Identifying the Quality of Mushrooms Using Machine Learning

Name:	**Sameer Verma**
Registration No./Roll No.:	21236
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

Introduction

There are 7311 training instances and 814 test instances available for the given problem, which is a two-class problem. The classes are 'edible' denoted by 'e' and 'poisonous' denoted by 'p'. Out of 7311 training instances 3787 instances belong to class 'e' and 3524 instances belong to class 'p'. This is a classification problem as the target variable here is discrete i.e. class 'e' and class 'p'. As part of the pre-processing steps, it was observed that the 'veil-type' feature contains only one unique value 'p' across all instances, making it redundant for model training. Consequently, this feature was dropped from the data set. Column named 'stalk-root' has stalk-root missing representation by '?', consequently it was treated as both a missing value and a feature representation value one by one , when treated as missing value it was imputed and replaced by the mode of that column namely 'b', after evaluation keeping '?' as a feature value resulted in better score (in terms of F1 score). To be precise, there was an improvement in reducing the number of False Positives (this can be verified by un-commenting lines 126,146,147,308,309).

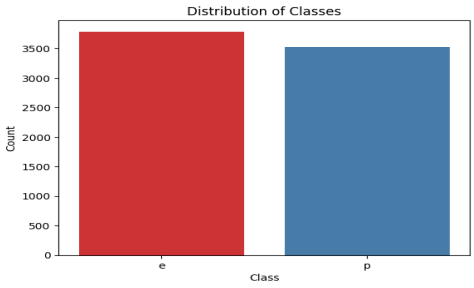

Figure 1: Overview of Data Set

Method

To facilitate the machine learning models' understanding of categorical features, one-hot encoding technique was employed. This process expanded categorical variables into binary columns, each representing a unique category. One hot encoding resulted in increase in number of features. Dimensionality reduction was performed using PCA to mitigate the computational burden associated with a large number of features. The elbow method (explained variance ratio) was employed to determine the optimal number of principal components (n components) for PCA. Various values were experimented with, and the one that yielded the highest model performance was selected. In this case, the elbow

method indicated that a value of 5 for n components provided a good balance between preserving information and reducing dimensionality, leading to enhanced model efficiency. The study incorporated several machine learning algorithms, namely Naive Bayes, Logistic Regression, SVM, KNN, AdaBoost, and RandomForest. These algorithms were implemented through a pipeline. Hyperparameter tuning via GridSearchCV was conducted to optimize model performance, with the F1 score serving as the evaluation metric, focusing on the positive class 'e'. Following the successful execution of the pipeline, the top 3 models based on their F1 scores were considered. These models were then used to create a voting-based classifier with hard voting. Combining the predictions of multiple models through a voting mechanism resulted in a more robust and accurate overall classification performance.To counter the overfitting, the K-fold Cross Validation technique was employed with 10 as the value of K [1].

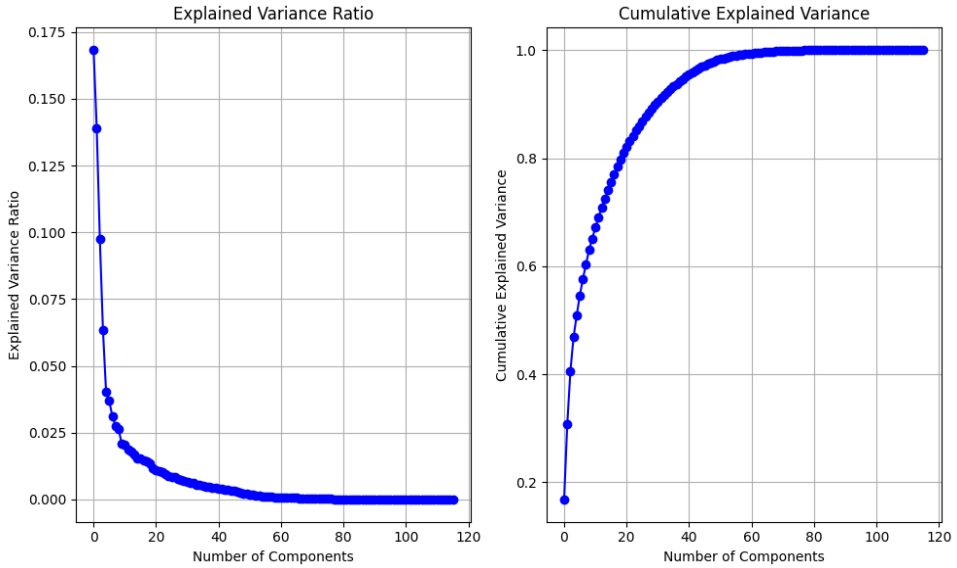

Figure 2: PCA analysis by elbow method.

