# MUSHROOM ANALYSIS USING CLASSIFICATION ALGORITHMS

**A PROJECT REPORT**

*for*

**DATA MINING TECHNIQUES (ITE2006)**

*in*

**B.Tech (Information Technology)**

*by*

**ESHA KUMAR (16BIT0396)**

**TANISHQ GUPTA (16BIT0380)**

**Winter Sem, 2019**

*Under the Guidance of*

**Prof. B. VALARMATHI**

Associate Professor, SITE


VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

**School of Information Technology and Engineering**

APRIL, 2019

## DECLARATION BY THE CANDIDATE

We here by declare that the project report entitled **"MUSHROOM ANALYSIS USING CLASSIFICATION ALGORITHMS"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **Prof. B.Valarmathi.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date :


VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

**School of Information Technology & Engineering [SITE]**

## CERTIFICATE

This is to certify that the project report entitled **"MUSHROOM ANALYSIS USING CLASSIFICATION ALGORITHMS"** submitted by **Esha Kumar (16BIT0396), Tanishq Gupta (16BIT0380)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

**Prof. B.Valarmathi**

**GUIDE**

**Asso. Professor, SITE**

# Mushroom Analysis Using Classification Algorithms

**Esha Kumar[1], Tanishq Gupta[2]**

[1]**Department of Information Technology, VIT University, Vellore, Tamil Nadu, India**

[2]**Department of Information Technology, VIT University, Vellore, Tamil Nadu, India**

**Abstract**

This project will investigate the data mining of mushroom data in order to create one or more classification models which are capable of identifying whether mushroom attribute is edible or poisonous. The data used in this project were sourced and presented by "UCI Machine Learning" for mushroom classification. We are interested to learn how well can we predict whether the mushroom is edible or poisonous. Eight different data mining algorithms Random Forest, KNN, Support Vector Machines, Logistic Regression, Linear SVC, Stochastic Gradient Descent, Naïve Bayes classification and Perceptron are used to answer the same question. A description of the predictive significance of each attribute, interesting or useful patterns which were found and any transformations applied to the data is provided with all proposed models. First, we will analyse the data, by showing it's characteristics, and then we will classify it to achieve maximum accuracy and precision.

**Keywords** – mushroom, Random Forest, KNN, Support Vector Machines, Logistic Regression, Linear SVC, Stochastic Gradient Descent, Naive Bayes, Perceptron, classification

## I. INTRODUCTION

Mushroom is fleshy and edible fruit bodies of several species of fungi members that usually grow in ground surface. The number of species of mushroom that has been known until now is less than 69.000 out of the estimation of 1.500.000 species in the world and in Indonesia, there are less than 200.000 species. This million species of mushroom, in general, can be divided into two types, namely edible and poisonous mushrooms. The Family of Mushrooms wildly live in the open spaces; both with various shapes, colors, and characteristics which are not known by many people are poisonous. The Family of poisonous Mushrooms can cause illness for one who consumes and also can cause death. These mushrooms that are living wildly can be consumed and even used as medicines. Hence, it is necessary to classify them as edible or poisonous.

## II. BACKGROUND

1. Random Forest Classification Algorithm:

This comes under the category of supervised classification algorithm. In this algorithm, a forest is created with a number of trees. And, the more the trees, the more robust the forest. The forest that is built is an ensemble

of Decision trees, which is trained with the 'bagging' method most of the times. The biggest advantage of Random Forest is that it can be used for both classification and regression problems as well. Random forest adds an extra randomness to the model while growing the trees. Instead of searching for the most importantly necessary feature while splitting a node, it searches for the best feature among a random subset of features. As a result of this, there is a wide variety obtained that generally results in a better model.

2. KNN:

KNN which stands for k nearest neighbours algorithm is a non parametric methodology used for classification and regression. And, in both the cases it takes the input as k closest training examples in the feature space. And the output is obtained depending on whether it was used for classification or regression. In k-NN , the output obtained is class membership. An object gets classified by a plurality vote of its neighbours, with the object getting assigned to the class most common among its k nearest neighbours. The output is the property value for the object. This value is average of the values obtained by k nearest neighbours.

3. Support Vector Machine:

Support Vector Machine is a kind of discriminative classifier, which is defined by a separating hyperplane. When the training data are labelled, the algorithm outputs an optimal hyperplane which then categorises new examples. In the case of a two dimensional space, this hyperplane is a line which divides a plane in two parts where in each class lay in either side.

4. Logistic Regression:

Logistic Regression is a kind of classification algorithm which is used to assign observations to a discrete set of classes. However, unlike linear regression which outputs continuous number values, logistic regression transforms the obtained output using the logistic sigmoidal function in order to return a probability value which can then be mapped to two or more discrete classes. Logistic Regression can be binary, multi, ordinal. Using the knowledge of sigmoid functions and the decision boundaries, a prediction function can be written. A prediction function in logistic regression returns the probability of our observation being positive.

5. Linear SVC:

Linear SVC is an extension of Support Vector Machines and it is very widely used. The main objective of a Linear SVC is to fit to the data provided, returning a 'best fit' hyperplane which divides or categories the data available. Then, after getting the hyperplane, the features can be feeded to the classifier in order to see what the ' predicted' class is.

6. Stochastic Gradient Descent:

Stochastic gradient descent is also known as incremental gradient descent. It is an iterative method in order to optimise a differentiable objective function, a stochastic approximation of gradient descent optimisation. Stochastic Gradient Descent uses a batch sixe of 1 per iteration. However it is noisy.

7. Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is a family of algorithms where all of them share a common principle, i.e pair of features being classified is independent of each other. The major fundamental of Naive Bayes assumption is that each feature makes an independent and equal contribution to the outcome. Bayes Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

8. Perceptron

Algorithm used for learning a linear threshold function is called as Perceptron algorithm. It is an algorithm developed primarily for supervised learning of binary classifiers. A binary classifier is a function which can decide whether or not an input, represented by a vector of numbers, belong to some specific class. It is a kind of linear classifier, which means a classification algorithm that makes its predictions based on a linear predictor function, combining a set of weights with the feature vector.

## III. Literature Survey

A numerous watermarking algorithms have been discussed and proposed in either spatial and frequency domain.

[1]. Indonesia has 13% types of mushroom on the planet, anyway there is restricted data on deciding eatable or harmful mushroom. In one of the works, there is an examination done between three characterization calculations, Decision Tree(C4.5), Naive Bayes and Support Vector Machine. The investigation technique utilized is try different things with helped device of WEKA that has been trying in the examination of the three calculations. The consequences of the testing signified that C4.5 calculation has a similar exactness level to the SVM by 100%, in any case, with regards to part of speed, C4.5 calculation works quicker than SVM.

[2]. For AI applications, characterization is certainly the initial phase in gathering , partitioning, categorisation, and detachment of dataset which depends on highlight vectors. A large number of the calculations are in execution for characterization of datasets that incorporates Bayes, lethargic, capacities, meta tree and standard classifiers. In this work, diverse classifiers calculation in particular Naive Bayes, Multilayer perceptron Instance based K-Nearest Neighbors (IBK), J48 Decision Tree, Simple Cart, ZeroR, CVParameter and Filtered Classifier execution is broke down. All the recognized grouping calculations are dissected and looked at as far as arrangement exactness and execution time under various datasets.

[3]. Learning vector quantization (LVQ) is a regulated neural system strategy material in non-straight detachment issues and generally utilized for information characterization. Existing LVQ calculations are generally centered around numerical information. This paper shows a bunch type LVQ calculation utilized for characterizing information with straight out qualities. The cluster learning rules make conceivable to develop the learning strategy for information in unmitigated nonvector spaces. Trials on UCI informational collections exhibit the proposed calculation is viable to improve the ability of standard LVQ to deal with information with all out qualities

[4]. Mushrooms have high advantages in the human body. In any case, not all mushrooms are palatable. While some have medicinal properties to fix malignant growth, some different sorts of mushrooms may contain infections that convey irresistible maladies. This paper is set to ponder mushroom social highlights, for example, the shape, surface and shade of the top, gill and stalk, just as the scent, populace and natural surroundings of the mushrooms. The Principal Component Analysis (PCA) calculation is utilized for choosing the best highlights for the arrangement analyze utilizing Decision Tree (DT) calculation. The grouping precision, coefficient metric, and time taken to construct an arrangement display on a standard Mushroom dataset were estimated. The conduct highlight of 'smell' was chosen as the most astounding positioned include that add to the high grouping precision.

[5]. Mushroom chasing is winding up progressively well known as a relaxation movement, it is foremost that we have some method for arranging them as harmful or non-toxic. Utilizing administered AI models on the dataset that UCI makes accessible of different attributes of mushrooms, we can get a forecast framework that can group mushrooms. The motivation of this venture is to comprehend which AI models work best on the dataset and which highlights are most characteristic of toxic mushrooms.

[6]. Scoring frameworks are direct arrangement models that just expect clients to include, subtract and increase a couple of little numbers so as to make a forecast. These models are in across the board use by the restorative network, yet are hard to gain from information since they should be precise and scanty, have coprime whole number coefficients, and fulfill various operational limitations. We present another strategy for making information driven scoring frameworks called a Supersparse Linear Integer Model (SLIM). Thin scoring frameworks are worked by utilizing a whole number programming issue that legitimately encodes proportions of precision (the 0– 1 misfortune) and sparsity (the ℓ0-seminorm) while confining coefficients to coprime whole numbers. Thin can flawlessly fuse a wide scope of operational requirements identified with exactness and sparsity, and can deliver satisfactory models without parameter tuning in light of the immediate control gave over these amounts. We give limits on the testing and preparing exactness of SLIM scoring frameworks, and present another information decrease strategy that can improve versatility by disposing of a bit of the preparation information in advance. Our paper incorporates results from a cooperation with the Massachusetts General Hospital Sleep Laboratory, where SLIM is being utilized to make an exceptionally custom fitted scoring framework for rest apnea screening.

