International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 5, July-August 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Comparative Study on Mushroom Classification using Supervised Machine Learning Algorithms

Kanchi Tank

Department of Information Technology, Bharati Vidyapeeth College of Engineering, University of Mumbai, Navi Mumbai, Maharashtra, India

ABSTRACT

Mushroom hunting is gaining popularity as a leisure activity for the last couple of years. Modern studies suggest that some mushrooms can be useful to treat anemia, improve body immunity, fight diabetes and a few are even effective to treat cancer. But not all the mushrooms prove to be beneficial. Some mushrooms are poisonous as well and consumption of these may result in severe illnesses in humans and can even cause death. This study aims to examine the data and build different supervised machine learning models that will detect if the mushroom is edible or poisonous. Principal Component Analysis (PCA) algorithm is used to select the best features from the dataset. Different classifiers like Logistic Regression, Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes and Random Forest are applied on the dataset of UCI to classify the mushrooms as edible or poisonous. The performance of the algorithms is compared using Receiver Operating Characteristic (ROC) Curve.

KEYWORDS: *Mushroom Classification, Principal Component Analysis, Logistic Regression, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Random Forest*

How to cite this paper: Kanchi Tank "A Comparative Study on Mushroom Classification using Supervised Machine Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-5, August 2021, pp.716-723, URL: www.ijtsrd.com/papers/ijtsrd42441.pdf

IJTSRD42441

INTRODUCTION

Mushrooms being the most sustainably produced foods, not only have good taste but also hold a great nutritional value [8]. They contain proteins, vitamins, minerals, and antioxidants. These can have various health benefits [6]. Consumption of mushrooms helps to fight different types of diseases such as cancer, helps to regulate blood cholesterol levels, and thus helps to fight diabetes. Mushrooms aid in strengthening our immune system and also help us to lose weight. They are a beguiling mixture of lucrative as well as speculative features.

But aside from the healthy mushrooms, there also exists poisonous and wild mushrooms whose consumption may result in severe illnesses in humans and can even cause death. It is not easy for a layman to differentiate wild mushrooms from healthy mushrooms [6]. This study aims to classify mushrooms into edible or poisonous using different supervised learning models on the dataset of UCI that makes available various specifications of mushrooms like cap shape, cap color, gill color, odor, etc.

RELATED WORK

In recent years, many researchers around the globe worked in classification and predictive analytics in various domains. Classification is most useful as it can make predictions about values of data using known results found from the different data [16]. Previous researchers have employed classification techniques in making predictions in various studies. For example, [19] applied six different Machine Learning algorithms namely, Decision Tree, SVM, KNN, Random Forest, Logistic Regression and Naïve Bayes for predicting diabetes in humans. [9] used several machine-learning algorithms like Random Forest, Naïve Bayes, Support Vector Machines SVM, and K-Nearest Neighbors to predict breast cancer among the women. [12] focused on the Data Mining techniques to discover information in student's raw data using different algorithms such as KNN, Naïve Bayes, and Decision Tree. [13] did a study on "Behavioral malware detection using Naïve Bayes classification techniques". The results showed that data mining is more efficient for detecting malware.

Classification of malware behavioral features can be a convenient method in developing a behavioral antivirus. [5] applied seven different algorithms namely Decision Table, Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Neural Networks (Perceptron), JRip and Decision Tree (J48) using Waikato Environment for Knowledge Analysis (WEKA) machine learning tool on the diabetes dataset. The research shows that time taken to build a model and precision/accuracy is a factor on one hand while kappa statistic and Mean Absolute Error (MAE) is another factor on the other hand. Therefore, ML algorithms require precision, accuracy and minimum error to have supervised predictive machine learning.

Furthermore, the results of a survey conducted by [15] identified the models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models as the most popular among the researchers for predicting Cardiovascular Diseases. A study by [7] on "Behavioral features for mushroom classification" - This paper is set to study mushroom behavioral features such as the shape, surface and color of the cap, gill and stalk, as well as the odor, population and habitat of the mushrooms. The Principal Component Analysis (PCA) algorithm is used for selecting the best features for the classification experiment using the Decision Tree (DT) algorithm. The results showed that the Decision tree using the J48 classifier produced 23 leaves and the size of the tree is 28. [10] discusses data mining algorithms specifically ID3, CART, and HoeffdingTree (HT) based on a decision tree. Hoeffding Tree provides better results with the highest accuracy, low time and least error rate when compared with ID3 and CART. A study by [11] focuses on developing a method for the classification of mushrooms using its texture feature, which is based on the machine learning approach. The performance of the proposed approach is 76.6% by using an SVM classifier, which is found better concerning the other classifiers like KNN, Logistic Regression, Linear Discriminant, Decision Tree, and Ensemble classifiers. [14] used the Decision Tree classifier to develop a classification model for edible and poisonous mushrooms. The results of the model's effectiveness evaluation revealed that the model using the Information Gain technique alongside the Random Forest technique provided the most accurate classification outcomes at 94.19%.

