Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754746

COMPUTER VISION AND MACHINE LEARNING BASED MUSHROOM TYPES CLASSIFICATION

Prashant Sharma
M.Tech Scholar, Department of Computer Science & Engineering, Bansal Institute of Engineering & Technology, Lucknow, India. prashant.jimkanpur@gmail.com

Dr. C.L.P. Gupta
Professor, Department of Computer Science & Engineering, Bansal Institute of Engineering & Technology, Lucknow, India. clpgupta@gmail.com

Abstract— Given their abundance in vital elements, proteins, minerals, and vitamins, it is advisable that we include mushrooms in our diet. Many farmers grow mushrooms in greenhouses, where the climatic conditions are specific and the space is constrained. Among the millions of varieties of mushrooms that exist on earth, there are two distinct types: edible mushrooms and harmful mushrooms. Many people get food poisoning because they are not aware that the mushrooms are poisonous. Even in some other countries, cases of poisoning from poisonous mushrooms have been documented. The difference between harmful and edible mushrooms might be challenging to make because of their similar characteristics and abundance. By assisting in the discovery of significant patterns from millions or even billions of data records, data mining techniques, specifically classification, can be used to determine the type of mushrooms. This paper offers a comparative investigation of different classification techniques for recognizing a dataset of mushrooms.

Current research on classifying mushrooms focuses on using ML techniques individually, and some systems outperform others in terms of accuracy. Research on the classification of mushrooms is scarce. Based on this research, an integrated model was proposed that would combine rather than treat separately the decisions generated by the most accurate methodologies. The mushroom dataset was downloaded from the UC Irvine repository. The results show that the accuracy performance of the integrated model is 95% better than that of other techniques.

Index Terms— Machine learning, classification algorithm, edible mushroom, inedible mushroom, decision tree, Support vector machine, KNN etc

I. INTRODUCTION

A significant nutritional supplement for many tribal societies worldwide is wild edible mushrooms (WEM). They are employed in traditions not simply as foods but also for other things, the most prevalent of which is medicine. Due to their distinctive fruit body formation, wide availability, and the specialized abilities needed for recognizing and identifying edible varieties, wild mushrooms stand out from other creatures. Additionally, it causes powerful, opposing emotions in people, ranging from intense dislike to profound attraction. Typically, these emotions are a component of society and tradition [1]. Through international ethnomycological research that examined the variety of beneficial mushroom species over the toxic ones and raised awareness of their traditional applications in various cultures. It has

recently been observed that the growing interest in using mushrooms on a global scale can be attributed to their significance as functional foods, or foods that provide medicinal effects. People generally favor functional meals because using them reduces reliance on drugs, promotes health without side effects, helps avoid disease, and offers a means of self-care. Wild edible mushrooms have the inherent potential to serve as a major resource for the creation of unique, high-value products for both food and medicine.

Due to the current rise in human population, economic globalization, deforestation, and cultural homogenization, wild foods and their applications are in danger. Due to the slow pace of ethnomycological study, particularly in the southern regions of India, there is a relative lack of conversion of information about mushrooms' ecology and traditional applications as food and medicine. The use of mushrooms as meals and remedies in tribal traditions has not been well studied in Kerala, which is important in the present period since modernization and the death of elderly people cause a reduction in the state's traditional knowledge. Given that consumers now favor functional foods that are natural, wholesome, and pleasant, the relationship between food and health is crucial [2].

The goal of the current study, "Ethnomedicinal and Ecological Studies of Wild Edible Mushrooms Used by Selected Tribes in Palakkad and Wayanad Districts of Kerala," was to examine the biodiversity of wild edible mushrooms and its impact on the food security, health care, culture, identity, and environment of the study's six chosen tribal communities. This study also aims to nutritionally profile and rate a number of wild edible mushrooms.

According to Deacon (2006) [3], basidiomycetes and ascomycetes' fruit bodies are classified as "wild edible fungi," "macrofungi," or "mushrooms." Since mushrooms are not green organisms and lack chlorophyll, they cannot generate their own energy to grow and must always interact with other living things. They eat complex organic compounds that are found in the living or dead tissue of other species. By decomposing plant and animal waste, mushrooms contribute significantly to the environment. Other plants and animals utilize the remnants once they have degraded. Many mushroom species exchange energy from plants for a variety of nutrients that the plants themselves are unable to supply in sufficient amounts. Mushrooqms are often divided into three classes based on their environment. They are parasitic fungi that hurt the hosts, symbiotic/mutualistic organisms that grow in conjunction with other species, and saprophytes that grow on decaying organic materials.

The majority of mushroom species are saprophytes. By destroying dead organic matter like wood, leaves, needles, and manure, they purify the environment. Wood and plant waste are transformed into soft, fibrous, or crumbly material by mushrooms. The process releases a lot of water. Wild edible plants like Agaricus and Amanita are found in pastures and other grassy regions. Fungi that create symbiotic partnerships with other species benefit both parties. Mycorrhizae are symbiotic relationships between mushrooms and tree roots that typically grow on the forest floor. The mycelium of the fungus creates a covering that covers the outermost fine roots. Numerous mycorrhizal fungi are host-specific, meaning they can only flourish on particular tree species. Many of the important edible species that are foraged in the wild, such as cantherelles, russula, boletus, etc., are among the macrofungi that often produce ectomycorrhiza (Mohanan, 2014) [4]. The mycorrhizal association aids in the exchange of nutrients between the mycorrhizal fungus and the tree. As a byproduct of photosynthesis, the tree provides the fungus with sugar, and in exchange, the fungus provides the tree with other

nutrients like nitrogen and phosphorus, which it absorbs via its hyphae from the tiniest soil pores. Additionally, mycorrhizae can shield trees from the harmful impacts of pollution. The fungi store heavy metals that the tree may absorb and deposit in the fungus' fruiting body. The trees that have mycorrhizal associations exhibit enhanced tolerance to various stress conditions, making them less prone to frost and gaining more protection from soil-borne pathogenic microbes.

Among fungi, parasitic mushrooms are the most lethal. They multiply on active trees and other plants, stealing their nutrients and slowly killing their hosts. A parasite continues to break down the organic matter after the host plant dies, turning the situation into a saprophytic one. Examples of this group include polypores and shelf fungus, which grow on the sides of trees and feed on the living wood (Deacon, 2006) [3].

Due to their edibility, therapeutic worth, poisonousness, psychoactive nature, mycorrhizal and parasitic relationship with forest trees, and economic and ecological significance, mushrooms. Today, growing mushrooms is regarded as an agricultural technique for generating cash.