[7]. We consider regulated learning with irregular choice trees, where the tree development is totally arbitrary. The strategy was utilized as a heuristic functioning admirably practically speaking in spite of the straightforwardness of the setting, yet with no hypothetical certifications. The objective of this paper is to reveal new insight into the whole worldview. We furnish solid hypothetical assurances in regards to learning with irregular choice trees. We present and think about three distinct variations of the calculation that have insignificant memory necessities: lion's share casting a ballot, limit averaging and probabilistic averaging. The irregular structure of the tree empowers us to adjust our setting to the differentially-private situation in this way we likewise propose differentially-private renditions of each of the three plans. We give upper limits on the speculation blunder and scientifically clarify how the precision relies upon the quantity of arbitrary choice trees. Besides, we demonstrate that just logarithmic number of autonomously chosen arbitrary choice trees get the job done to effectively group the vast majority of the information, notwithstanding when differential-protection ensures must be kept up. Such an investigation has never been finished. We exactly demonstrate that larger part casting a ballot and limit averaging give the best exactness, likewise for moderate clients requiring high protection ensures. Specifically, a straightforward dominant part casting a ballot rule, that was not considered before with regards to differentially-private learning, is a particularly decent contender for the differentially-private classifier since it is considerably less delicate to the decision of backwoods parameters than different strategies.

[8]. The Waikato Environment for Knowledge Analysis (Weka) is a complete suite of Java class libraries that execute many cutting edge AI and information mining calculations. Weka is unreservedly accessible on the World-Wide Web and goes with another content on information mining [1] which records and completely clarifies every one of the calculations it contains. Applications composed utilizing the Weka class libraries can be kept running on any PC with a Web perusing ability; this enables clients to apply AI strategies to their very own information paying little heed to PC stage. Apparatuses are accommodated pre-preparing information, encouraging it into an assortment of learning plans, and dissecting the subsequent classifiers and their execution. A critical asset for exploring through Weka is its on-line documentation, which is naturally produced from the source. The essential learning strategies in Weka are classifiers, and they initiate a standard set or choice tree that models the information.

[9]. This investigation centers around the utilization of information mining procedures to break down a recently acquired informational collection. The examination will likewise expand past research at Pace University into the employments of a human-machine interface to build the precision of AI. To this end, the investigation will utilize an ostensible informational index, the Mushroom Database, and the information mining apparatus Weka. Different information mining calculations are utilized against the Mushroom Database, including an unpruned choice tree, a casted a ballot perceptron calculation, a covering calculation that creates just right standards, and the closest neighbor classifier. At long last, an unpruned tree is utilized to build up a human-machine intelligent application.

[10]. Grouping is one of the utilizations of feed-forward Artificial Neural Network (ANN). Characterization can delineate to predefined classes or gatherings. It is alluded to as a regulated learning, on the grounds that before inspecting information the classes are constantly decided. Multi-Layer Perception, is an administered nonpartisan systems demonstrate that is use to prepare and test information to assemble a model. In this investigation. Multi-Layer Perception is utilized to prepare the Data set to deliver a model to make forecast of grouping .After setting up the Mushrooms information for preparing, just 8124 of dataset occurrences used to be train. Programming used to mining information in this task is Neural Connection Version 2.0. This report, for the most part clarifying the Classification, Multi-Layer Preceptor, Back proliferation, Mushrooms, and subtleties on the mining action done to the chose datasets, to decide if Mushroom's quality is palatable or Poison..

[11]. Picking the wild mushrooms from the wild and woodlands for nourishment reason or for no particular reason has turned into an open issue inside the most recent years on the grounds that numerous sorts of mushrooms are toxic. Appropriate assurance of mushrooms is one of the key wellbeing issues in picking exercises of it, which is generally spread, in nations. This commitment proposes a novel way to deal with help assurance of the mushrooms through utilizing a proposed framework with cell phones. Some portion of the proposed framework is a versatile application that effectively utilized by a client - mushroom picker. Subsequently, the mushroom type assurance procedure can be performed at any area dependent on explicit characteristics of it. The mushroom type assurance application keeps running on Android gadgets that are generally spread and sufficiently modest to empower wide misuse by clients. This paper created Mushroom Diagnosis Assistance System (MDAS) that can be utilized on a cell phone. Two classifiers are utilized which are Naive Bays and Decision Tree to characterize the mushroom types. The proposed methodology chooses the best of the definitely realized mushroom qualities, and after that indicate the mushroom type. The utilization of explicit highlights in mushroom assurance process accomplished precise outcomes.

[12]. Information Mining is the programmed look for fascinating and valuable connections between characteristics in databases. One noteworthy obstruction to powerful Data Mining is the size and unpredictability of the objective database, both as far as the list of capabilities and the example set. While many Machine Learning calculations have been connected to Data Mining applications. There has been specific enthusiasm for the utilization of Genetic Algorithms (GAs) for this reason because of their accomplishment in vast scale inquiry and enhancement issues. This standard per investigates how GAs are being utilized to improve the execution of Data Mining grouping and arrangement calculations and analyzes techniques for improving these methodologies.

[13]. The examination work presents K-modes bunching calculation in taking care of Data digging issues for agro-based dataset. Mushroom informational index accessible from the UCI information archive were broke down to recognize diverse blends of characteristics that are critical in gathering the mushroom information as harmful or consumable. Bunching is unsupervised discovering that goes for parceling an informational collection into gatherings of comparable things. The objective is to make groups of information objects where the inside bunch similitude is amplified (intra-group likeness) and the between-bunch closeness is limited (between bunch comparability). K-modes grouping calculation was utilized to bunch mushroom datasets. The framework was created utilizing Java programming language and Object Oriented Methodology (OOM) was connected in light of the fact that, OOM advances effortlessness, reusability, unwavering quality and increment the improvement speed. The bunched consequence of mushroom dataset depicted theoretical examples comparing to 23 types of gilled mushrooms in the Agaricus and Lepiota family. Every specie was recognized as unquestionably consumable, certainly harmful or of obscure edibility and not suggested. What's more, we likewise locate the most toxic Trichoderma green that form disease in palatable basidiomycetes. It has been known to have the capacity to cause a drastical decline underway or even whole harvests can be cleared out. The Nigerian Agri-Business is by and by searching for scientists to accomplice so as to advance agro-based items free from any sort of deformities, which the paper attempted to accomplish.

[14]. This examination will concentrate on the utilization of Data Mining procedures on recently investigated informational collections. The information mining device Weka will be utilized. Weka represents Waikato

condition for information investigation, and "is a well known suite of AI programming written in Java, created at the University of Waikato. WEKA is free programming accessible under the GNU General Public License". The motivation behind the investigation is to broaden past examinations by running new informational collections of stylometry, keystroke catch, and mouse development information through Weka utilizing different information mining calculations. The examination will likewise expand past research at Pace University into the employments of a human-machine interface to build the exactness of AI. To this end, the investigation will utilize an ostensible informational collection, the Mushroom Database

[15]. This examination will concentrate on the utilization of Data Mining procedures on recently investigated informational indexes. The information mining device Weka will be utilized. Weka represents Waikato condition for information examination, and "is a mainstream suite of AI programming written in Java, created at the University of Waikato. WEKA is free programming accessible under the GNU General Public License". The reason for the investigation is to broaden past examinations by running new informational indexes of stylometry, keystroke catch, and mouse development information through Weka utilizing different information mining calculations. The examination will likewise expand past research at Pace University into the employments of a human-machine interface to build the precision of AI. To this end, the examination will utilize an ostensible informational collection, the Mushroom Database

[16]. Formal Concept Analysis (FCA) is a developing information innovation that has applications in the visual investigation of huge scale information. Be that as it may, informational indexes are regularly excessively huge (or contain such a large number of formal ideas) for the subsequent idea grid to be decipherable. This paper supplements existing work here by portraying two techniques by which valuable and sensible grids can be gotten from substantial informational collections. This is accomplished however the utilization of a lot of uninhibitedly accessible FCA apparatuses: the setting maker FcaBedrock and the idea excavator In-Close, that were created by the creators, and the cross section manufacturer ConExp. In the primary strategy, a sub-setting is created from an informational collection, offering ascend to a discernible grid that centers around qualities of intrigue. In the second strategy, a setting is dug for 'extensive' ideas which are then used to re-compose the first setting, in this manner diminishing 'clamor' in the unique circumstance and offering ascend to an intelligible cross section that clearly depicts a reasonable diagram of the substantial arrangement of information it is gotten from.

[17]. This paper presents grouping procedures for breaking down mushroom dataset. Fake Mushroom dataset is made out of records of various kinds of mushrooms, which are palatable or non-eatable. Aritificial Neural Network and Adaptive Nuero Fuzzy deduction framework are utilized for execution of the grouping strategies. Diverse strategies utilized for grouping like ANN, ANFIS and Naïve Bayes are utilized to order extraordinary mushrooms as consumable or non-palatable. The execution of the distinctive systems is assessed utilizing exactness, MAE, kappa measurement. In the wake of examining the outcomes it was discovered that Adaptive Nuero Fuzzy derivation System outflanked different strategies with most elevated exactness, least mean outright mistake and ANN is the second best entertainer. In the event that size of preparing set is expanded, the precision additionally expanded as for preparing set.