The remaining of this paper proceeds as follows. Section III presents the materials and methods applied to achieve the objective of this research. Subsequent sections IV and V present the results and conclusion of the study.

MATERIALS AND METHODS

Data mining is one of the major and important technologies that is currently being used in the industry for performing data analysis and gaining insight into the data. It uses different data mining techniques such as Machine Learning, Artificial Intelligence, and statistical analysis. In this study, machine learning techniques are used for mushroom classification. Machine learning provides a pool of tools and techniques, using these tools and techniques raw data can be converted into some actionable, meaningful information by computers. In this paper, supervised machine learning algorithms are used.

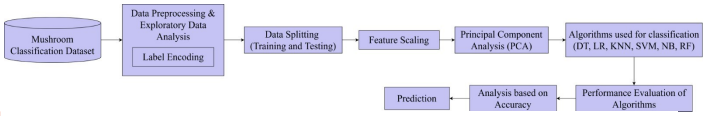

Figure 1 Methodology for Mushroom Classification

A. Dataset and Attributes

This research paper uses an openly available dataset that is downloaded from the UCI machine learning repository. This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one [4].

This dataset contains 22 attributes with 8124 instances of mushrooms. Figure 2 gives the attribute information of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   class                     8124 non-null   object
 1   cap-shape                 8124 non-null   object
 2   cap-surface               8124 non-null   object
 3   cap-color                 8124 non-null   object
 4   bruises                   8124 non-null   object
 5   odor                      8124 non-null   object
 6   gill-attachment           8124 non-null   object
 7   gill-spacing              8124 non-null   object
 8   gill-size                 8124 non-null   object
 9   gill-color                8124 non-null   object
 10  stalk-shape               8124 non-null   object
 11  stalk-root                8124 non-null   object
 12  stalk-surface-above-ring  8124 non-null   object
 13  stalk-surface-below-ring  8124 non-null   object
 14  stalk-color-above-ring    8124 non-null   object
 15  stalk-color-below-ring    8124 non-null   object
 16  veil-type                 8124 non-null   object
 17  veil-color                8124 non-null   object
 18  ring-number               8124 non-null   object
 19  ring-type                 8124 non-null   object
 20  spore-print-color         8124 non-null   object
 21  population                8124 non-null   object
 22  habitat                   8124 non-null   object
dtypes: object(23)
memory usage: 1.4+ MB
```

