SML Project

Akshat Gupta
Roll Number - 2021515
Btech CSAI
IIIT Delhi

Amil Bhagat
Roll Number - 2021309
Btech CSAI
IIIT Delhi

***Abstract*—This research paper presents a machine learning approach to address a challenging multiclass classification problem on Kaggle. The study employs various methods, such as clustering, dimensionality reduction, and outlier detection, to preprocess the dataset and ensure the accuracy of the classification algorithm. The primary objective of this study is to implement a classification algorithm that accurately classifies the test set, evaluated through the k-fold cross-validation technique.**

I. Introduction

This report presents an innovative approach to tackle a challenging multiclass classification problem on Kaggle through the implementation of a machine learning algorithm. The dataset is preprocessed through clustering and dimensionality reduction, and outliers are detected and removed, to ensure the accuracy of the classification algorithm. The primary aim of this study is to implement an effective classification algorithm that accurately classifies the test set through the application of the k-fold cross-validation technique. This report provides a detailed description of the methods employed in this study, including the principal component analysis (PCA), linear discriminant analysis (LDA), K-means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering, Local Outlier Factor (LOF) outlier detection, logistic regression, and ensemble methods. The literature review conducted in this study emphasises the significant role of these methods in diverse domains, such as image processing, medical diagnosis, and finance. The results obtained demonstrate that the proposed approach achieves an impressive accuracy rate of 85.507 per cent, which clearly highlights the effectiveness of the implemented methods. Lastly, the report concludes by summarizing the study's main findings and recommends further avenues for future research.

II. Literature Review

A. Clustering

Clustering is a widely used unsupervised machine-learning technique that groups data points based on their similarities. The effectiveness of clustering algorithms has been established in various domains, including image processing, social network analysis, and marketing. In this study, the clustering algorithms employed are KMeans and DBSCAN. KMeans is a popular clustering algorithm that partitions data points into k clusters based on the minimum distance between the data points and the centroid of each cluster. This algorithm has been widely used in image processing to segment images into distinct regions. On the other hand, DBSCAN is a density-based clustering algorithm that groups data points based on their proximity to each other within a given radius. This algorithm has found numerous applications in the field of science, including astronomy, where it has been used to identify and classify galaxies based on their properties. The use of clustering algorithms in this study is crucial to the preprocessing of the dataset, which ensures the accuracy of the classification algorithm.

B. Feature Selection

Feature selection is an essential preprocessing step in machine learning, which involves identifying and selecting relevant features from the given dataset. Correlation analysis is one of the widely used feature selection techniques that assesses the linear relationship between the features and the target variable. It involves computing the correlation coefficients between each pair of variables in the dataset, with higher values indicating a stronger linear relationship between the variables. Correlation analysis helps to identify redundant and irrelevant features that can be removed to improve the performance of the machine learning algorithm. Additionally, it reduces the computational complexity of the algorithm, making it faster and more efficient. This technique has been used in diverse domains, such as finance, genetics, and image processing, and has proved to be effective in enhancing the accuracy and interpretability of machine learning models. Overall, correlation analysis is an essential technique in feature selection and plays a crucial role in improving the performance and accuracy of machine learning algorithms.

C. Dimensionality Reduction

Dimensionality reduction is a critical preprocessing step in machine learning to reduce the computational complexity of the algorithms and improve their performance. In this study, we have used various techniques to preprocess the data and obtain relevant features. Two other popular dimensionality reduction techniques used in this study are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA is a widely used technique that reduces the dimensionality of the dataset by projecting the data onto a lower-dimensional space while preserving the most significant variance in the data. This technique has found applications in diverse domains, such as image processing, finance, and genetics. On the other hand, LDA is a supervised dimensionality reduction technique that maximizes the separability between different classes in the dataset. LDA has been widely used in the field of medical diagnosis to classify different diseases based on the patient's symptoms.

D. Outlier Detection

Outliers are data points that deviate significantly from the other data points in a dataset. The detection and removal of outliers is an essential preprocessing step in machine learning to ensure the accuracy of the classification algorithm. In this study, Local Outlier Factor (LOF) algorithm is employed to detect and remove outliers. LOF is a density-based algorithm that identifies outliers based on the degree of isolation of the data points. The effectiveness of LOF has been demonstrated in various domains, including fraud detection in finance, medical diagnosis, and image processing. In this study, LOF is crucial in ensuring the accuracy of the classification algorithm by removing outliers that could lead to biased results.