[18]. Information mining assumes an essential job in our every day life period. Every one of the information has been digitalized so we have to break down it to make helpful data for our insight. Order and Clustering are the two essential real methods utilized for extricating the information from the database. Bunching is known as the unsupervised realizing which is segment a dataset in to a gathering by their similitudes. The goal of this paper is to assess the execution of various grouping calculation, for example, Expectation Maximization (EM), Farthest Fast and K-implies by accurately bunched occurrences and time taken to construct the model for mushroom dataset utilizing information mining instrument WEKA (Waikato condition for Knowledge Analysis). The mushroom dataset comprises of 8124 occurrences and 22 properties with two classes whether it is palatable or harmful. The dataset is gathered from the UCI AI archive.

[19]. In this paper, an information digging application is presented for choosing exceedingly powerful factors or side effect of various infection finding in mushroom yield. The investigation additionally centers around a few variables causing a particular mushroom infection. Exceedingly potential manifestations among a few variables were engaged out for better administration in such manner. That is the reason information mining systems are being utilized for positioning among side effects. This paper centers around recognizing explicit illnesses among a few infections utilizing an information mining grouping based methodologies. Genuine information has been taken from mushroom ranch and from there on sanitization of potential elements is done through information mining approaches. The arrangement method and sickness forecast of mushroom dataset were readied utilizing Naïve Bayes, SMO and RIDOR calculations. A measurable examination has been created so as to locate the best indications required for mushroom sickness analysis. Other than this, it looks through the best performing arrangement calculation among all.

[20]. Extraction of information in country information is a trying undertaking, from discovering plans likewise, associations additionally, elucidation. In solicitation to procure possibly interesting structures likewise, associations from this information, it is consequently imperative that a procedure be made additionally, exploit the arrangements of existing techniques likewise, instruments open for information mining additionally, information very in databases. Information mining is moderately another methodology in the field of horticulture. Exact information in depicting crops relies upon climatic, geological, common likewise, different components. These are exceptionally basic contributions to make depiction likewise, desire models in information mining. In this examination, a powerful information mining system dependent on kNN is investigated, presented additionally, executed to depict country crops. The methodology attracts moves up to demand issues by using Principal Components Analysis (kNN) as a pre getting ready technique additionally, a changed Genetic Algorithm (GA) as the limit streamlining agent. The health limit in GA is changed appropriately using powerful partition measure.

The following table suggest the diverse applications in watermarking scheme available offered by various techniques.

| **Name of technique** | **Domain** | **Host image** | **Advantages of techniques** | **Disadvantages of techniques** |
|---|---|---|---|---|
| Classification Algorithm for Edible Mushroom Identification using C4.5 | Spatial | colour | The algorithm is used for building smaller or larger, more accurate decision trees and the algorithm is quite time efficient. | The traditional k-means algorithm does not perform well for categorical data sets where there is natural ordering among values. K is a positive integer number. A small variation in data can lead to different decision trees when using C4.5.For a small training set, C4.5 does not work very well. |
| Performance analysis of Classification Algorithms under Different Datasets, using multilayer perceptron network, IBK, Naive Bayes | spatial | colour | The MLP algorithm is a very good algorithm to use for the regression and mapping. It can be used to map an N-dimensional input signal to an M-dimensional output signal, this mapping can also be non-linear. | The main limitation of the MLP algorithm used in mushroom analysis earlier, is that, because of the way it is trained, it can not guarantee that the minima it stops at during training is the global minima. |

| | | | | |
|---|---|---|---|---|
| Extending Learning Vector Quantization for Classifying Data with Categorical Values, using LVQ | spatial | colour | An advantage of LVQ is that it creates prototypes that are easy to interpret for experts in the respective application domain. LVQ systems can be applied to multi-class classification problems in a natural way. It is used in a variety of practical applications. | The LVQ network was less accurate. The Gaps have been identified and we are looking for more accurate and precise results with ensemble methods |
| Behavioural Features for Mushroom Classification, using PCA and decision tree | spatial | colour | PCA's key advantages are its low noise sensitivity, the decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimensions. A major decision tree analysis advantages is its ability to assign specific values to problem, decisions, and outcomes of each decision. Incorporation of monetary values to decision trees help make explicit the costs and benefits of different alternative courses of action. | Decision Trees provide less information on the relationship between the predictors and the response, are biased toward predictors with more variance or levels, can have issues with highly collinear predictors and can have poor prediction accuracy. |
| Mushroom Classication using Logistic Regression | spatial | colour | It is more robust, the independent variables don't have to be normally distributed, or have equal variance in each group. It does not assume a linear relationship between the IV and DV. It may handle nonlinear effects. | classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be potentially very bad |
| Supersparse linear integer models for optimized medical scoring systems | spatial | colour | The model is really parctical and interpretable and gives good results. | have issues with highly collinear predictors and can have poor prediction accuracy. |

| | | | | |
|---|---|---|---|---|
| Differentially- and non-differentially-private random decision trees | spatial | colour | A major decision tree analysis advantages is its ability to assign specific values to problem, decisions, and outcomes of each decision.<br>Incorporation of monetary values to decision trees help make explicit the costs and benefits of different alternative courses of action. | Decision Trees provide less information on the relationship between the predictors and the response, are biased toward predictors with more variance or levels, can have issues with highly collinear predictors and can have poor prediction accuracy. |
| Machine learning tools for data classification using Weka Tool. | spatial | colour | Free availability under the GNU General Public License. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. | These tools provide less accurate Predictions. |
| Data Mining on a Mushroom Database using IBK and PRISM | spatial | colour | It is quite efficient and gives good results. | Makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be potentially very bad. |
| Classifying poisonous and edible mushrooms in the Agaricus and Lepiota family using multi layer perception - | spatial | colour | The MLP algorithm is a very good algorithm to use for the regression and mapping. It can be used to map an N-dimensional input signal to an M-dimensional output signal, this mapping can also be non-linear. | The clustering process in the modified algorithm is faster |
| Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices, using Naive Bayes | spatial | colour | super simples, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. | The first disadvantage is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Another problem happens due to data scarcity. |

| | | | | |
|---|---|---|---|---|
| Application of Genetic Algorithms to Data Mining-Application on mushroom dataset | spatial | colour | Genetic algorithms search parallel from a population of points. Therefore, it has the ability to avoid being trapped in local optimal solution like traditional methods, which search from a single point.<br>Genetic algorithms use probabilistic selection rules, not deterministic ones. | No guarantee of finding global maxima. A lot of time taken for convergence. It has incomprehensible solutions. |
| K-Modes Clustering Algorithm in Solving Data Mining Problems for Mushroom Dataset | spatial | colour | If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. | Difficult to predict K-Value. With global cluster, it didn't work well. Different initial partitions can result in different final clusters. It does not work well with clusters (in the original data) of Different size and Different density |
| Data Mining: The Mushroom Database, using Naive bayes and Priori | spatial | colour | This is the most simple and easy-to-understand algorithm among association rule learning algorithms. The resulting rules are intuitive and easy to communicate to an end user. It doesn't require labeled data as it is fully unsupervised; as a result, you can use it in many different situations because unlabeled data is often more accessible.<br>The algorithm is exhaustive. | The algorithms scans the database too many time which reduces its overall performance. |
| Analysis of Large Data Sets using Formal Concept Lattices, using FCA | spatial | colour | Efficient | Existing measures and criteria , proposed here and found in literature, do not identify what aspects of knowledge get preserved. |

| | | | | |
|---|---|---|---|---|
| Mushroom Classification Using ANN and ANFIS Algorithm | spatial | colour | ANNs have the ability to learn and model non-linear and complex relationships, which is really important because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex. ANFIS also uses the ANN's ability to classify data and identify patterns. ANFIS model is more transparent to the user and causes less memorisation errors. | Performance is low when the dataset is small and its is high when it is large. |
| Clustering Techniques for Mushroom Database | Spatial | Colour | The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. | Disadvantages of clustering are complexity and inability to recover from database corruption. |
| An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach using Ripple down rule learner | spatial | colour | The algorithm is quite complex but gives average efficient results. | The certain limitations of the algorithm results in the less accuracy than expected. |
| Applying Data Mining Techniques to Evaluate Rural Crops Using kNN | spatial | colour | If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. | Difficult to predict K-Value. With global cluster, it didn't work well. Different initial partitions can result in different final clusters. It does not work well with clusters (in the original data) of Different size and Different density |