Figure 2 Attribute Information

B. Data Preprocessing And Exploratory Data Analysis

The dataset contains two classes i.e., edible and poisonous. To check the balance of each, a bar graph is plotted. Since the data is categorical, Label Encoder is used to convert it to ordinal. Label Encoder converts each value in a column to a number [18]. Figure 3 shows the count of each class whereas Figure 4 shows the dataset after label encoding.

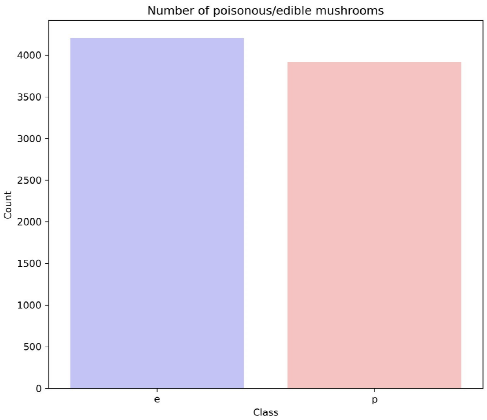

Figure 3 Bar plot to visualize the count of edible and poisonous mushrooms

	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat
0	5	2	4	1	6	1	0	1	4	0	...	2	7	7	0	2	1	4	2	3	5
1	5	2	9	1	0	1	0	0	4	0	...	2	7	7	0	2	1	4	3	2	1
2	0	2	8	1	3	1	0	0	5	0	...	2	7	7	0	2	1	4	3	2	3
3	5	3	8	1	6	1	0	1	5	0	...	2	7	7	0	2	1	4	2	3	5
4	5	2	3	0	5	1	1	0	4	1	...	2	7	7	0	2	1	0	3	0	1

5 rows × 22 columns

Figure 4 Label Encoding

A violin plot is a part of EDA that is used to show the distribution of quantitative data across several levels of one or more categorical variables in such a way that those distributions can be compared. A violin plot is used here to represent the distribution of the classification characteristics.

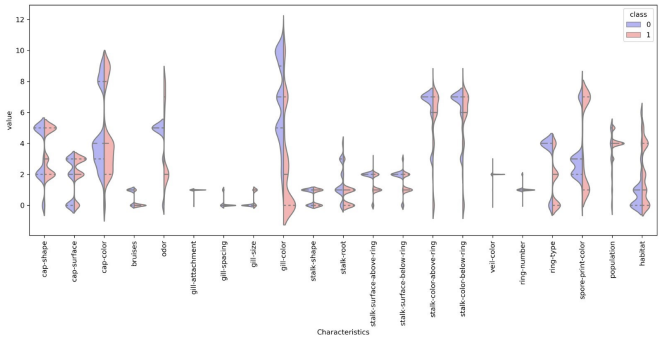

Figure 5 Violin plot representing the distribution of the classification characteristics

Since the dataset contains categorical variables, we apply the get_dummies() method to convert the categorical data into dummy or indicator variables. Figure 6 shows the dummy/indicator variables of the dataset. The conversion of categorical variables into dummy variables leads to the formation of the two-dimensional binary matrix where each column represents a particular category, in our case, 0 is for edible mushroom whereas 1 is for poisonous.

	cap-shape_1	cap-shape_2	cap-shape_3	cap-shape_4	cap-shape_5	cap-surface_1	cap-surface_2	cap-surface_3	cap-color_1	cap-color_2	...	population_2	population_3	population_4	population_5	habitat_1	habitat_2
0	0	0	0	0	1	0	1	0	0	0	...	0	1	0	0	0	0
1	0	0	0	0	1	0	1	0	0	0	...	1	0	0	0	1	0
2	0	0	0	0	0	0	1	0	0	0	...	1	0	0	0	0	0
3	0	0	0	0	1	0	0	1	0	0	...	0	1	0	0	0	0
4	0	0	0	0	1	0	1	0	0	0	...	0	0	0	0	1	0

5 rows × 95 columns

Figure 6 Dummy/indicator variables

Correlation matrices are a requisite tool of exploratory data analysis. It is convenient to understand the relationship among variables/columns. A heatmap is plotted to represent the correlation between the variables.

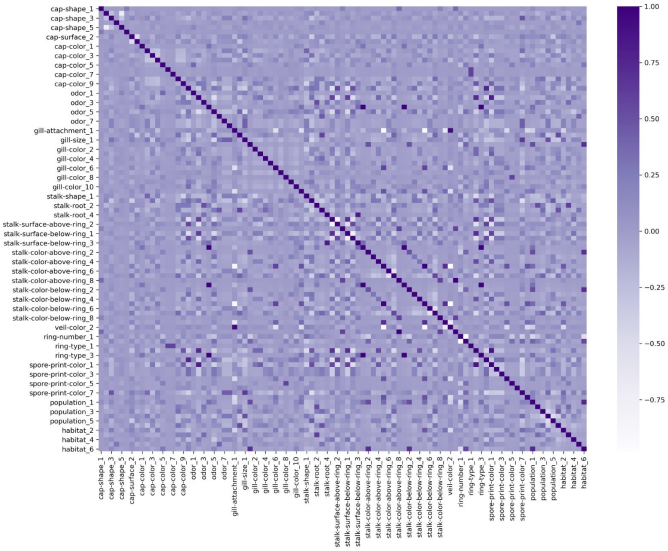

Figure 7 Heatmap representing the correlation between the dummy/indicator variables

C. Data Splitting