Data Mining

Data mining is the process of obtaining important information from big datasets that are stored in databases and other data repositories and contain associations, trends, and anomalies. Data warehouses and other data storage facilities contain large databases. Data cleansing, data integration, selection, transformation, mining, pattern evaluation, and knowledge display are all essential steps in knowledge discovery, which is a component of data mining. The term "data cleaning" refers to the process of taking out noise and missing values from the dataset, as well as learning more about the model that was used to access the noise and taking into account any adjustments that were made. Data from different sources are combined in the phase known as "data integration," which is a key objective. To retrieve the necessary data, a selection of the data is made. A process known as data transformation must initially comprise several data preparation methods in order to make the data suitable for mining [5]. These procedures include normalization and aggregation, for instance.

Based on the objectives that each specific activity seeks to achieve, the tasks related to data mining can be broadly categorized into two types. The terms "descriptive tasks" and "prediction tasks" are used to describe these two categories of labor, respectively. Descriptive data mining tasks are those that characterize the general characteristics of the data [6]. On the other hand, data mining tasks known as predictive data mining tasks carry out inference on an existing data set in order to predict how a new data set will behave.

Data mining includes a wide range of tasks, such as categorization, forecasting, time series analysis, associating, grouping, and summarizing data. Each of these positions relates to the descriptive or predictive aspects of data mining. Each of the aforementioned tasks can be completed singly or in combination by a data mining system as a part of data mining.

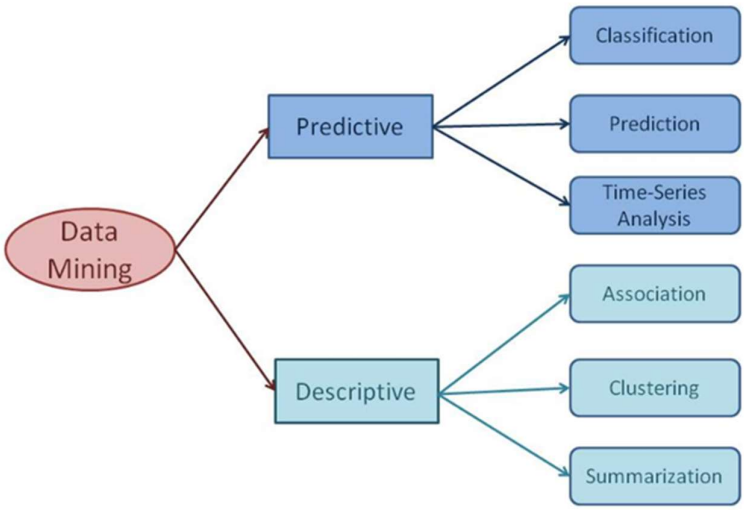


Figure 1: Data Mining Tasks

Machine Learning

The terms used in machine learning have been divided into the groups like supervised learning, unsupervised learning, reinforcement learning. The training data are examined by supervised learning, which produces the desired output that may be applied to mapping provided data [7]. Unsupervised learning is a machine learning technique that uses unlabeled data to derive a function to describe an unknown structure. Machine learning that uses behaviorist psychology to make decisions is called reinforcement learning.

Classification

The goal of classification is to create a model that can classify an object based on its characteristics. There will be a collection of records that can be accessed, each of which will have a unique set of attributes. One of the attributes will be a class attribute, and the classification task's goal will be to accurately ascertain the class attribute for the new batch of records.

On the basis of a training data set where the classification is known, the classification algorithm determines the categories to which a set of new items belongs. The task of classification involves supervised learning. Neural networks, support vector machines, k-nearest neighbors, naive Bayes, decision trees, and radial basis functions are some of the most popular classifiers. Unlike neural networks for classification, decision tree classifiers do not have a strong theoretical foundation, but they perform incredibly well in actual use.

The decision trees are known to perform better than the majority of the other classifiers, especially when employed alongside techniques like Random Forests. The procedure mentioned above is the first step of classification, which is a supervised data mining technique. to create a classification model using data on customers who have and have not defaulted over an extended period of time as building blocks (or training data). According to which group they most closely resemble, the algorithms will search for clients whose qualities fit the attribute patterns of prior defaulters and non-defaulters. Using these classifications to determine which clients are most likely to default. In a classification model, the target attribute may have more than two alternative values.

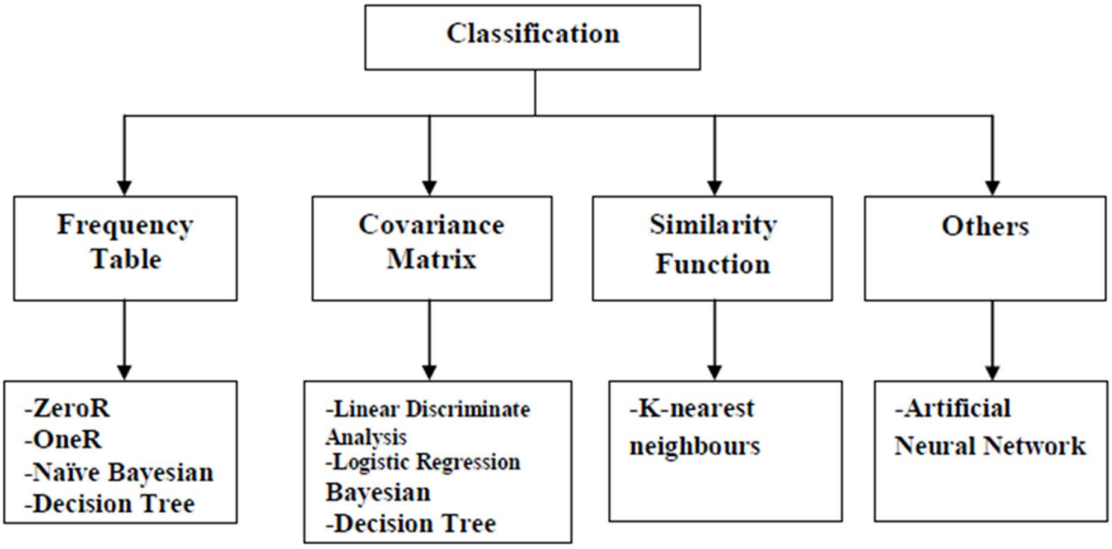


Figure 2: Classification Methodologies

II. METHODLOGIES

On the basis of a training set of data that includes observations whose category is known, classification is the process of determining a new observation category section. The cluster analysis technique is used to classify the objects based on how similar they are. Studies have been conducted to contrast the various classification techniques that have been created thus far. Various classifiers, including Nave Bayes, K Means, Decision Tree, Support Vector Machine (SVM), and K Nearest Neighbour (KNN) classifier, have been investigated in this article.