## IV. DATASET DESCRIPTION & SAMPLE DATA

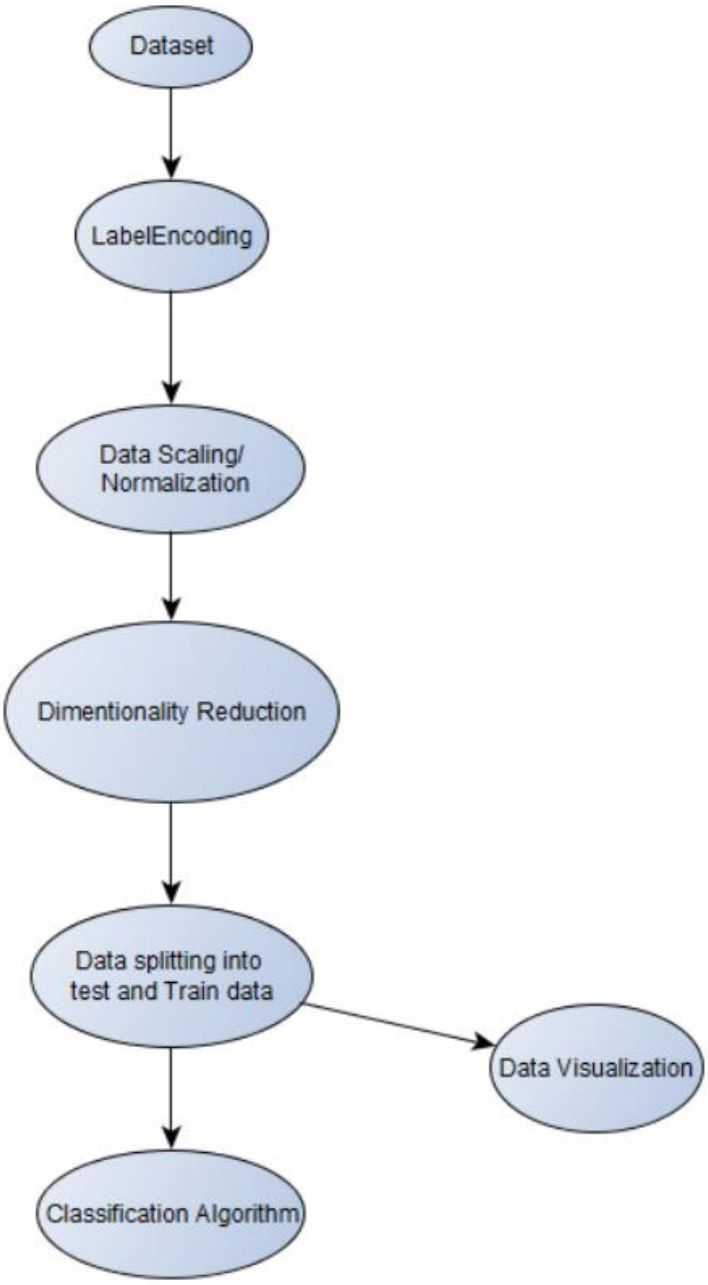
The dataset for the mushroom was acquired from the UCI repository. This dataset contains samples of mushrooms from the Agaricus and Lepiota Family and then they are classified as definitely edible, definitely poisonous or of unknown edibility and not recommended. This mushroom dataset (Table 1) contain 8124

number of instances with 22 number of attributes. There are 2 class labels where definitely edible become one class label as ‘e’ and definitely poisonous or of unknown edibility and not recommended form one class label as ‘p’. The dataset has very even class distribution with 51.8% are edible and 48.2% are poisonous.
Mushroom dataset had been split into training and testing where 90% of mushroom dataset was used for training and the remaining used for testing purpose. Only training dataset will undergo several process of analysis where testing dataset was preserved after the process in order to check either classifiers algorithm overfit or not.
cap-shape:
bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
cap-surface:
fibrous=f, grooves=g, scaly=y, smooth=s
cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
bruises: bruises=t, no=f
odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
gill-attachment:
attached=a, descending=d, free=f, notched=n
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
stalk-shape: enlarging=e, tapering=t
stalk-root:
bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
stalk-surface-above-ring:
fibrous=f, scaly=y, silky=k, smooth=s
stalk-surface-below-ring:
fibrous=f, scaly=y, silky=k, smooth=s
stalk-color-above-ring:
brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
stalk-color-below-ring:
brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
veil-type:
partial=p, universal=u
veil-color:
brown=n, orange=o, white=w, yellow=y
ring-number: none=n, one=o, two=t
ring-type:
cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
spore-print-color:
black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
population:
abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
habitat:
grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 8124 | **Area:** | Life |
| **Attribute Characteristics:** | Categorical | **Number of Attributes:** | 22 | **Date Donated** | 1987-04-27 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 380661 |

Number of Instances: 8124 Number of Attributes: 2

## V. PROPOSED ALGORITHM WITH FLOWCHART




## VI. EXPERIMENTS RESULTS

**Logistic Regession:**

```
Training results:

Accuracy Score: 0.9179

Classification Report:
             precision    recall  f1-score   support

          0       0.90      0.94      0.92      2951
          1       0.94      0.89      0.91      2735

avg / total       0.92      0.92      0.92      5686
```

```
Test results:

Accuracy Score: 0.9135

Classification Report:
             precision    recall  f1-score   support

          0       0.89      0.94      0.92      1257
          1       0.94      0.88      0.91      1181

avg / total       0.91      0.91      0.91      2438


Confusion Matrix:
[[1187   70]
 [ 141 1040]]
```

**Support Vector Machine:**

```
Training results:

Accuracy Score: 1.0000

Classification Report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      2951
          1       1.00      1.00      1.00      2735

avg / total       1.00      1.00      1.00      5686


Confusion Matrix:
[[2951    0]
 [   0 2735]]

Average Accuracy:       0.9993

Standard Deviation:     0.0016

 Test results:

 Accuracy Score: 1.0000

 Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1257
           1       1.00      1.00      1.00      1181

 avg / total       1.00      1.00      1.00      2438


 Confusion Matrix:
 [[1257    0]
  [   0 1181]]
```

**KNN:**

```
    Training results:

    Accuracy Score: 0.9996

    Classification Report:
                 precision    recall  f1-score   support

              0       1.00      1.00      1.00      2951
              1       1.00      1.00      1.00      2735

    avg / total       1.00      1.00      1.00      5686


    Confusion Matrix:
    [[2951    0]
     [   2 2733]]

    Average Accuracy:       0.9991
Test results:

Accuracy Score: 0.9996

Classification Report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      1257
          1       1.00      1.00      1.00      1181

avg / total       1.00      1.00      1.00      2438


Confusion Matrix:
[[1257    0]
 [   1 1180]]
```

**GaussianNB**

```
Training results:

Accuracy Score: 0.9260

Classification Report:
             precision    recall  f1-score   support

          0       0.91      0.95      0.93      2951
          1       0.94      0.90      0.92      2735

avg / total       0.93      0.93      0.93      5686


Confusion Matrix:
[[2795  156]
 [ 265 2470]]

Average Accuracy:       0.9261

Standard Deviation:     0.0051

  Test results:

  Accuracy Score: 0.9192

  Classification Report:
               precision    recall  f1-score   support

            0       0.90      0.94      0.92      1257
            1       0.94      0.89      0.91      1181

  avg / total       0.92      0.92      0.92      2438


  Confusion Matrix:
  [[1187   70]
   [ 127 1054]]
```