Data splitting is a process used to separate a given dataset into at least two subsets called 'training' and 'test'. This step is usually implemented after data preprocessing. Using train_test_split() from the data science library scikit-learn, the data is split into subsets i.e. training and test which contains 70% and 30% data respectively. This minimizes the potential for bias in the evaluation and validation process.

D. Feature Scaling and Principal Component Analysis

Feature Scaling is done to standardize the independent features present in the data in a fixed range. We have used StandardScaler() to perform feature scaling. It performs the task of Standardization [1].

$$x_{new} = \frac{x - \mu}{\sigma}$$

StandardScaler() will normalize the features i.e. each column of X, individually, so that each feature/column will have μ = 0 and σ = 1. The Standard Scaler assumes data is normally distributed within each feature and scales them such that the distribution centered around 0, with a standard deviation of 1 [17].

The Principal Component Analysis (PCA) algorithm is used to select the best features from the mushroom dataset. PCA is a technique from linear algebra that can be used to automatically perform dimensionality reduction. Reducing the number of features in a dataset can reduce the risk of overfitting and also improves the accuracy of the model [20]. We have used PCA with n_components = 2 for reducing the dimensions of the dataset.

E. Classification Modelling

After the feature extraction and selection, the supervised machine learning methods are applied to the data obtained. The machine learning methods to be applied, as discussed previously, are:

- Logistic Regression (LR)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Naïve Bayes (NB)
- Random Forest (RF)

F. Performance Evaluation of Algorithms

In this step, evaluation of the prediction results using various evaluation metrics like confusion matrix, classification accuracy, precision, recall, f1-score, etc. is done.

- **Confusion Matrix -**

It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another. It describes the complete performance of the model.

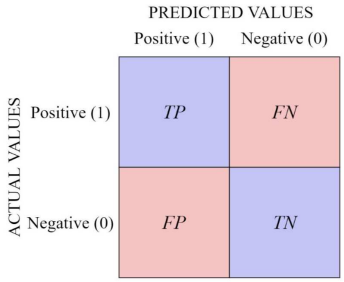

Figure 8 Confusion Matrix

Where *TP* = True Positives,
TN = True Negatives,
FP = False Positives,
FN = False Negatives.

- **Classification Accuracy -**

It is the ratio of the number of correct predictions to the total number of input samples. It is given as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision -**

Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. It attempts to answer the question: What proportion of positive identifications is actually correct? Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall / Sensitivity / True Positive Rate (TPR) -**

It is the number of correct positive results divided by the number of all relevant samples. Recall attempts to answer the question: What proportion of actual positives is identified correctly? Mathematically, recall is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score -**

It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for the F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, the F1 Score is defined as follows:

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

F1 Score tries to find the balance between precision and recall.

- **False Negative Rate (FNR) -**

False Negative Rate (FNR) tells us what proportion of the positive class got incorrectly classified by the classifier [2]. Mathematically, the FNR is given by:

$$FNR = \frac{FN}{TP + FN}$$

- **Specificity / True Negative Rate (TNR) -**

Specificity tells us what proportion of the negative class got correctly classified [2]. Mathematically, it is given by:

$$Specificity = \frac{TN}{TN + FP}$$

- **False Positive Rate (FPR) -**

False Positive Rate (FPR) tells us what proportion of the negative class got incorrectly classified by the classifier [2]. Mathematically, it is given by:

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

RESULTS

In this experimental study, six machine learning algorithms were used. These algorithms are LR, DT, KNN, SVM, NB, and RF. All these algorithms were applied to the UCI Mushroom Classification Dataset. Data was divided into two portions, training data, and testing data, both these portions consisting of 70% and 30% data respectively. Feature scaling using StandardScaler() was performed. The Principal Component Analysis (PCA) algorithm was used with n_components = 2 for reducing the dimensions and selecting the best features from the dataset [3]. All six algorithms were applied to the same dataset and results were obtained. Predicting accuracy is the main evaluation parameter that is used in this work. Accuracy is the overall success rate of the algorithm.

True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP) predicted by all the algorithms are presented in Table 1. In our case, TP means actual edible mushrooms. TN, actual poisonous mushrooms. FP, actually poisonous but predicted to be edible. FN, actually edible but predicted to be poisonous.

Algorithm	TP	FN	FP	TN
LR	2849	102	432	2303
DT	2951	0	0	2735
KNN	2873	78	244	2491
SVM	2893	58	374	2361
NB	2850	101	477	2258
RF	2951	0	3	2732

Table 1 TP, FN, FP, TN predicted by algorithms on the training set

