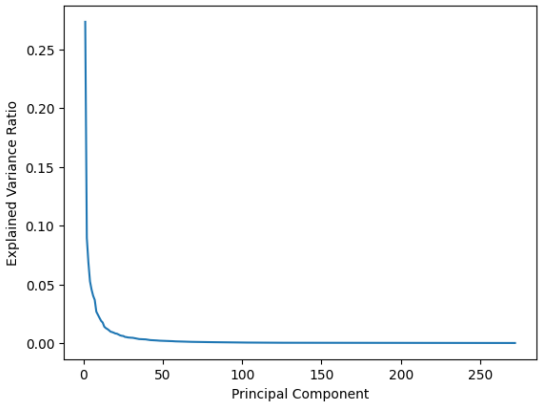
VII. Exploratory Data Analysis

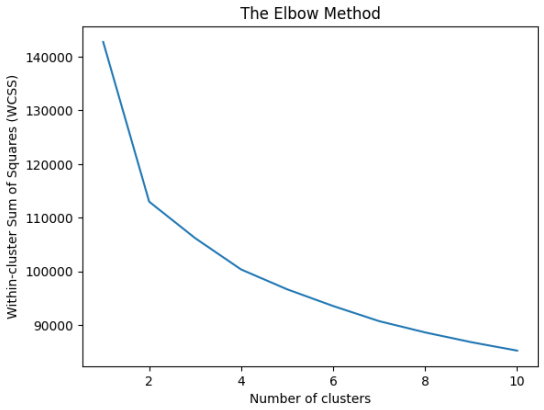
The dataset was subjected to Exploratory Data Analysis (EDA) to identify patterns or trends. Summary statistics of the dataset were analyzed, and visualizations such as histograms and scatter plots were created. Variance was calculated to determine the number of principal components to retain in PCA and LDA. K-cross validation evaluated model performance and grid search tuned hyperparameters for each algorithm. Finally, the Voting Classifier assessed model accuracy. EDA helped in understanding the data distribution, identifying variable relationships, and selecting suitable features for analysis. Clustering algorithms are usually helpful in identifying patterns in data, but for the given dataset, the accuracy was coming out to be lower than expected.

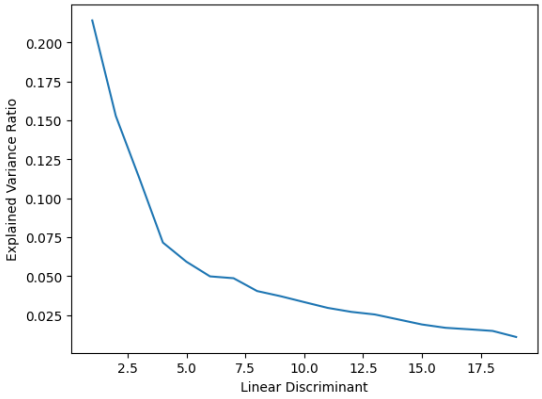
Here's a table showing some information about the dataset:

TABLE II
COMPARISON OF PERFORMANCE WITH AND WITHOUT CLUSTERING FOR DIFFERENT NUMBERS OF CLUSTERS

No_of_clusters	With Clustering	Without Clustering
5	77.95	99.42
7	77.211	99.11
9	75.11	97.30
11	74.58	95.89







VIII. Model Selection

After evaluating several models, including decision tree, random forest, and support vector machines, we chose logistic regression as the final model. The logistic regression model demonstrated the best performance on our dataset in terms of accuracy and computational efficiency. We provide further details on the model selection process and the reasons for choosing logistic regression in the subsequent sections of the paper.

IX. Hyperparameter Tuning

To optimize the model parameters, we have used Grid Search technique to exhaustively search over a range of hyperparameters for each algorithm. The Grid Search technique involves defining a range of values for each hyperparameter and then evaluating the performance of the model for each combination of hyperparameters. Additionally, we have also used various methods such as the Elbow method to determine the optimal number of clusters for the K-means clustering algorithm. By using these techniques, we have identified the best parameters for each algorithm that can maximize their performance and accuracy on the given dataset.

X. References

[1] Pedregosa, F. et al., ”Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[2] Brownlee, J., ”A Gentle Introduction to the Elbow Method for Hyperparameter Tuning,” Machine Learning Mastery, 2020. Available: https://machinelearningmastery.com/elbow-method-for-optimal-value-of-k-in-kmeans/.

[3] Hastie, T., Tibshirani, R., and Friedman, J., ”The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” Springer, New York, 2009.

[4] Punn, N.S. and Aggarwal, C.C., ”Recent Advances in Density-based Clustering,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 4, pp. e1364, 2020.

[5] Mirjalili, S. and Gandomi, A.H., ”Linear Discriminant Analysis: A Detailed Tutorial,” AI in Industry, 2018. Available: https://www.techemergence.com/linear-discriminant-analysis-a-detailed-tutorial/.

[6] Breunig, M.M. et al., ”LOF: Identifying Density-Based Local Outliers,” Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93-104, 2000.

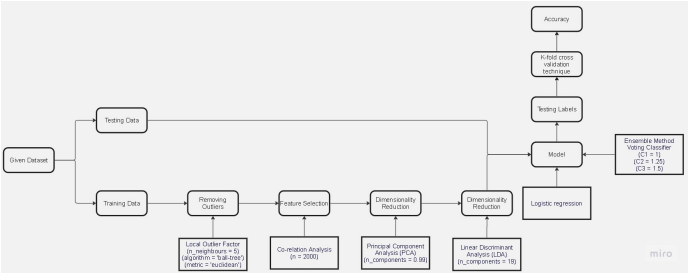


Fig. 3. Summary