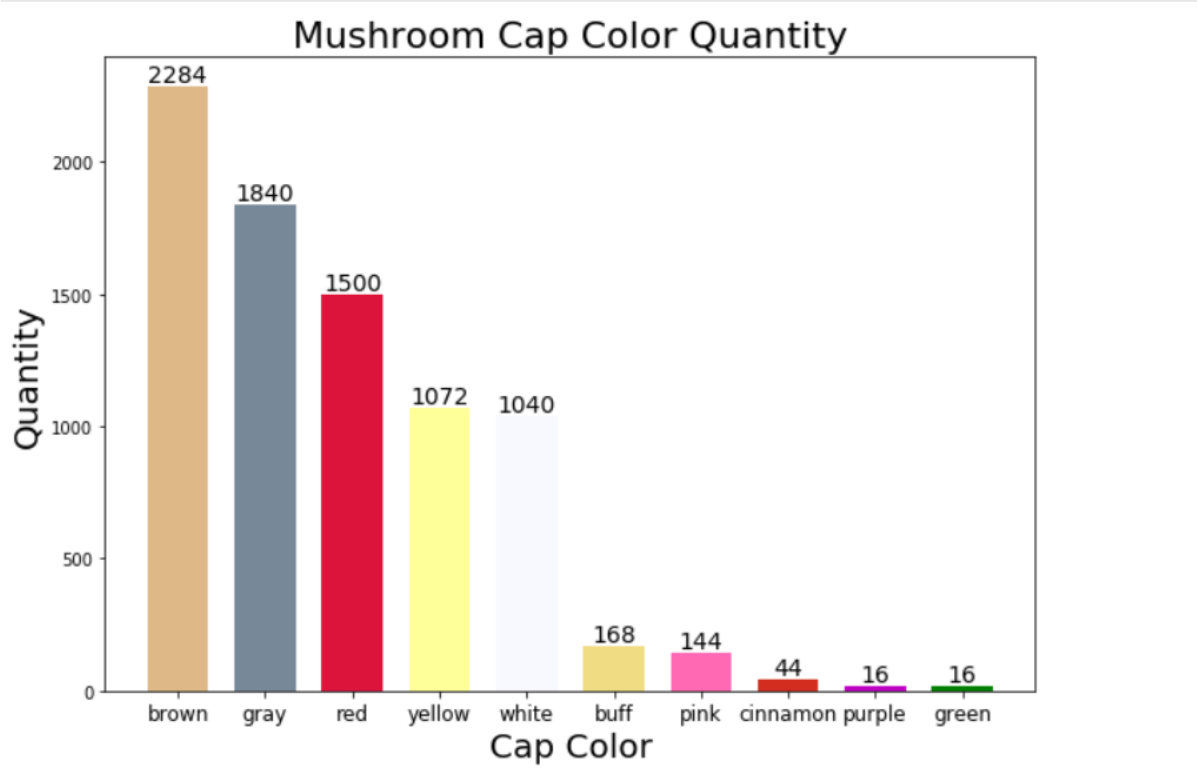
**Decision Tree**

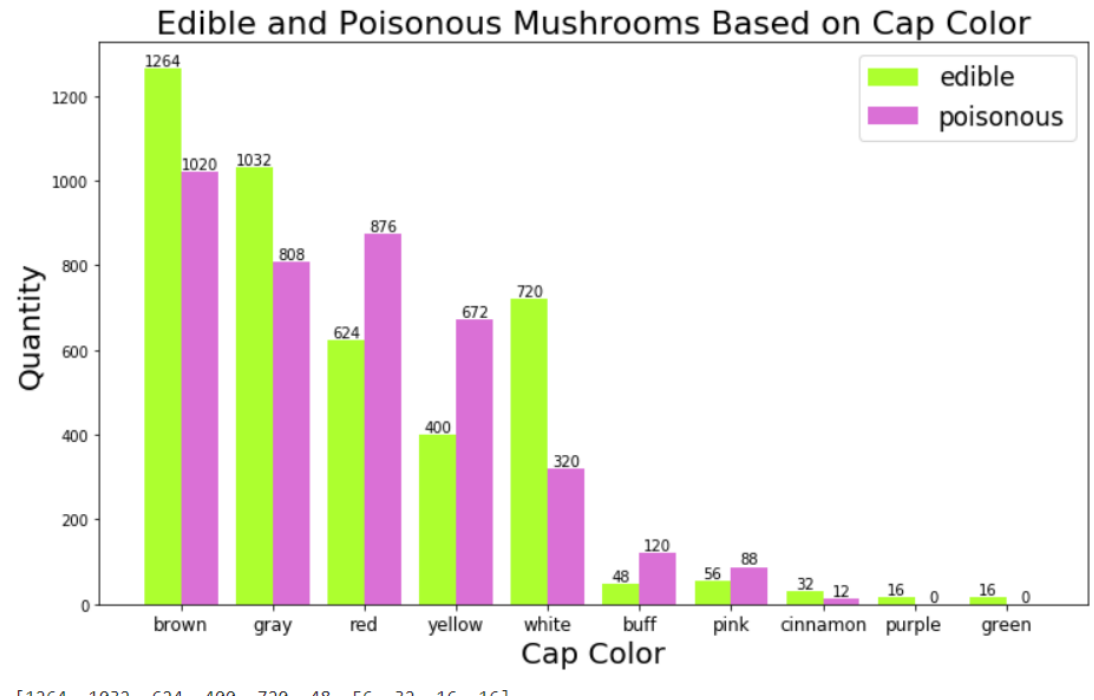
```
Training results:

Accuracy Score: 1.0000

Classification Report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      2951
          1       1.00      1.00      1.00      2735

avg / total       1.00      1.00      1.00      5686


Confusion Matrix:
[[2951    0]
 [   0 2735]]

Average Accuracy:       0.9949

Standard Deviation:     0.0045


Test results:

Accuracy Score: 0.9943

Classification Report:
             precision    recall  f1-score   support

          0       1.00      0.99      0.99      1257
          1       0.99      1.00      0.99      1181

avg / total       0.99      0.99      0.99      2438


Confusion Matrix:
[[1247   10]
 [   4 1177]]
```

**Random Forest**

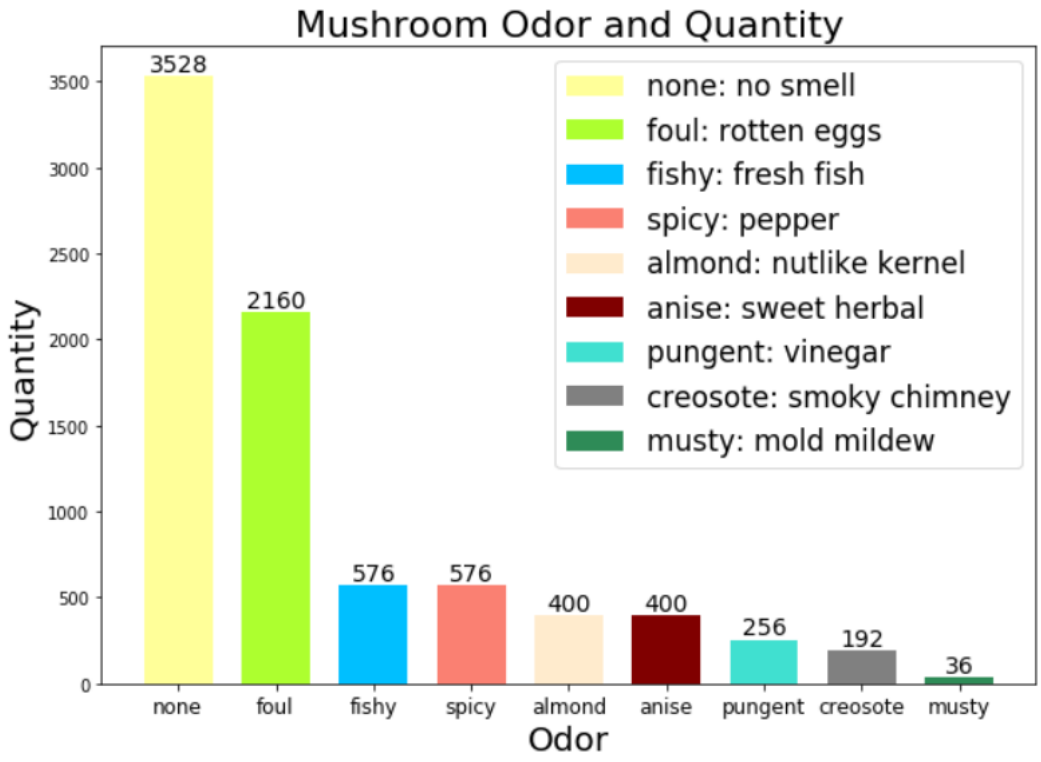
```
Training results:

Accuracy Score: 1.0000

Classification Report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      2951
          1       1.00      1.00      1.00      2735

avg / total       1.00      1.00      1.00      5686


Confusion Matrix:
[[2951    0]
 [   0 2735]]

Average Accuracy:       0.9996

Standard Deviation:     0.0007


Test results:

Accuracy Score: 1.0000

Classification Report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      1257
          1       1.00      1.00      1.00      1181

avg / total       1.00      1.00      1.00      2438


Confusion Matrix:
[[1257    0]
 [   0 1181]]
```