Algorithm	TP	FN	FP	TN
LR	1218	39	198	983
DT	1130	127	128	1053
KNN	1206	51	143	1038
SVM	1234	23	174	1007
NB	1218	39	211	970
RF	1206	51	139	1042

Table 2 TP, FN, FP, TN predicted by algorithms on the test set

The training and test set visualizations are given below:

- **Logistic Regression**

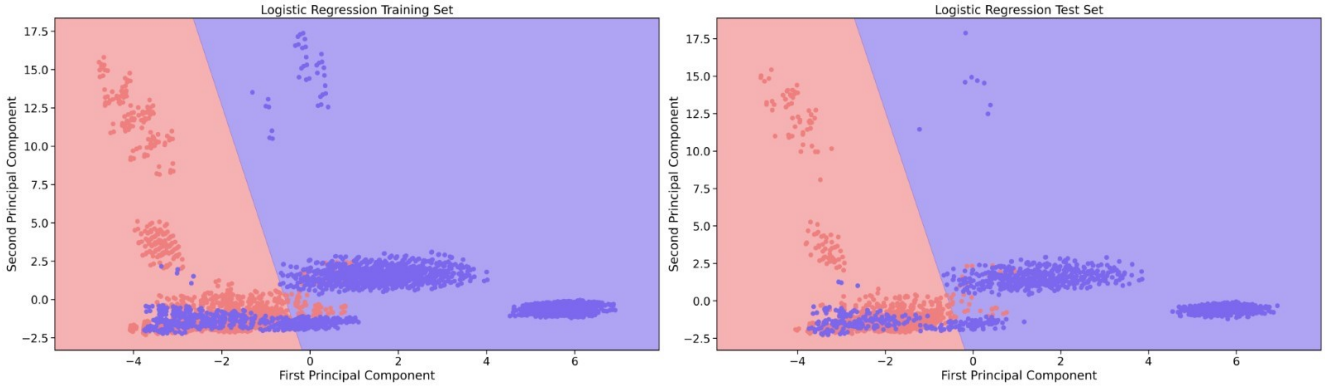

Figure 9 Logistic Regression Training and Test Set – PCA

Decision Tree

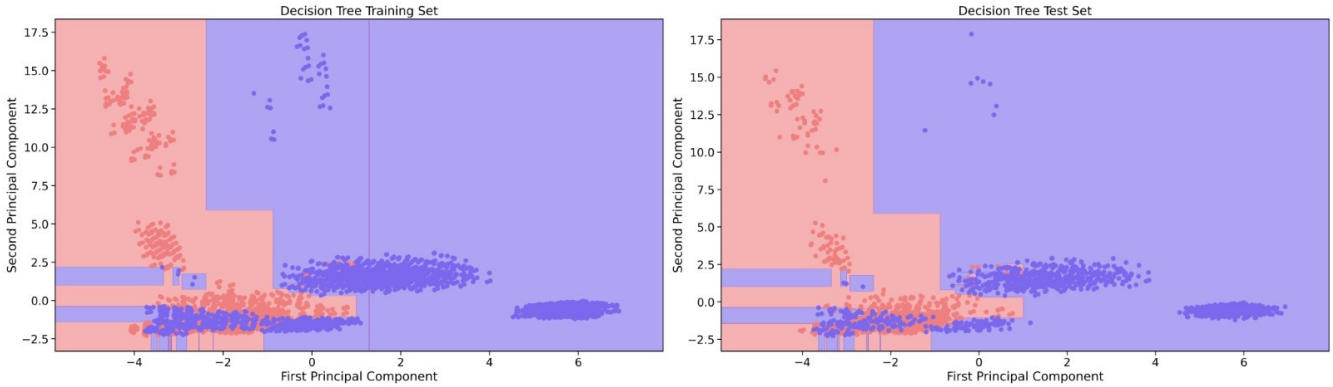

Figure 10 Decision Tree Training and Test Set – PCA

- **K-Nearest Neighbor**

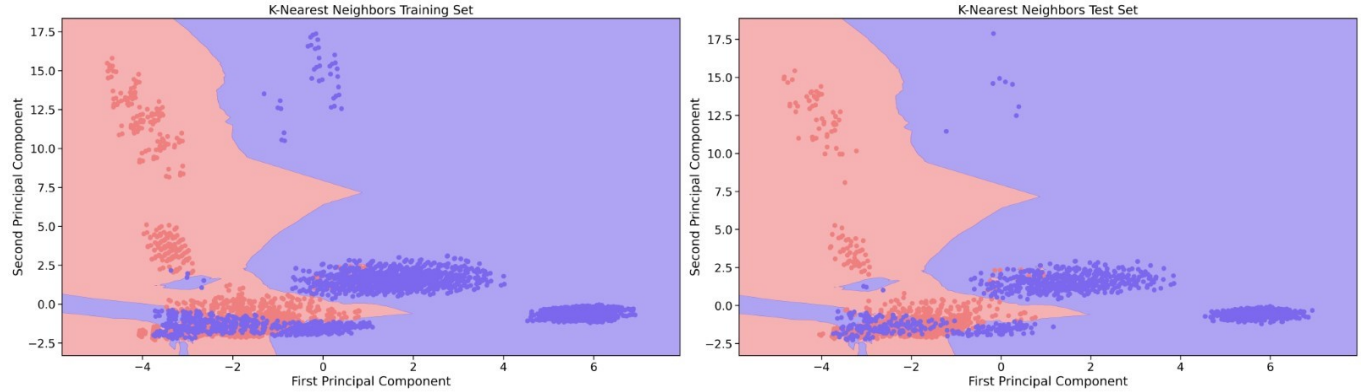

Figure 11 K-Nearest Neighbor Training and Test Set – PCA

Support Vector Machine

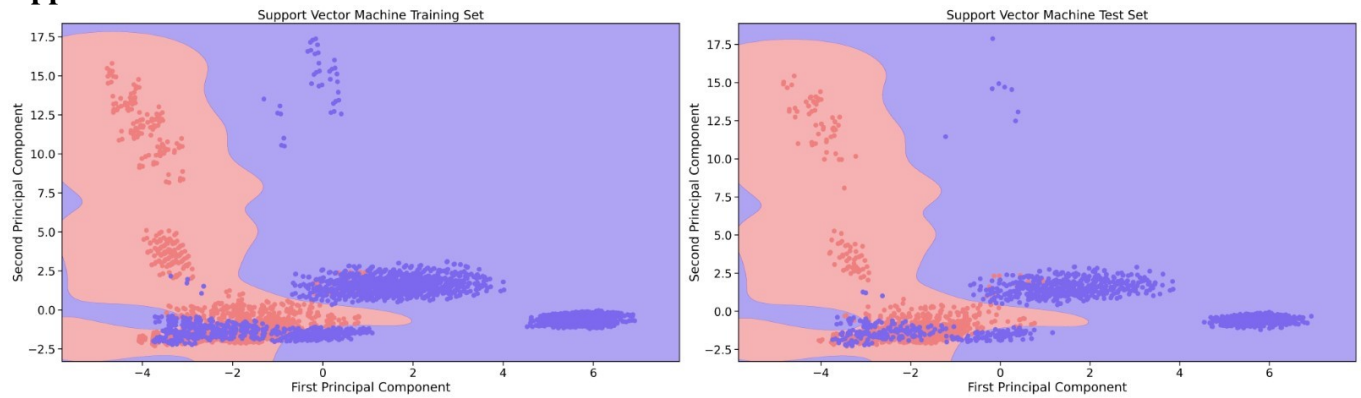

Figure 12 Support Vector Machine Training and Test Set – PCA

Naïve Bayes

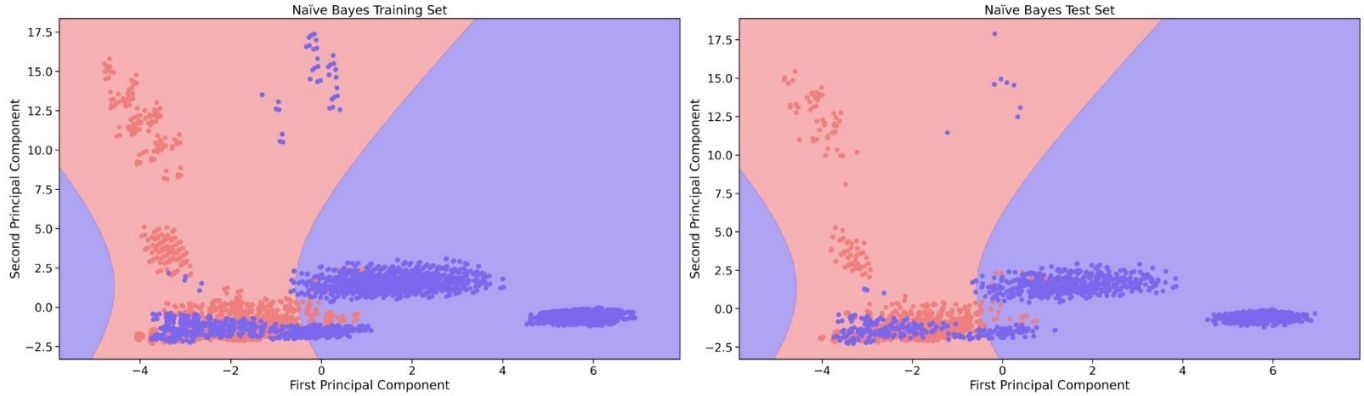

Figure 13 Naïve Bayes Training and Test Set – PCA

Random Forest

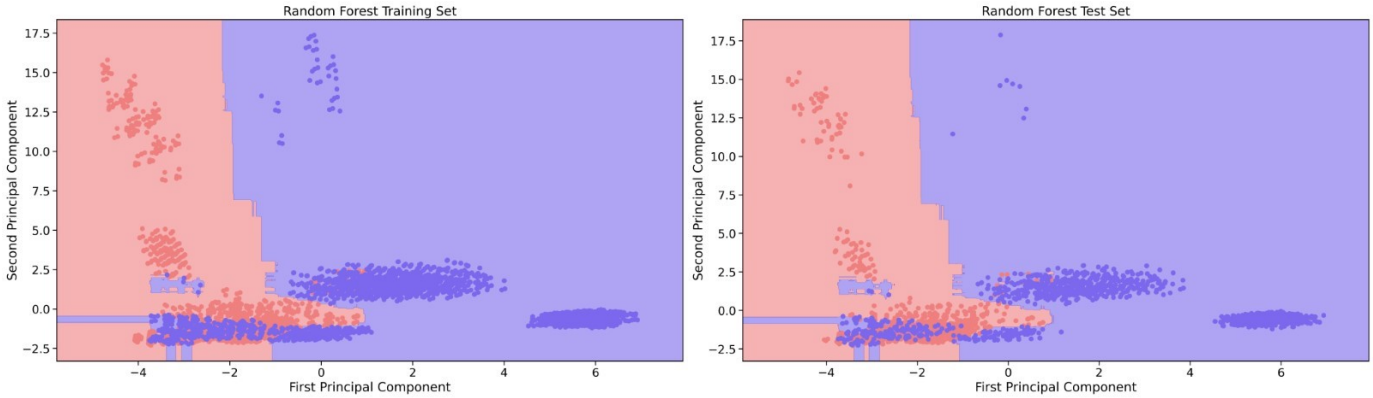

Figure 14 Random Forest Training and Test Set – PCA

We plotted a Receiver Operator Characteristic (ROC) curve which is an evaluation metric for binary classification problems, in our case, mushroom classification. It is a probability curve that plots the TPR against

FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

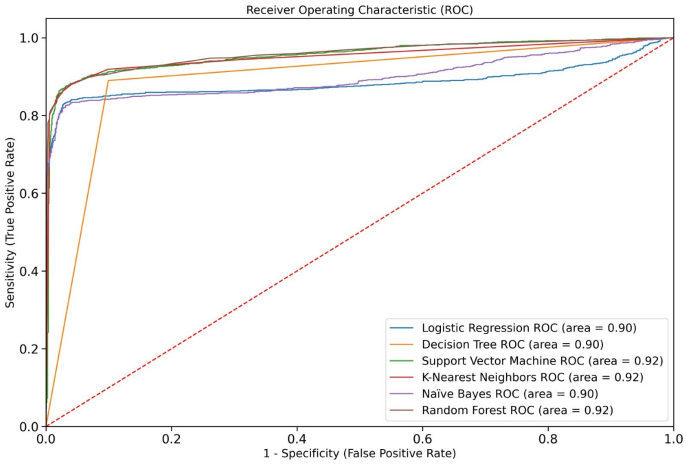

Figure 15 ROC Curve

It is evident from the plot that the AUC for the Random Forest and K-Nearest Neighbor ROC curve is higher than others. Therefore, we can say that Random Forest and KNN performed better than other classifiers. The training accuracy score, average accuracy score, standard deviation and test accuracy score of all six algorithms is given in the following table:

Algorithm	LR	DT	KNN	SVM	NB	RF
Training Accuracy	0.9061	1.0000	0.9434	0.9240	0.8983	0.9995
Average Accuracy	0.9066	0.8871	0.9291	0.9235	0.8987	0.9226
Standard Deviation	0.0103	0.0146	0.0103	0.0104	0.0113	0.0119
Test Accuracy	0.9028	0.8954	0.9204	0.9192	0.8975	0.9221

Table 3 Training accuracy, Average accuracy, Standard Deviation and Test Accuracy of algorithms

CONCLUSION

In this paper, six popular supervised machine learning algorithms are used for classifying mushrooms into edible or poisonous. These include LR, DT, KNN, SVM, NB and RF. Predictions were made about mushrooms (whether edible or poisonous) on the UCI mushroom classification dataset consisting of 8124 records. Principal Component Analysis (PCA) algorithm is used with n_components = 2 for reducing the dimensions of the dataset. There are a total of 23 categorical variables in this dataset which were converted into dummy/indicator variables. These 