Logistic Regression (LR) Algorithm

Statistical experts and researchers use logistic regression (LR), one of the most significant statistical and data mining approaches, to analyze and categorize binary and proportional response data sets [2, 8]. One of LR's main benefits is that it can extend to multi-class classification problems and automatically provide probability [3]. The fact that most methods used in LR model analysis follow the same underlying principles as those in linear regression is another advantage. Furthermore, LR is amenable to the majority of unconstrained optimization approaches [5]. Using methods like the shortened Newton, LR importance has recently returned to popularity. Truncated Newton methods have been used to successfully resolve complex optimization problems.

Unfortunately, the conventional binary methods, consisting of LR, are contradictory in the presence of unbalanced and infrequent events data, limited samples, and/or specific sampling techniques (such choice-based sampling). Prior correction and weighting are the two most used correction methods. when there is little chance of interest among the population, Halili and Kamberi [9] showed that these modifications can still have an impact by applying them to the LR model.

If there are n instances, then there are d characteristics (also known as parameters or qualities), and y is a binary outcomes vector, then $X \in Rn \times d$ is a data matrix. With respect to each occurrence of xi Rd (a row vector in X), yi = 1 or yi = 0 is the outcome for i = 1... n. Let situations where yi = 1 indicate that an event occurs belong to the positive class, while situations where yi = 0 indicate that an event does not occur. It is important to determine if

instance xi is positive or negative. A Bernoulli experiment with a probability pi or an anticipated value E[yi] can be used to conceptualize an instance (the random component). Such an issue would have a matrix shape in a linear model.

$$y = X\beta + \varepsilon, \quad (1)$$

where ε denotes the error vector and

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1\, x_{11}\, x_{12} \cdots x_{1d} \\ 1\, x_{21}\, x_{22} \cdots x_{2d} \\ \vdots\ \vdots\ \ \vdots\ \ \ \vdots \\ 1\, x_{n1}\, x_{n2} \cdots x_{nd} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}, \text{and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

The unknown parameter vector has the properties xi ← [1, xi] and β ← [β0, βT]. From this point forward, it is assumed that the vector contains the intercept. Since y has a probability distribution and is a Bernoulli random variable, let's go on.

$$P(y_i) = \begin{cases} p_i, & \text{if } y_i = 1; \\ 1 - p_i, & \text{if } y_i = 0; \end{cases} \quad (3)$$

so the response's anticipated value is

$$E[y_i] = 1(p_i) + 0(1 - p_i) = p_i = \mathbf{x}_i\boldsymbol{\beta} \quad (4)$$

With a variation, too

$$V(y_i) = p_i(1 - p_i). \quad (5)$$

From the linear model, it follows

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \quad (6)$$

Therefore

$$\varepsilon_i = \begin{cases} 1 - p_i, & \text{if } y_i = 1 \text{ with probability } p_i; \\ -p_i, & \text{if } y_i = 0 \text{ with probability } 1 - p_i; \end{cases} \quad (7)$$

As a result, εi distribution has an anticipated value

$$E[\varepsilon_i] = (1 - p_i)(p_i + (-p_i)(1 - p_i) = 0 \quad (8)$$

and a variance

$$V(\varepsilon_i) = E[\varepsilon_i^2] - E[\varepsilon_i] = (1 - p_i)^2(p_i) + (-p_i)^2(1 - p_i) - (0) = p_i(1 - p_i). \quad (9)$$

The least squares method cannot be used because the response's expected value and variance are not constant (they are heteroskedastic), and the mistakes are not distributed regularly. Additionally, since yi ∈ {0, 1}, linear regression would result in values that are either above or below 1. Consequently, the logistic response function, illustrated in Figure 3, is the suitable one when the response vector is binary.

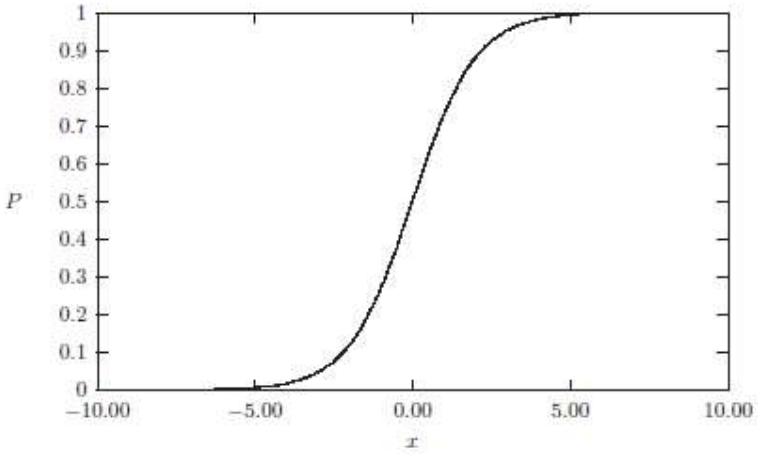


Figure 3: Logistic Response Function

Random Forest Algorithm

One of the greatest introductions to the Random Forest algorithm is provided here. The author gives examples in four distinct sectors to help newcomers understand about this method after introducing it with a real-life scenario. The author starts the essay by highlighting a feature of the Random Forest method that intrigues him: the fact that it can be applied to both classification and regression issues. The other benefits of the Random Forest algorithm will be listed at the conclusion of the paper. The classification exercise was the author's choice for this post because it will be simpler for a beginner to understand. The application issue in the upcoming post will be regression.

As a supervised classification technique, the Random Forest algorithm is the first. We can tell from the name that it aims to generate a random forest in some way. The more trees in the forest, the more accurate its results will be. There is a direct correlation between the quantity of trees in the forest and the results it can produce. However, it is important to keep in mind that developing the forest is not the same as making a choice based on information acquisition or an index method.

The author provides four resources to assist anyone dealing with decision trees for the first time in learning about and comprehending them properly. The decision tree is a tool for decision-support. It shows the potential results using a graph that looks like a tree. If you give the decision tree a training dataset with targets and features, it will produce a set of rules. You can make predictions using these guidelines. The author gives the following example to demonstrate his point: You should gather information on your daughter's past likes for animated movies and take into account specific traits if you want to predict whether your daughter would appreciate an animated movie. The decision tree approach is then used to generate the rules. You can then enter the movie's features to find out if your daughter will like it. Information gain and Gini index computations are used to determine these nodes and create the regulations.

A description of the random forest system was given. He is essentially a metal architect, but because of his ad hoc classification system, Random Tree, Weka is included in the decision-making tree method. A common training machine for natural trees creates a distinct choice matrix and frequently results in high risk variables in each cycle of the hauling method [46].