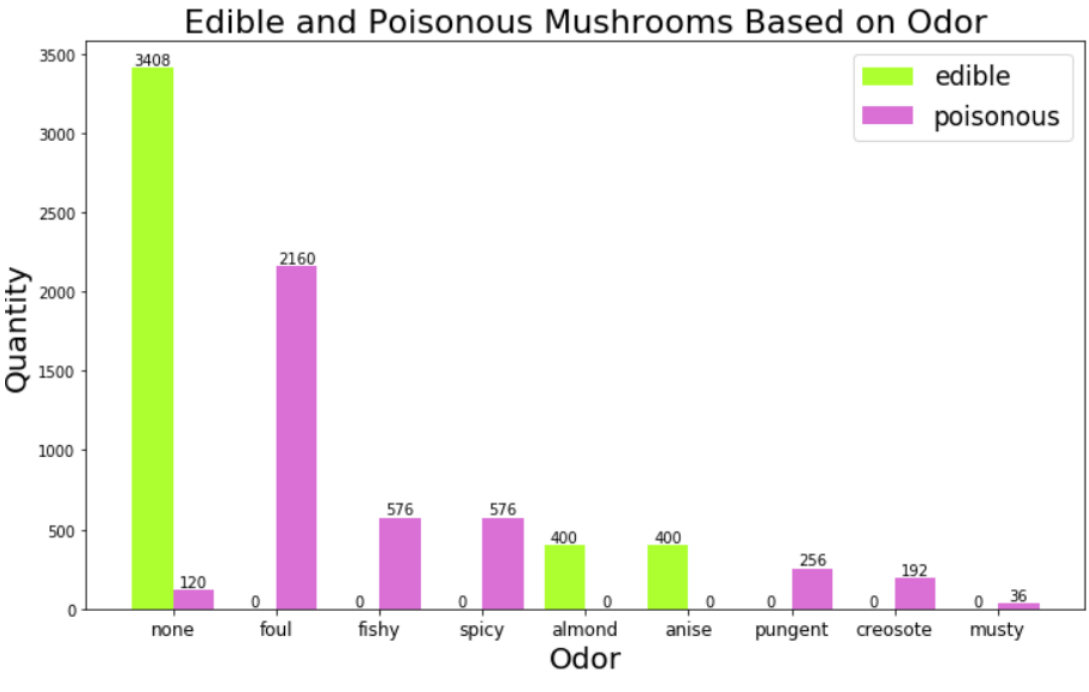
**Total Mushrooms for each cap color**



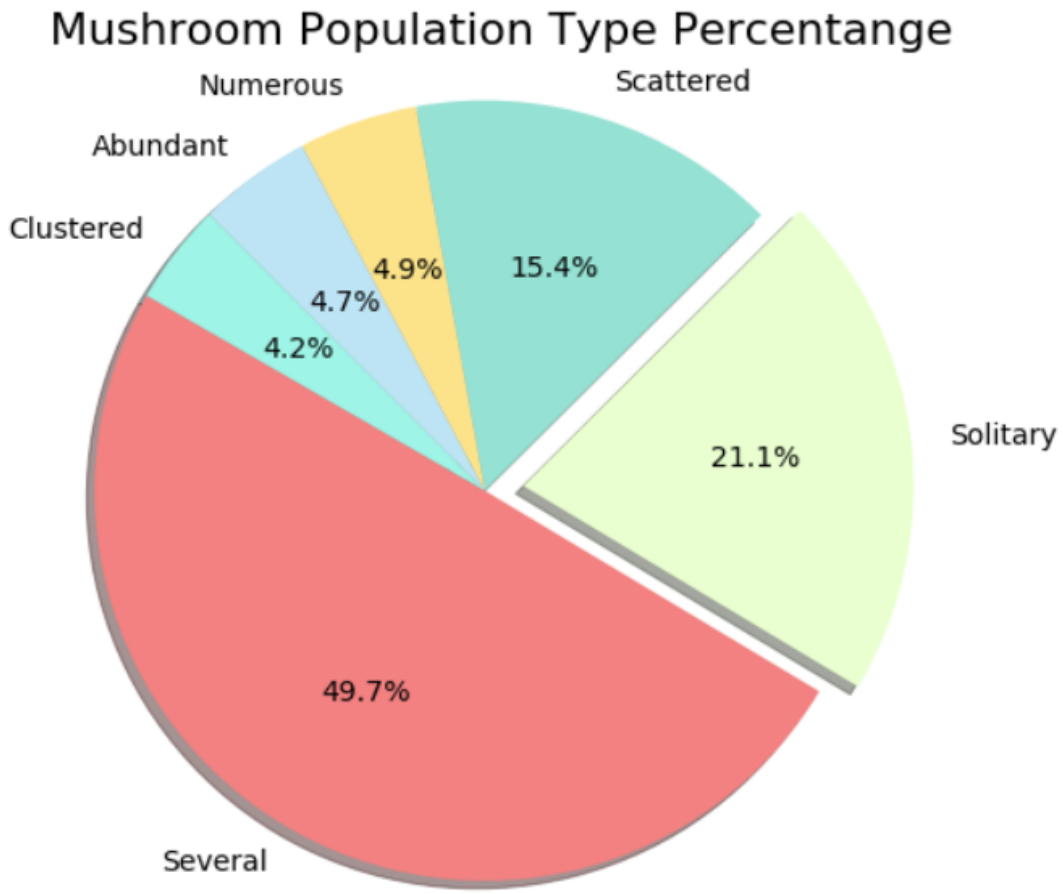



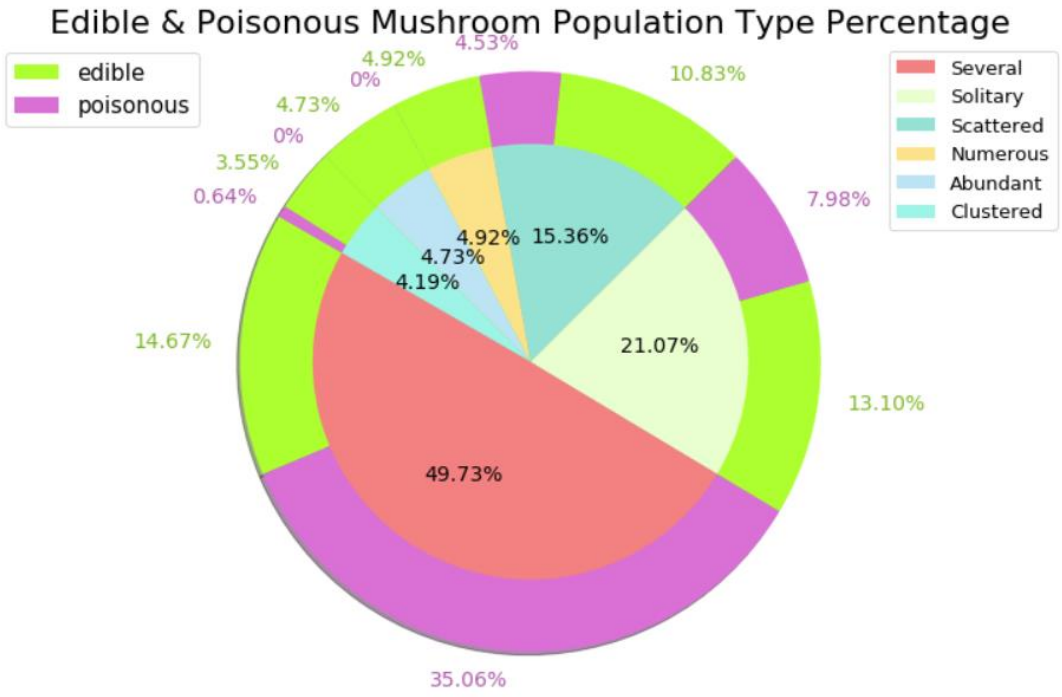
**Total mushrooms for each odor type**



**Total percentage of mushrooms for each population category**

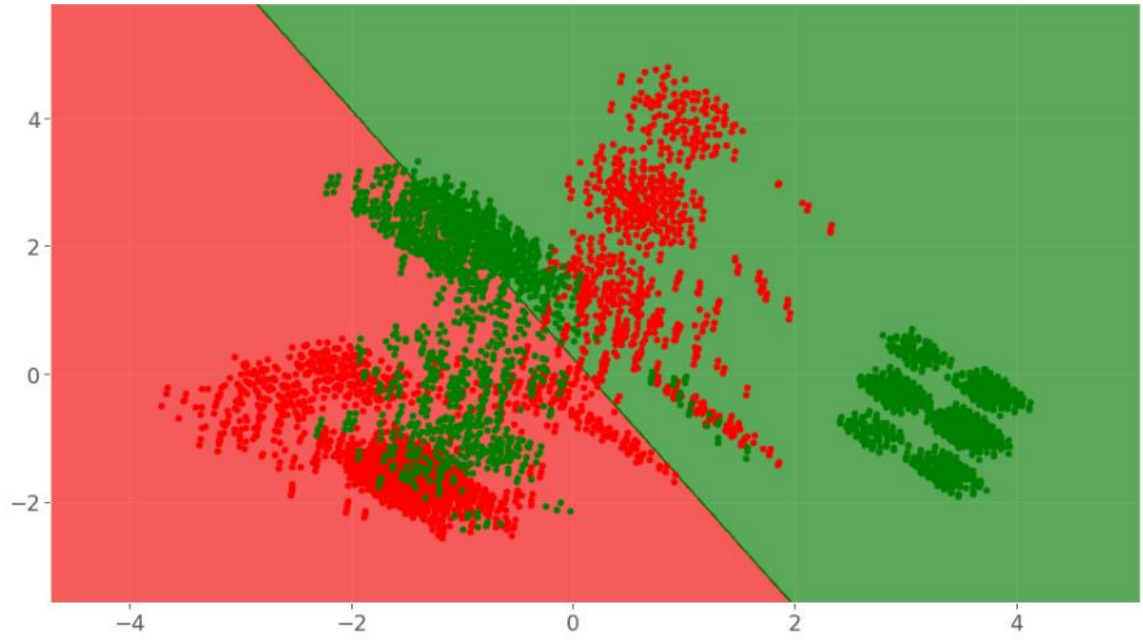


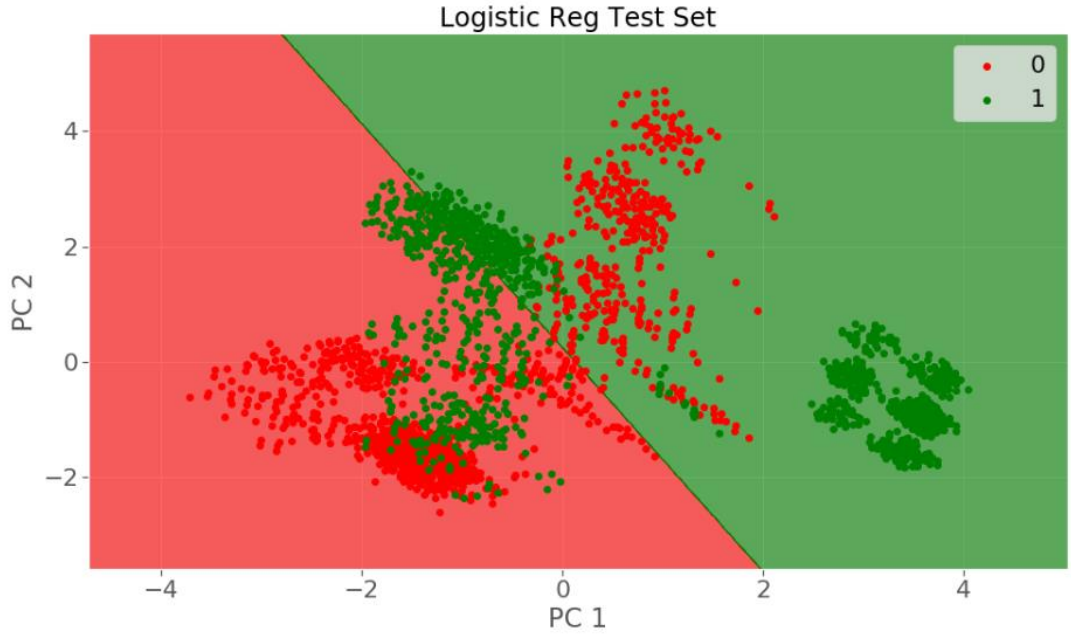