23 variables (which became 95 after conversion), were reduced to only 2 variables i.e. Principal Components. All six classification models were trained over these two principal components. From the experimental results obtained, it can be seen that Random Forest and K-Nearest Neighbor gave the highest test accuracy of 92.21% and 92.04% followed by Support Vector Machine with 91.92% test accuracy, Logistic Regression with 90.28% test accuracy, Naïve Bayes with 89.75% test accuracy and Decision Tree with 89.54% test accuracy.

REFERENCES

[1] Bhandari, Aniruddha. "Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization." Analytics Vidhya, 2020. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ (accessed Dec. 24, 2020).

[2] Bhandari, Aniruddha. "AUC-ROC Curve in Machine Learning Clearly Explained - Analytics Vidhya." Analytics Vidhya, 2020. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/ (accessed Feb. 02, 2021).

[3] Brownlee, Jason. "Principal Component Analysis for Dimensionality Reduction in Python." Machine Learning Mastery, 2020. https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/ (accessed Jan. 21, 2021).

[4] D. Dua and C. Graff, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

[5] F.Y, Osisanwo, Akinsola J.E.T, Awodele O, Hinmikaiye J. O, Olakanmi O, and Akinjobi J. "Supervised Machine Learning Algorithms: Classification and Comparison." International

Journal of Computer Trends and Technology 48, no. 3 (June 25, 2017): 128–38. https://doi.org/10.14445/22312803/ijctt-v48p126.

[6] Firdous, DrxHina. "Health Benefits Of Mushroom, Uses And Its Side Effects." Lybrate, 2020. https://www.lybrate.com/topic/benefits-of-mushroom-and-its-side-effects (accessed Sep. 24, 2020).

[7] Ismail, Shuhaida, Amy Rosshaida Zainal, and Aida Mustapha. "Behavioural Features for Mushroom Classification." In ISCAIE 2018 - 2018 IEEE Symposium on Computer Applications and Industrial Electronics, 412–15. IEEE, 2018. https://doi.org/10.1109/ISCAIE.2018.8405508.

[8] Kalač, Pavel. "A Review of Chemical Composition and Nutritional Value of Wild-Growing and Cultivated Mushrooms." Journal of the Science of Food and Agriculture 93, no. 2 (January 30, 2013): 209–18. https://doi.org/10.1002/jsfa.5960.

[9] Khourdifi, Youness, and Mohamed Bahaj. "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification." In 2018 International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2018, 1–5. IEEE, 2019. https://doi.org/10.1109/ICECOCS.2018.8610632.