The tree has now reached its full potential and is not cut. A new dataset's tree is compressed downward. When the command line node is finished, the teaching example is assigned to the tag. This process, called a Random Forest Production, is carried out across all woods.

We've employed the Random Forest algorithm Random Forest is a versatile and user-friendly software technique that, most of the time, gives excellent results without the need for superfluous parameters. It is also among the most widely used approaches since it is simple to use and can be applied to both classification and regression. The supervised learning algorithm Random Forest. It's making woodlands and somehow randomizing it. The decision trees used in The Forest's construction are mostly taught using bagging techniques. The main idea behind bagging approach is that a variety of training designs improve overall outcome. To put it simply: Random tree produces and integrates numerous choice trees to produce more accurate and reliable forecasts. One important advantage of random forests is that they may be applied to both ranking and regression problems.

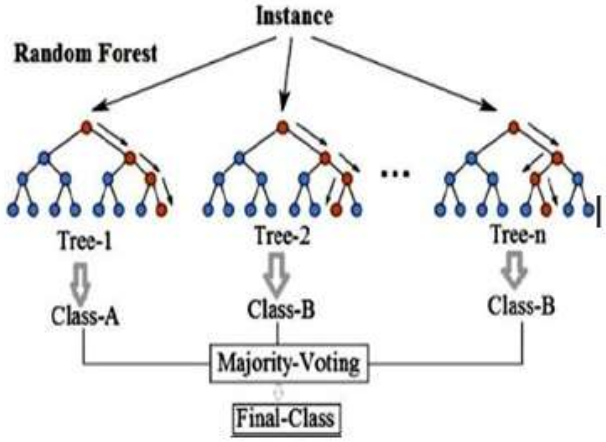


Figure 4: Random Trees Classifier

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a binary classifier that creates an optimal margin hyper-plane to divide the input points into two groups [10–12]. It solves a challenging quadratic programming problem (QPP) for this reason. As opposed to the other existing classification methods, SVM offers a worldwide answer by creating a special hyper-plane to divide the data points of various classes. Because SVM adheres to the Structural Risk Minimization (SRM) principle, it increases generalization capacity while reducing risk during the training phase. Originally designed for binary classification, SVM was later effectively expanded by academics to multi-class classification issue scenarios [13–17]. The training dataset for binary classification can be shown as:

$$T = \{(x1, y1), (x2, y2), \dots, (xl, yl)\} \quad (10)$$

Where xi ∈ Rn, i = 1,2, … , l , points of the input data in n-dimensional space in real time R and yi ∈ {+1,-1} complies with the class designation. 'l' stands for the training dataset's size. The training dataset for multi-class classification has the similar representation. The only difference is that the data points belong to multiple class labels i.e., $yi \in \{1,2, \dots, K\}$ where K is the total number of class labels present in the dataset. Consider a matrix Xi consists of

data points which are linearly separable in Rn. SVM solves following primal QPP for the classification of data points:

$$\min_{w,b} \frac{1}{2} w^T w \quad (11)$$

$$\text{s.t.} \quad X_i w \geq 1 - b \quad \text{for } y_i = 1$$

$$X_i w \leq -1 - b \quad \text{for } y_i = -1$$

and generates following hyper-plane

$$w^T x + b = 0 \quad (12)$$

which lies in between two bounding hyper-planes

$$w^T x + b = 1 \text{ and } \quad w^T x + b = -1 \quad (13)$$

Where $w \in Rn$ and $b \in R$ are normal vector and bias term respectively. The plane separates the two classes' data points with a margin of $2/\|w\|2$. Here $\|.\|2$ represents the L2 norm. The data points lie on the hyper-plane are also known as support vectors. Primal problem of SVM for the data points which are not strictly linearly separable can be defined as:

$$\min_{w,b,\xi_i} \frac{1}{2} w^T w + ce^T \xi_i$$

$$\text{s.t.} \quad X_i w + \xi_i \geq 1 - b \quad \text{for } y_i = 1$$

$$X_i w - \xi_i \leq -1 - b \quad \text{for } y_i = -1, \xi_i > 0, i = 1,2, \ldots, l \quad (14)$$

Here ξi represents the error variable associated with the ith data point. The above equation can be solved by obtaining wolfe dual of (14) as:

$$\max_{\alpha} e^T \alpha - \frac{1}{2} \alpha^T y_i X_i X_i^T y_i \alpha$$

$$\text{s.t. } e^T y_i \alpha = 0, 0e \leq \alpha \leq ce \quad (15)$$

Where $\alpha \in Rn$ is lagrangian multiplier. After solving above equations, the hyper-plane parameters are obtained as:

$$w = X_i^T y_i \alpha \quad \text{and} \quad b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} y_i - X_i w \quad (16)$$

Where NSV represents number of support vectors. SVM classifier assigns the class to a test data point according to the sign of following decision function:

$$f(x) = sign(w^T . x + b) \quad (17)$$

If $(x) > 0$, the +1 class has been given to the test data point. while for the opposite case it is assigned to class -1. Figure 4 shows the geometric representation of traditional SVM in two-dimensional feature space. Different symbols are used to symbolize the data points of each class. The plane represented in red color is the separating plane; which assigns the two classes' data points the most degree of margin:

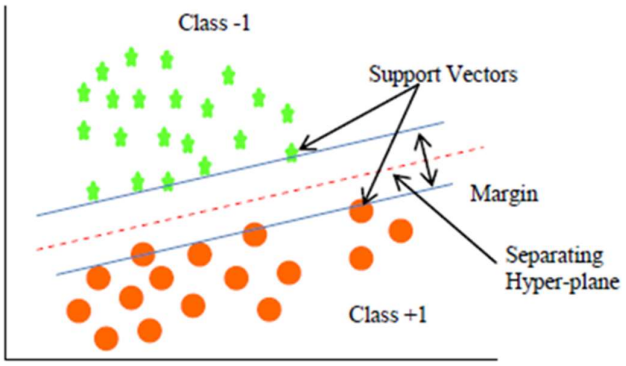


Figure 4: Geometric representation of linear SVM for binary classification

SVM also works well for the data points which are not separable by linear class boundary by using the concept of kernel trick. For this purpose, the data points are transformed into higher-dimensional feature space by using kernel function. Several kernel functions such as Polynomial, Gaussian, Sigmoid etc. are used by the researchers for the mapping of data points into higher dimensional space in order to make their separation easier. SVM is one of the most popular classification methods because it has demonstrated superior performance when compared to other machine learning approaches. It has applications in a wide range of fields, including text categorization, defect prediction, speech recognition, intrusion detection, disease detection, bankruptcy prediction, face identification, emotion detection, time series forecasting, and others.