**Training and Testing dataset graphs**

**Logistic Regression**

**Testing Data**

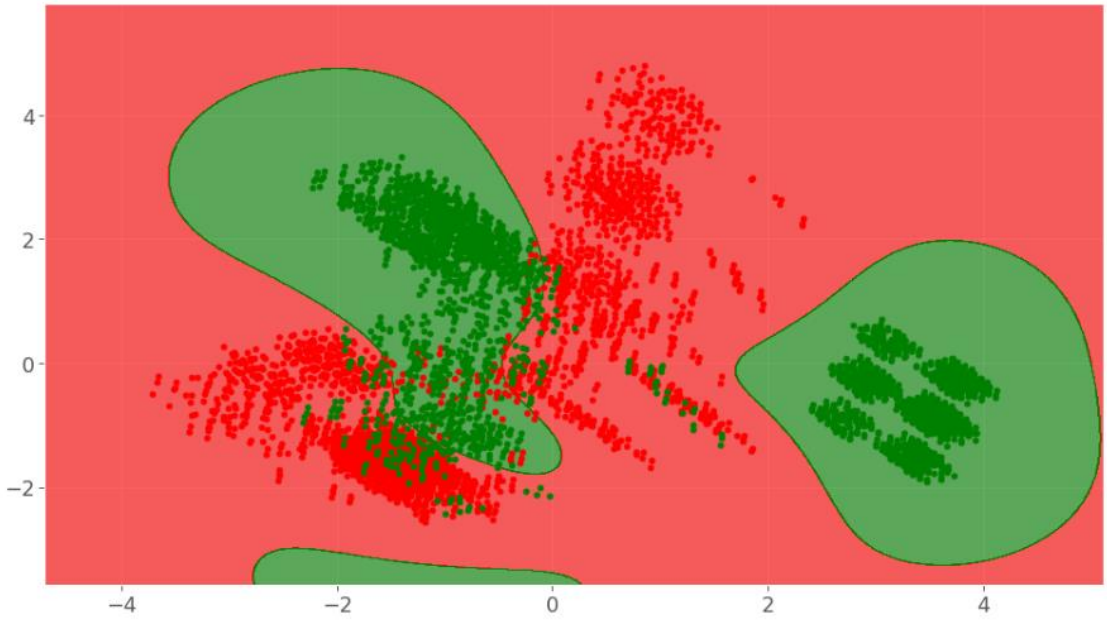



**Training Data**

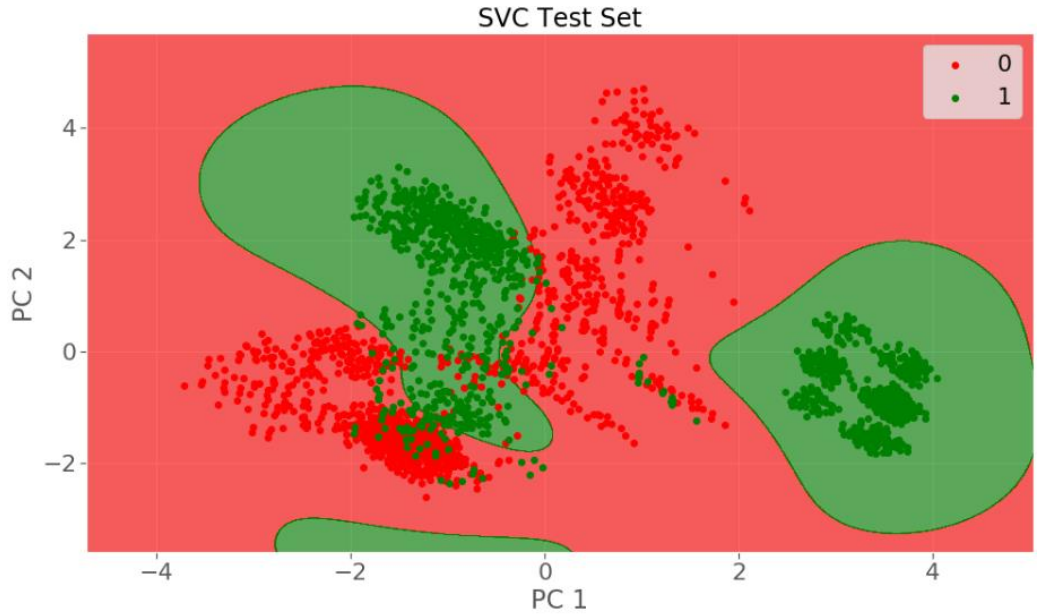


**SVC**

**Training Dataset**

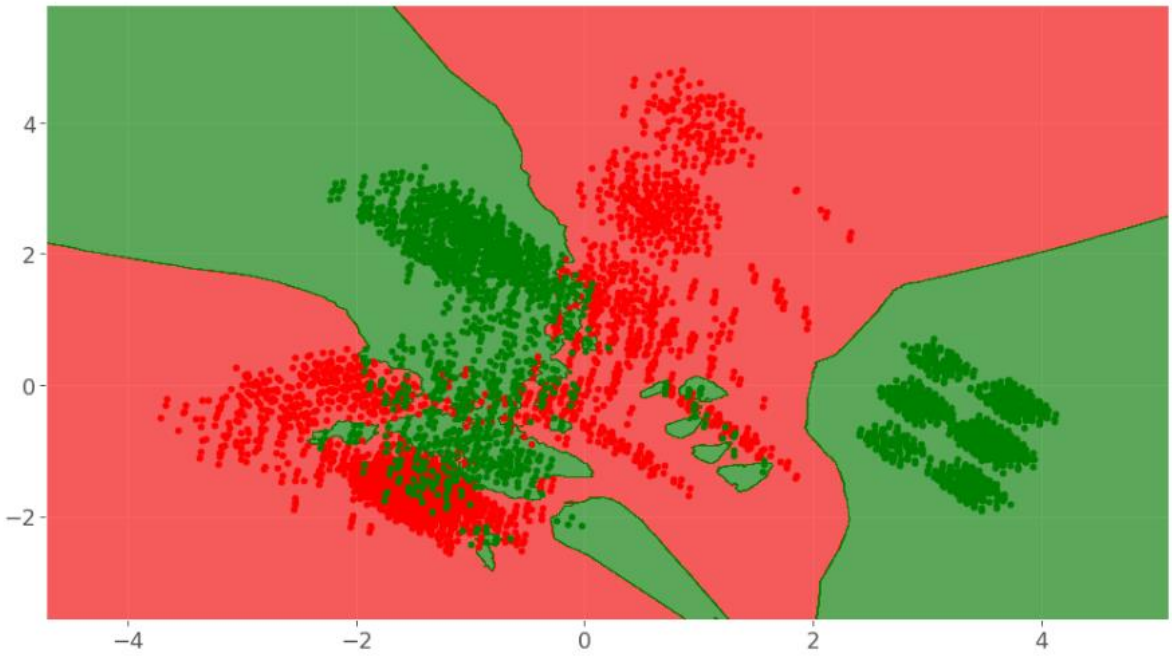



**Testing Dataset**

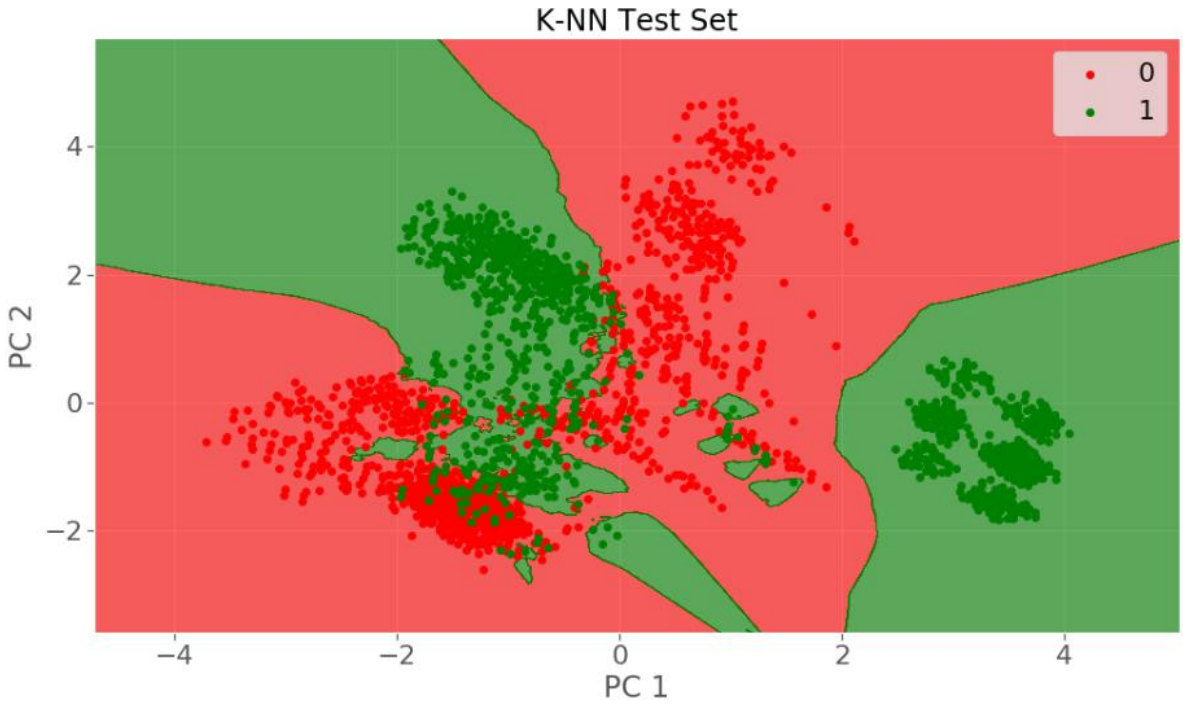