Evaluation

The evaluation of the top three models yielded insightful results. The Random Forest Classifier [2] exhibited an exception macro averaged F1 score of 1.00, with macro averaged precision and recall scores of 0.99 and 1.00, respectively. KNN performed almost equally well, achieving an impressive macro averaged F1 score of 0.99. It demonstrated a balanced precision-recall trade-off, with macro averaged precision and recall values both exceeding 0.99. The Support Vector Machine (SVM) model showcased exceptional performance, achieving a perfect macro averaged F1 score of 1.00. The SVM model displayed impeccable macro averaged precision of 1.00 and recall of 0.99, indicating flawless classification across both classes. Following the execution of the individual models, a voting-based classifier was crafted by aggregating predictions from Random Forest, KNN, and SVM through hard voting. This ensemble approach harnessed the collective strengths of these models. The resulting ensemble classifier not only maintained but surpassed the individual models, attaining a perfect macro averaged F1 score of 1.00. Precision (1.00)and recall(0.99) values for both classes were consistently elevated, underscoring the potency of combining diverse modeling strategies. The voting classifier, a synergistic amalgamation of Random Forest, KNN, and SVM, stands out as a robust solution, exemplifying the prowess of ensemble learning in enhancing predictive accuracy. The main motive behind this amalgamation was to reduce the number of False Positives, in which it succeeded, and was backed by an impressive confussion matrix which showed only 1 False Postive out of 1463 instances. ROC-AUC curve were also to visualize the performance and the ensemble classifier showed an AUC of 0.9997 [3].

Table 1: Performance Of Different Classifiers (macro averaged)

Classifier	Precision	Recall	F-measure
Adaptive Boosting	0.99	0.99	0.99
K-Nearest Neighbor	0.99	0.99	0.99
Naive Bayes	0.91	0.90	0.90
Logistic Regression	0.95	0.95	0.95
Random Forest	0.99	1.00	1.00
Support Vector Machine	1.00	0.99	1.00
Ensemble classifier	1.00	1.00	1.00

Table 2: Confusion Matrices of Different Classifiers

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	740	7
Poisonous	2	714

Adaptive Boosting

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	741	6
Poisonous	0	716

SVM

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	728	19
Poisonous	132	584

Naive Bayes

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	743	4
Poisonous	0	716

K-Nearest Neighbor

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	716	31
Poisonous	48	668

Logistic Regression

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	744	3
Poisonous	1	715

Random Forest

Actual Class	Predicted Class	
	Edible	Poisonous
Edible	744	3
Poisonous	0	716

Ensemble Classifier

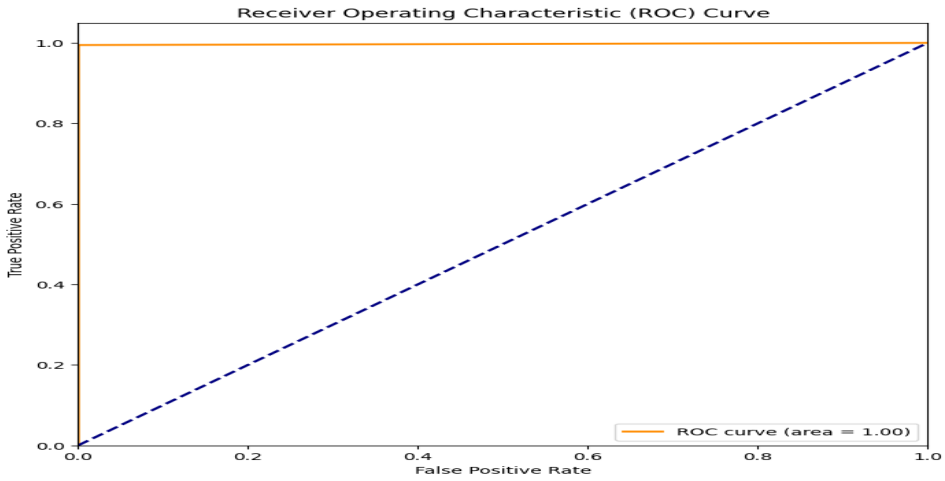

Figure 3: ROC-AUC for Ensemble Classifier

Discussion

In machine learning, the choice between Principal Component Analysis (PCA) and feature selection is often influenced by dataset characteristics. While feature selection is conventionally favored for sparse data, our experimentation on the mushroom classification problem revealed a surprising efficacy of PCA. Despite not being the typical choice for sparse datasets, PCA outperformed traditional feature selection methods. The optimal number of components was determined using the elbow method, and a value of 5 provided the best performance. This unexpected success of PCA in our scenario could be attributed to specific dataset characteristics, such as a moderate number of feature values. This emphasizes the importance of empirical validation and tailored approaches to accommodate dataset-specific nuances. The final ensemble model, created through a voting classifier integrating Random Forest, KNN, and SVM, demonstrated a substantial improvement in overall performance. This collective intelligence highlighted the synergy achieved when combining diverse modeling strategies. Despite its success, the framework has limitations. The reliance on empirical validation implies that its performance might be sensitive to dataset-specific characteristics. Furthermore, the interpretability of the ensemble model may be compromised due to the complexity introduced by combining multiple algorithms. The trade-off between interpretability and performance should be considered based on the application's requirements. There are several avenues for future exploration. Firstly, a more in-depth analysis of the dataset characteristics and their influence on the choice between PCA and feature selection methods could provide valuable insights. Additionally, exploring alternative ensemble strategies and incorporating advanced feature selection techniques might further enhance model performance. The framework's adaptability suggests potential applicability to diverse datasets, warranting exploration in different domains and problem contexts. Lastly, investigations into model explainability methods could address the interpretability concerns associated with ensemble models, ensuring transparency in decision-making processes.

References

[1] R. Kohavi. A study of cross validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, pages 1137-1145, 1995.

[2] R. A. Olshen L. Breiman, J. H. Friedman and C. J. Stone. Classification and regression trees. *Metrika*, 33:128–128, 1986.

[3] A.P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7) pages 1137-1145, 1997.

Github Link : https://github.com/SameerVermaDSE/MLProject