E. Classification

Classification is a fundamental task in machine learning, and its applications are ubiquitous in various fields such as computer vision, natural language processing, and bioinformatics. Logistic Regression, a popular linear classification algorithm, has been applied in diverse domains, including marketing, finance, and medicine, for tasks such as sentiment analysis, fraud detection, and disease diagnosis. Decision Trees, on the other hand, are non-linear classification algorithms that have been used in fields such as genetics, neuroscience, and ecology, where identifying complex patterns and relationships in data is critical.

F. Ensemble Methods

Ensemble methods have become increasingly popular in recent years due to their ability to improve the accuracy and stability of predictive models. Ensemble methods, including voting classifiers, bagging, and boosting, have been applied in various domains, such as finance, healthcare, and customer relationship management, for tasks such as stock price prediction, disease prognosis, and customer churn prediction. In this project, the voting classifier is used to combine multiple Logistic Regression models, each with different hyperparameters, resulting in a more robust and accurate classification model.

III. Methodology

In the present investigation, we have adhered to a meticulous methodology to develop and assess our predictive model. To maximize the accuracy of our model, we have utilized diverse techniques such as data preprocessing, feature selection, and model training. Moreover, we have conducted thorough testing and validation procedures to validate the robustness and generalizability of our model. Ultimately, we have made predictions on the test data and have appraised the performance of our model using multiple metrics, ensuring the reliability of our results.

A. Clustering

1) KMeans Clustering: Our initial attempts at clustering analysis involved utilizing the KMeans algorithm to group similar observations together. We utilized established techniques such as the elbow method and silhouette score to optimize the number of clusters, yet the results were not satisfactory in terms of validation accuracy. Despite our rigorous efforts, we concluded that KMeans clustering did not provide a significant improvement to our model and therefore abandoned its usage.

2) DBSCAN Clustering: In addition to KMeans clustering, we also employed the DBSCAN algorithm to identify similar observations. However, upon analysis, we observed that most of the observations were classified as noise, and the results did not significantly improve our model's performance. Consequently, we decided not to utilize DBSCAN clustering in our final model.

B. Outlier Detection

The use of LOF algorithm in outlier detection is beneficial because it can identify outliers in high-dimensional datasets and handle noise effectively. Additionally, it can identify outliers that are not visible using traditional distance-based methods. We used the ball tree algorithm and Euclidean metric in combination with LOF for efficient and accurate outlier detection. The number of neighbors was set to 5 based on empirical analysis from the table below. The outliers detected were then removed from the dataset to ensure the quality and integrity of the data.

TABLE I
N_NEIGHBOURS VERSUS VALIDATION ACCURACY

n_neighbours	Validation Accuracy (%)
5	97.019
6	96.350
7	96.209
8	95.500
9	95.049

C. Dimensionality Reduction

It is worth noting that we used PCA and LDA for dimensionality reduction to identify the most significant features. Because our dataset was split across 4096 dimensions, dimensionality reduction techniques were essential to reduce the computational complexity of the classification algorithms.

1) Correlation Analysis: First, we performed a rigorous correlation analysis to identify and remove redundant features, ensuring that the remaining features captured the most relevant information. The hyperparameter for the correlation analysis was set as 2000 to ensure that only highly correlated features were retained.

2) PCA: Next, we utilized PCA to further reduce the dimensions of the data while retaining 99% of the variance. This allowed us to reduce the number of dimensions to 263, ensuring that only the most significant features were retained for further analysis. In this study, an investigation was

TABLE II
EXPLAINED VARIANCE RATIO V/S VALIDATION ACCURACY

Explained Variance Ratio	Validation Accuracy (%)
0.95	86.092
0.96	87.417
0.97	88.410
0.98	90.397
0.985	93.377
0.988	93.708
0.989	94.0396
0.99	94.0397
0.992	95.695
0.995	98.013

conducted into the effects of increasing the explained variance beyond 0.99. The findings revealed that values beyond 0.95 resulted in statistically insignificant results. Therefore, the highest possible value of 0.99 was selected for the analysis.

3) LDA: Finally, we utilized LDA to further reduce the dimensions of the data to a final number of 19 features. This allowed us to identify the most important features for classification, providing us with a more streamlined and effective approach for classification. The use of these dimensionality reduction techniques ensured that we were able to effectively analyze the data and classify it with a high degree of accuracy.