[10] Lavanya, B. "Performance Analysis of Decision Tree Algorithms on Mushroom Dataset." International Journal for Research in Applied Science and Engineering Technology V, no. XI (November 13, 2017): 183–91. https://doi.org/10.22214/ijraset.2017.11029.

[11] Maurya, Pranjal, and Nagendra Pratap Singh. "Mushroom Classification Using Feature-Based Machine Learning Approach." In Advances in Intelligent Systems and Computing, 197–206. Singapore: Springer, 2020. https://doi.org/10.1007/978-981-32-9088-4_17.

[12] Mohammadi, Mehdi, Mursal Dawodi, Wada Tomohisa, and Nadira Ahmadi. "Comparative Study of Supervised Learning Algorithms for Student Performance Prediction." In 1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019, 124–27. IEEE, 2019. https://doi.org/10.1109/ICAIIC.2019.8669085.

[13] Norouzi, Monire, Alireza Souri, and Majid Samad Zamini. "A Data Mining Classification Approach for Behavioral Malware Detection." Journal of Computer Networks and Communications 2016 (2016): 1–9. https://doi.org/10.1155/2016/8069672.

[14] Nuanmeesri, Sumitra, and Wongkot Sriurai. "Development of the Edible and Poisonous Mushrooms Classification Model by Using the Feature Selection and the Decision Tree Techniques." International Journal of Engineering and Advanced Technology 9, no. 2 (2019): 3061–66. https://doi.org/10.35940/ijeat.b4115.129219.

[15] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart Disease Prediction Using Machine Learning Techniques: A Survey." International Journal of Engineering and Technology(UAE) 7, no. 2.8 Special Issue 8 (March 19, 2018): 684–87. https://doi.org/10.14419/ijet.v7i2.8.10557.

[16] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data Mining in Course Management Systems: Moodle Case Study and Tutorial." Computers and Education 51, no. 1 (August 2008): 368–84. https://doi.org/10.1016/j.compedu.2007.05.016.

[17] Roy, Baijayanta. "All about Feature Scaling." Towards Data Science, 2020. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35 (accessed Dec. 27, 2020).

[18] Roy, Baijayanta. "All about Categorical Variable Encoding." Towards Data Science, 2020. https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02 (accessed Dec. 14, 2020).

[19] Sarwar, Muhammad Azeem, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. 2018. "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare." In ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing, IEEE, 1–6. https://doi.org/10.23919/IConAC.2018.8748992.

[20] Widmann, Maarit, and Rosaria Silipo. "3 New Techniques for Data-Dimensionality Reduction in Machine Learning." The New Stack, 2019. https://thenewstack.io/3-new-techniques-for-data-dimensionality-reduction-in-machine-learning/ (accessed Jan. 18, 2021).