Decision Tree (DT)

One of the most popular algorithms, in both practical applications and scholarly research, is the decision tree. Non-parametric technique is adaptable. Robustness: Remains unchanged when each input variable is transformed in a monotone (strictly speaking) manner. Robust against the addition of unrelated input variables, according to feature selection. Interpretability: A number of smaller, local judgments can be used to approximate a larger, more complex decision. Speed: Greedy algorithms that use a divide-and-conquer tactic without going back.

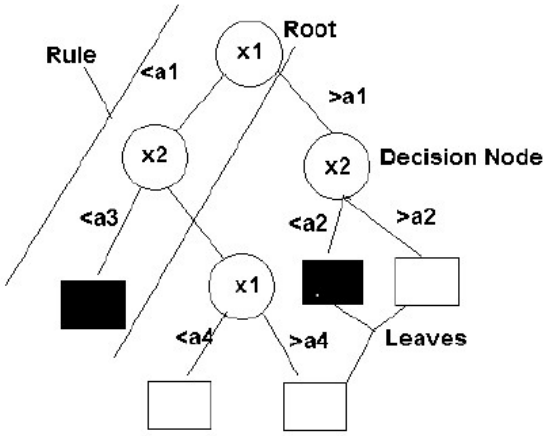


Figure 5: Decision Tree Diagram

Each decision node in Figure 5, the illustration of a decision tree comprises a test for one attribute, and each child branch corresponds to a potential attribute-value. Each leaf or terminal

node forecasts a class label. A categorization rule corresponds to each path from the root to the leaf.

k-Nearest Neighbors (kNN) Algorithm

Due to its ease of use and excellent accuracy, the K Nearest Neighbour (KNN) method has been utilized in a variety of data analysis applications, including pattern recognition, data mining, databases, and machine learning. It has been acknowledged as one of the top 10 data mining methods (Wu et al. 2008) [18]. A lazy learning technique used for classification is KNN. In machine learning, it is the simplest algorithm. Any form of label can be predicted using this strategy.

Instances are categorized using KNN classification based on similarity. It is an example of a lazy learning algorithm where computing is postponed until after classification and the function is approximated locally. KNN is mostly utilized for classifying and clustering data. Many scientists discovered that the KNN algorithm performs well in their tests using various datasets. The missing variables in the mushroom dataset make it difficult. The Euclidean Distance's neighboring column's equivalent values are used in the KNN method to fill in the missing values. It uses the value from the immediately following column if the matching value from the closest neighbor is also absent. Comparing this concept to other approaches, it is straightforward and extremely competitive. The weakness of KNN is the absence of probabilistic semantics, which enables the use of posterior prediction probabilities.

Numerous writers have improved KNN in order to increase its effectiveness. An algorithm known as the class-wise KNN (C-KNN) has been created and tested on the mushroom dataset. The lowest class-wise distance is used to classify the data in this instance. For C-KNN, accuracy is 78.16%. To categorize instances of the mushroom database, a KNN model combining K means and KNN algorithms has been developed. Here, eliminating the noise increases efficiency while also enhancing data quality. K-means removes instances that were improperly classified, while KNN is used for classification.

The KNN algorithm:

Stepi1: Each of theinew instances is checkediwith the already availableicases, based on distanceiassignment and classified usingik value.

Step2:.The distanceiwill be less, ifithe instances are moreisimilar and vice versa.i

Step 3: Observe theidistance, k -value andiinstance. Based on theseiobservations instances are assignedito a specific class.i

Step4: The prediction isibased on the k-value.iSo KNN classifier isik-dependent. Here k representsithe number of nearestineighbors and for differentivalues of k, outcomeimay not be theisame [24].

Step 5:iDetermine the value ofik for datssets foriclassification accuracy.

III. RESULT AND DISCUSSIONS

DATA SETDESCRIPTION

The mushroom dataset might be found at UCI mushroom data, where it was downloaded [19]. The data collection has a total of 8124 rows of data records and 22 different qualities to consider. Each type of mushroom is classified according to whether or not it is edible or harmful. These rows are broken up into 4208 edible mushrooms and 3916 toxic mushrooms. The attributes that are utilized to classify mushrooms are outlined in Table 1, which may be found here.

Before continuing with the process, the initial step is to organize the data. In order to prepare the data, it will be necessary to get rid of any null values and features that are repeated. Python packages that were utilized in the creation of datasets from raw data. The 22 characteristics are necessary for determining if a mushroom is edible or inedible. It is necessary for us to investigate these characteristics in order to determine which ones make a significant contribution to the classification procedure. We have come to the conclusion that "stalk-root" and "stalk-root" should both be removed. The reason for this is that the "stalk-root" feature is lacking 2480 data, and the "veil-type" feature only contains a single value, meaning that it cannot assist us in classifying anything. Furthermore, it is anticipated that "gill-color" will have the greatest contribution to the classification, so this factor needs to be taken into consideration.

Table 1: Mushroom dataset attribute information

No	Attribute	values
1	cap-shape	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2	cap-surface	fibrous=f,grooves=g,scaly=y,smooth=s
3	cap-color	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4	bruises	bruises=t,no=f
5	odor	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6	gill-attachment	attached=a,descending=d,free=f,notched=n
7	gill-spacing	close=c,crowded=w,distant=d
8	gill-size	broad=b,narrow=n
9	gill-color	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10	stalk-shape	enlarging=e,tapering=t
11	stalk-root	Bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12	stalk-surface-above-ring	fibrous=f,scaly=y,silky=k,smooth=s
13	stalk-surface-below-ring	fibrous=f,scaly=y,silky=k,smooth=s
14	stalk-color-above-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15	stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16	veil-type	partial=p,universal=u
17	veil-color	brown=n,orange=o,white=w,yellow=y
18	ring-number	none=n,one=o,two=t
19	ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n, pendant=p, sheathing=s, zone=z
20	spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21	population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22	habitat'	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

The descriptions of fictitious samples from 23 different species of gilled mushrooms in the Agaricus and Lepiota Family Mushrooms are included in this dataset. The Audubon Society Field Guide to North American Mushrooms, published in 1981, included the following descriptions. Each species is classified as either absolutely safe to consume, absolutely poisonous, or of unknown edibility and is not advised for consumption. This more recent class was brought together with the toxic one [4]. This dataset has 22 attributes, and there are 8124

different mushrooms included in it. The information regarding the dataset's attributes may be found in table 1.

Data Preprocessing

The dataset is divided into two categories: those that are edible and those that are poisonous. A bar graph is plotted in order to determine whether or not each is in balance. Because the data is categorical, the ordinal representation requires being converted via Label Encoder. Label Encoder makes a numerical representation of each value contained in a column [20]. Figure 6 depicts the total count for each class, while Figure 7 displays the dataset following the application of labels.