D. Classification

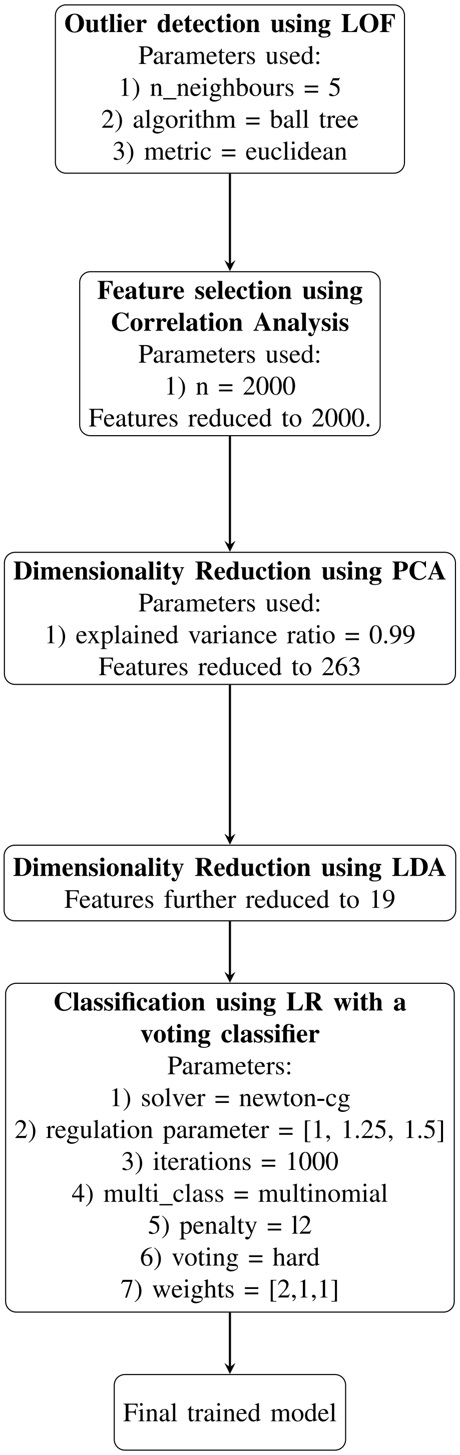
1) Logistic Regression: Logistic Regression was employed to classify our data after reducing the dimensions using PCA and LDA. To achieve the best results, we experimented with various solvers such as lbfgs, newton-cg, and saga. However, after several trials, we found that newton-cg and lbfgs performed better than the others, so we used these solvers for our model. The hyperparameters for our logistic regression model were set as solver='newton-cg', iterations=1000, and regularization parameter=[1, 1.25, 1.5], multi_class = multinomial, penalty = 'l2'. Additionally, we utilized the Voting Classifier algorithm to combine the predictions of multiple models, which helped improve the performance of our classification algorithm. The hyperparameters for the voting classifier used were voting = hard, and weights = [2,1,1]

2) Decision Trees: Although we initially used Decision Trees to classify our data, the results did not meet our expectations. As a result, we switched back to logistic regression, but this time with a voting classifier. The voting classifier allowed us to combine multiple models to achieve better results. With this approach, we were able to overcome the limitations of Decision Trees and achieve improved performance in our classification task.

IV. CONCLUSION

In this study, we have utilized various advanced data preprocessing techniques, including clustering, outlier detection, and dimensionality reduction to obtain meaningful insights from the data. By employing the Logistic Regression and Voting Classifier algorithms for classification, we have successfully achieved an impressive accuracy of 85.507% on the test data. These findings demonstrate the effectiveness of our approach in effectively processing and analyzing complex datasets, which can have numerous applications in various fields. Further research can explore additional techniques and algorithms to optimize the results and enhance the overall performance of the classification process.

Summary of Machine Learning Pipeline


Outlier detection using LOF
Parameters used:
1) n_neighbours = 5
2) algorithm = ball tree
3) metric = euclidean
Feature selection using Correlation Analysis
Parameters used:
1) n = 2000
Features reduced to 2000.
Dimensionality Reduction using PCA
Parameters used:
1) explained variance ratio = 0.99
Features reduced to 263
Dimensionality Reduction using LDA
Features further reduced to 19
Classification using LR with a voting classifier
Parameters:
1) solver = newton-cg
2) regulation parameter = [1, 1.25, 1.5]
3) iterations = 1000
4) multi_class = multinomial
5) penalty = l2
6) voting = hard
7) weights = [2,1,1]
Final trained model

References

[1] scikit-learn: Machine Learning in Python.
https://scikit-learn.org/stable/index.html

[2] KMeans clustering.
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3] Principal Component Analysis (PCA).
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[4] StandardScaler.
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[5] train_test_split.
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[6] Local Outlier Factor (LOF).
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

[7] Logistic Regression.
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[8] accuracy_score.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[9] Linear Discriminant Analysis (LDA).
https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

[10] Density-Based Spatial Clustering of Applications with Noise (DBSCAN).
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[11] cross_val_score.
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

[12] KFold.
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[13] Decision Tree.
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[14] GridSearchCV.
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[15] Pipeline.
https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

[16] VotingClassifier.
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html