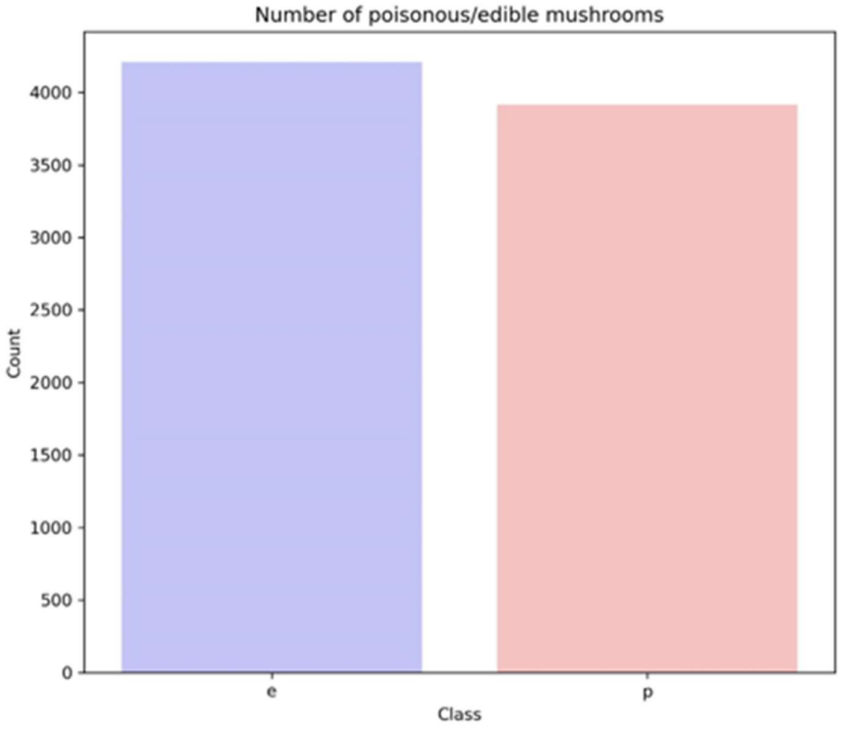


Figure 6: Count of edible and poisonous mushrooms

	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

Figure 7: Label Encoding

In exploratory data analysis, correlation matrices are an essential tool that must be used. It is helpful to have an understanding of the connection between the variables and the columns. In order to visually show the degree of correlation between the variables, a heatmap is displayed.

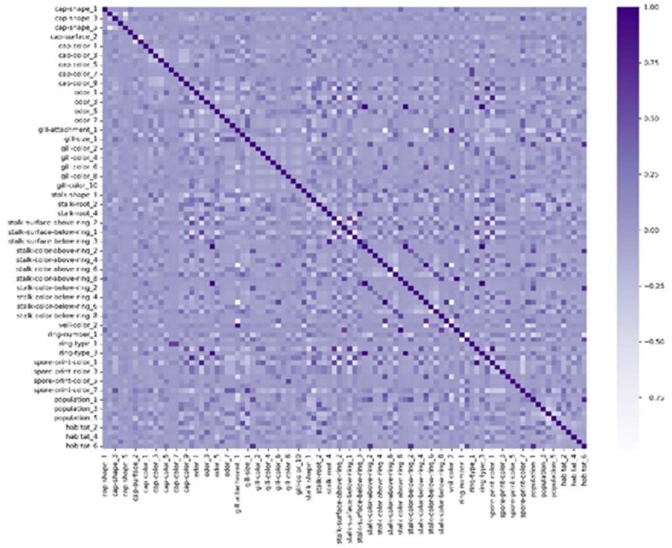


Figure 8: Correlation between the dummy/indicator variables

The visualizations of training and test set are given here:

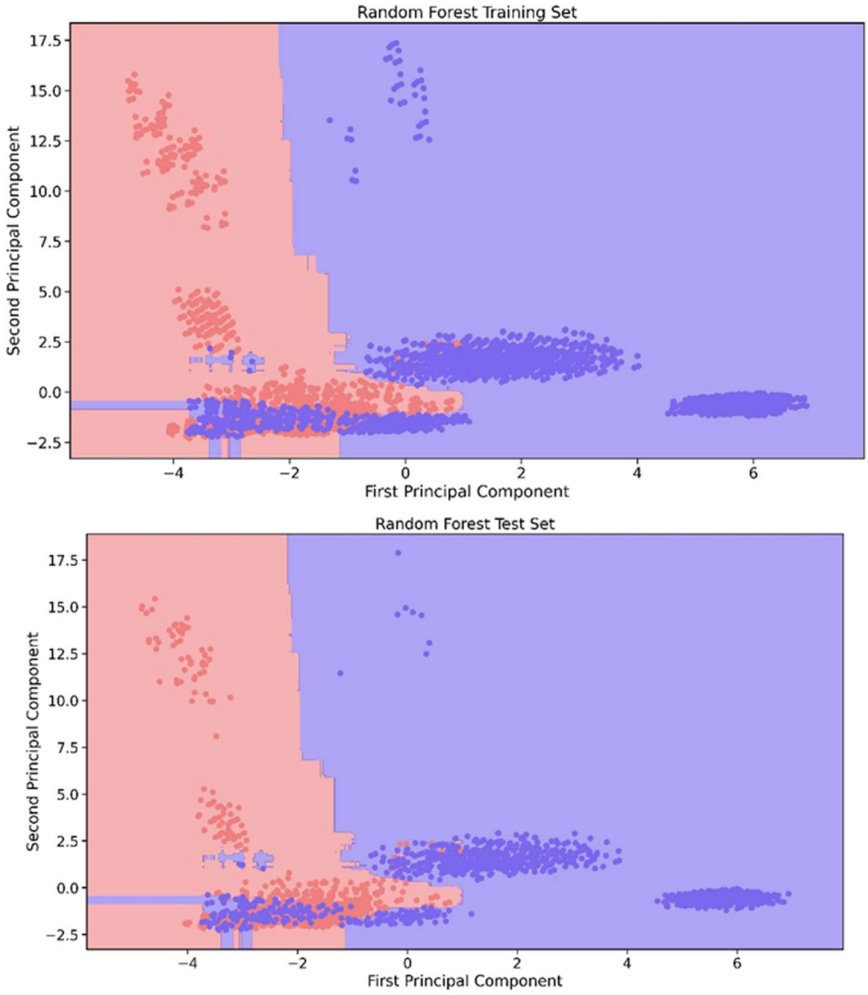


Figure 9: Random Forest (RF) Training and Test Set

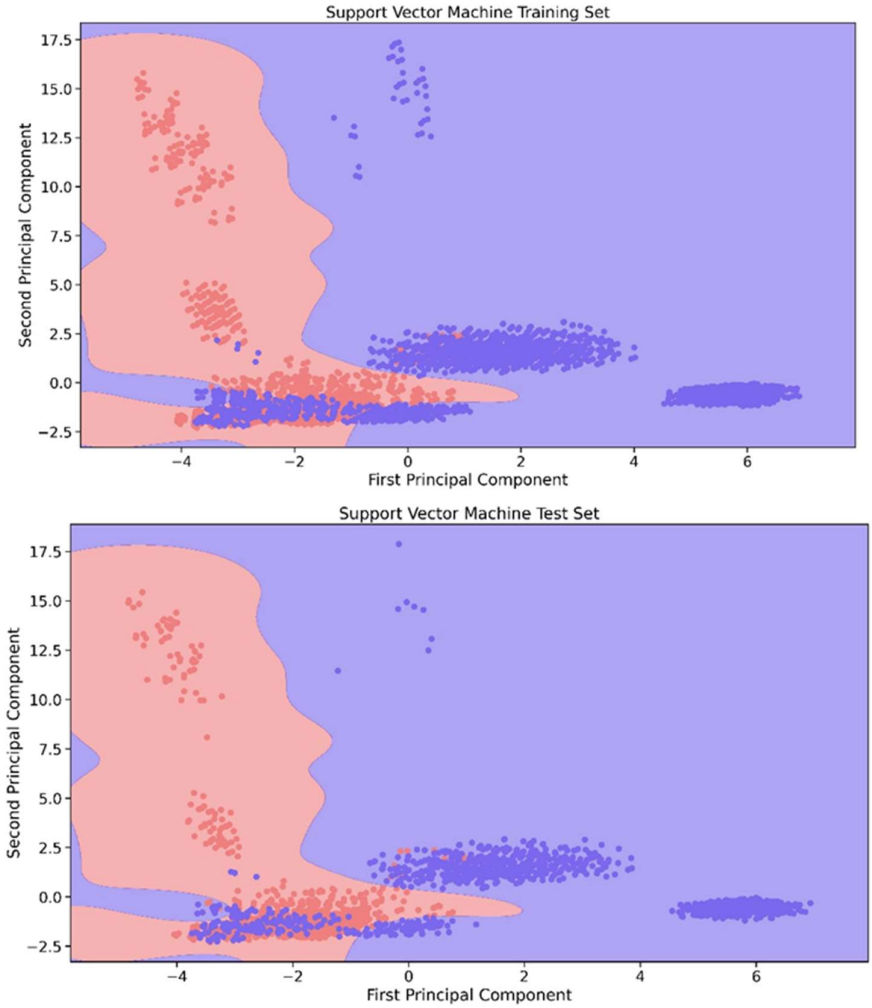
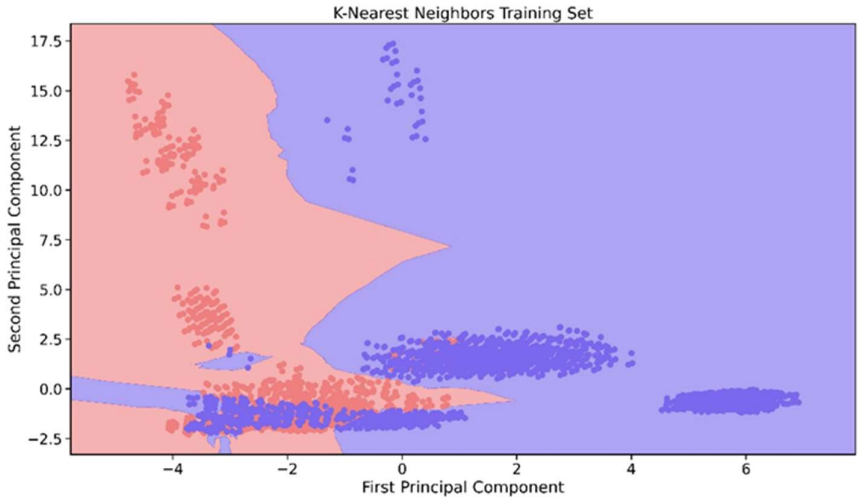


Figure 10: Support Vector Machine (SVM) Training and Test Set



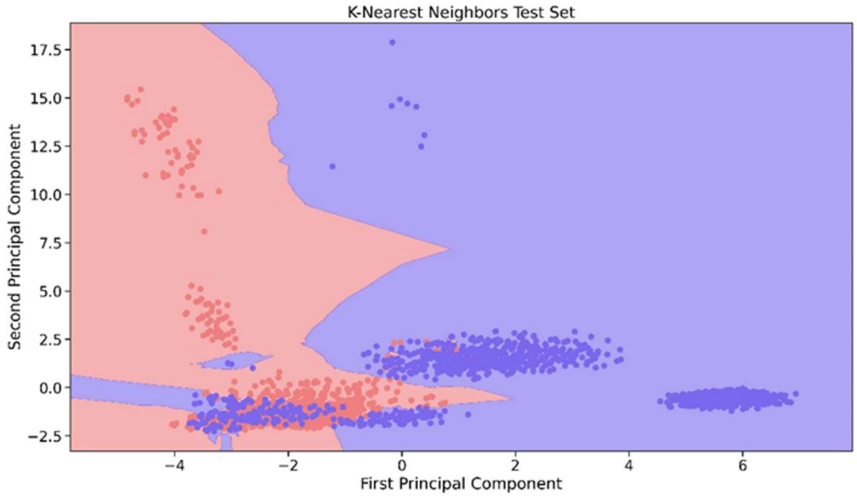


Figure 11: KNN Training and Test Set

The Training and Test Sets have been plotted in figure 9 to figure 11 for Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors algorithms.

A Receiver Operator Characteristic (ROC) curve, which is an evaluation metric for binary classification problems, specifically mushroom classification, was plotted by ourselves. It is a probability curve that plots the TPR against the FPR at a variety of threshold values and, in essence, differentiates the "signal" from the "noise." The area under the curve, often known as AUC, is a metric that is used to evaluate the capacity of a classifier to differentiate between different categories. A graph that displays the performance of a classification model at all classification thresholds is referred to as a ROC curve. A higher performance can be inferred from classifiers that produce curves that are located closer to the upper left corner. Hence, the ROC curve can be summarized using AUC and shown in figure 12. The highest ROC value is 0.9502 for KNN classifier that means KNN has best performance among all three classifiers.

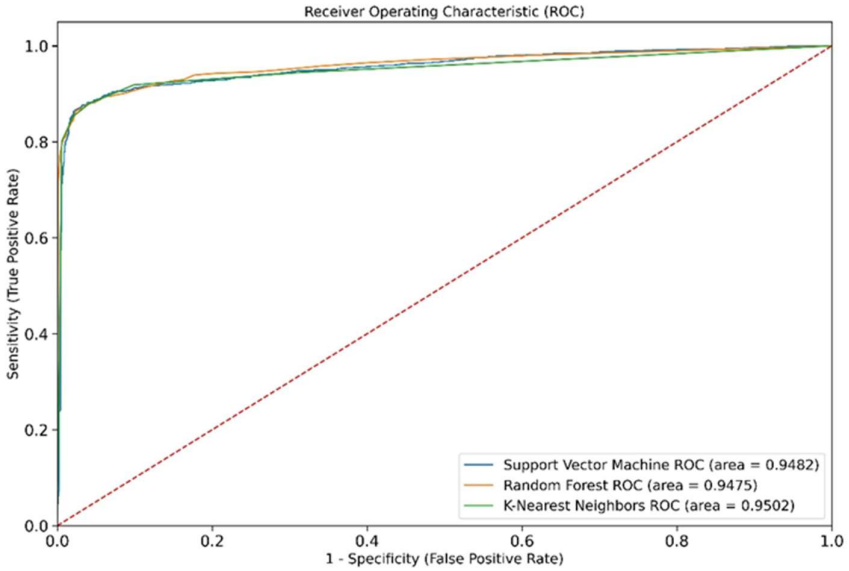


Figure 12: Receiver Operator Characteristic (ROC) Curve

IV. CONCLUSION

The mushroom is a significant contributor of both necessary proteins and vitamin content. On the other hand, the vast majority of identified mushroom types are toxic. ML learning was utilized to determine whether or not certain types of mushrooms were edible based on their properties. Previous research utilized ML techniques alone, and it was found that certain strategies performed significantly better in terms of accuracy than others. Consequently, it is difficult to decide which approaches should be taken because there are many viable options. Within the scope of this investigation, we have utilized a number of different machine learning classifiers on the "mushroom data" dataset. The dataset obtained from the repository on the UCI campus. Following an examination of the data, we have discovered that one feature known as "stalk-root" is missing a significant number of values, while another feature known as "veil-type" has the same values for each row. These two characteristics are removed so as to eliminate any potential impact they may have had on the classifications. In this case, the "odor_n" attribute is the most significant and influential aspect in the decision-making process. On the basis of the findings, we are able to draw the conclusion that an inedible mushroom is more likely to have an odor.

References

[1] Adejumo, T. O and Awosanya, O. B. 2005. Proximate and mineral composition of four edible mushroom species from South Western Nigeria. African Journal of Biotechnology, 4 (10), 1084-1088.

[2] Agrahar-Murugkar, D and Subbulakshmi, G. 2005. Nutritional value of edible wild mushrooms collected from the Khasi hills of Meghalaya. Food Chemistry, 89, 599-603.

[3] Deacon, J. 2006. Fungal Biology, Blackwell Publishing, Malden, pp. 33-34.

[4] Mohanan, C. 2014. Macrofungal diversity in the Western Ghats,Kerala, India: Members of Russulaceae. Journal of Threatened Taxa, 4 (6), 5636–5648.

[5] Lakshmi, KR & Kumar, SP 2013, 'Utilization of data mining techniques for prediction of diabetes disease survivability', International Journal of Scientific and Engineering Research, vol. 4, no. 6, pp. 933-942.

[6] Ashwinkumar UM & Anandakumar KR, 'Predicting early detection of cardiac and diabetes symptoms using data mining techniques', doi: 10.7763/ipcsit.2012.v49.20, pp. 106-115.

[7] Cwiklinska-Jurkowska, M 2009, 'Performance of the support vector machines for medical classification problems', Biocybernetics and Biomedical Engineering, vol. 29, no. 4, pp. 63-81

[8] Hamonangan, R., M.B. Saputro, and C.B.S.D.K. Atmaja, Accuracy of classification poisonous or edible of mushroom using naïve bayes and k-nearest neighbors. Journal of Soft Computing Exploration, 2021. 2(1): p. 53-60.

[9] Halili, F. and F. Kamberi, Performance analysis of classification Algorithms: A case study of Naïve Bayes and J48 in Big Data. Applied Mathematics and Computation. 2(2): p. 50-57.

[10] Nyanhongo, G. S., Gübitz, G., Sukyai, P., Leitner, C., Haltrich, D and Ludwig, R. 2007. Oxidoreductases from Trametes spp. in biotechnology: A wealth of catalytic activity. Food Technol. Biotechnol., 4, 250–268.

[11] Panda, A. K and Swain, K. C. 2011. Traditional uses and medicinal potential of Cordyceps sinensis of Sikkim. Journal of Ayurveda and Integrative Medicine. 2 (1), 9-13

[12] Lakshmi, KR & Kumar, SP 2013, 'Utilization of data mining techniques for prediction of diabetes disease survivability', International Journal of Scientific and Engineering Research, vol. 4, no. 6, pp. 933-942.

[13] Ashwinkumar UM & Anandakumar KR, 'Predicting early detection of cardiac and diabetes symptoms using data mining techniques', doi: 10.7763/ipcsit.2012.v49.20, pp. 106-115.

[14] Bellazzi, R, Ferrazzi, F & Sacchi, L 2011, 'Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews', Data Mining and Knowledge Discovery, vol. 1, no. 5, pp. 416-430.

[15] Hourali, M & Montazer, GA 2011, 'An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology', Advances in Fuzzy Systems, doi: 10.1155/2011/683976

[16] Arif, M & Akram, MU 2010, 'Pruned fuzzy K-nearest neighbor classifier for beat classification', Journal of Biomedical Science and Engineering, vol. 3, no. 4, pp. 380-389.

[17] Ashari, A, Paryudi, I & Tjoa, AM 2013, 'Performance comparison between naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool' vol. 4, pp. 33-39.

[18] Shirui, Pan., Jia, Wu., Xingquan, Zhu., and Chengqi, Zhang, IEEE., ―Graph Ensemble Boosting for Imbalanced Noisy Graph Stream Classification‖, IEEE Transactions on Cybernetics Vol. 45, No. 5, pp. 954 – 96, 2015.

[19] Ritika Raisanen. 2009. Dyes from lichens and mushrooms, In:Handbook of Natural Colorants, Thomas Bechtold and Rita Mussak (ed.) John Wiley and Sons, U.K, pp. 183–200.

[20] Kulshreshta, S., Mathur, N and Bhatnagar, P. 2014. Mushroom as a product and their role in mycoremediation. AMB Express, 4,29..