**KNN**

**Training Dataset**

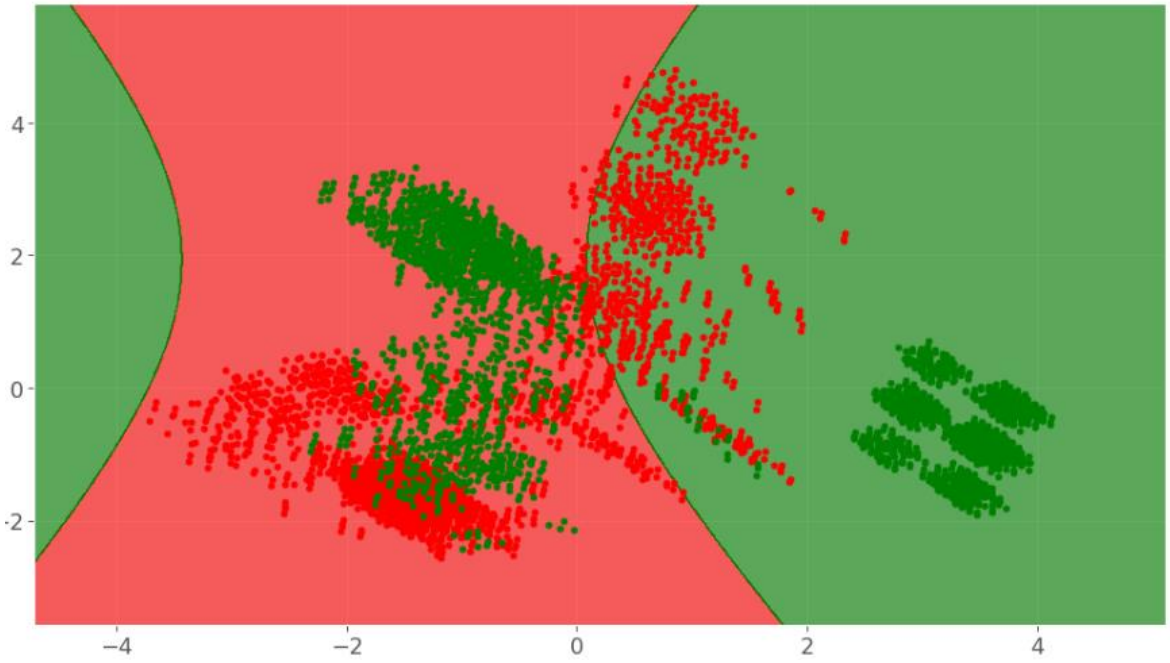



**Testing Dataset**

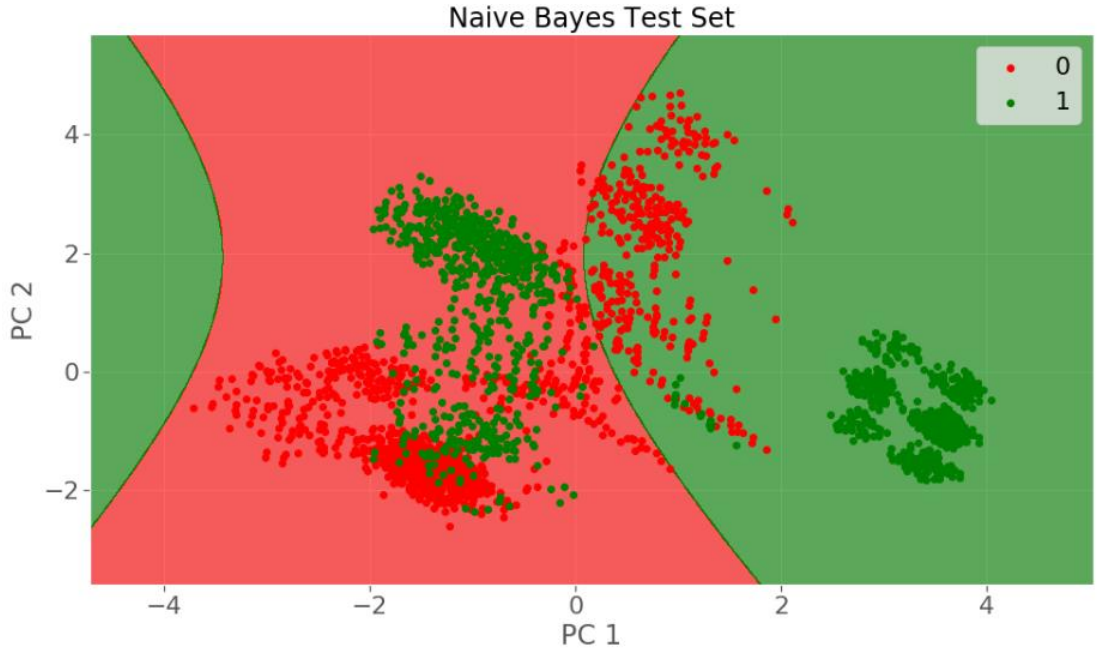


**Naïve Bayes**

**Training Dataset**

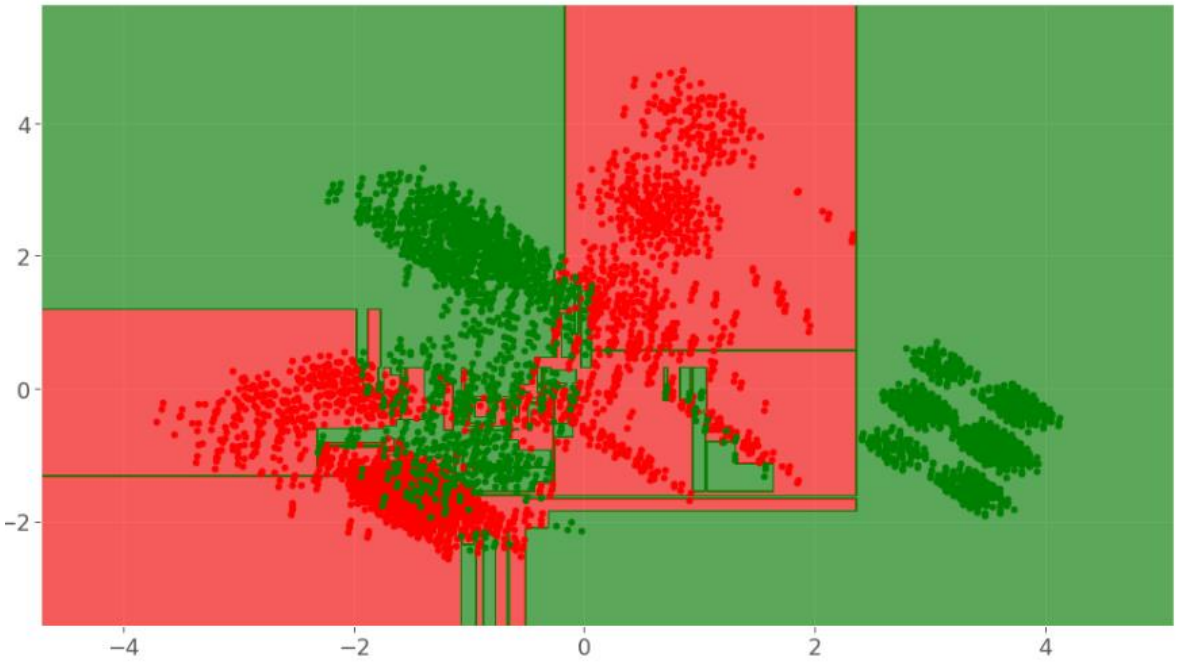



**Testing Dataset**




**Decision Tree**

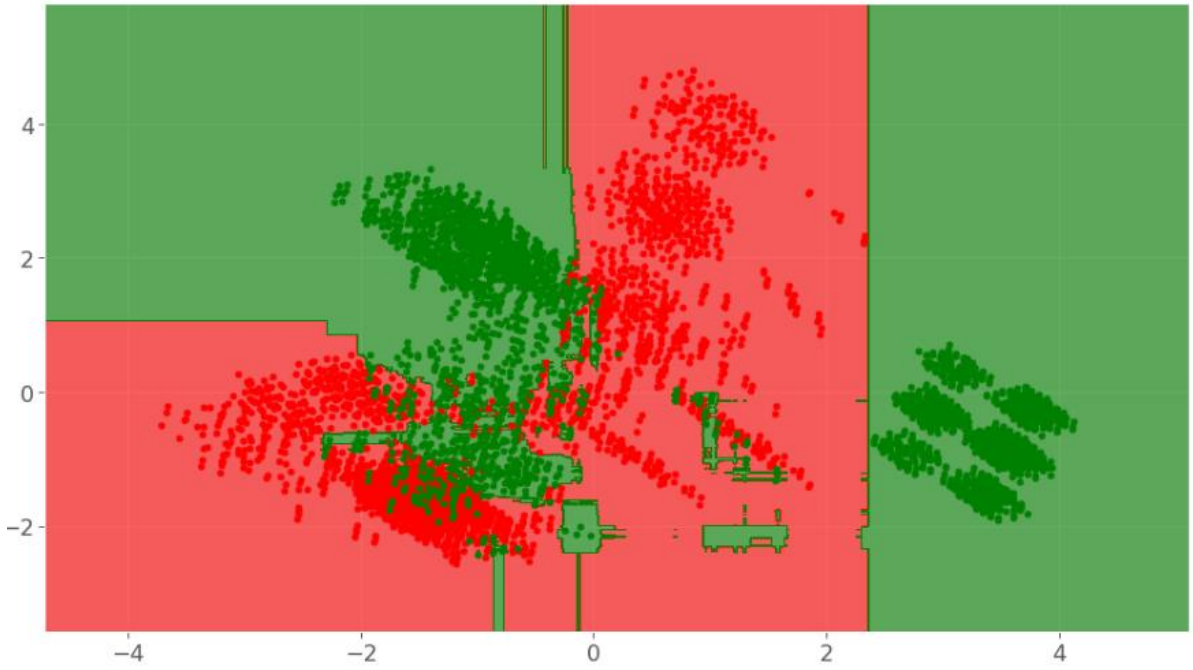
**Training Dataset**



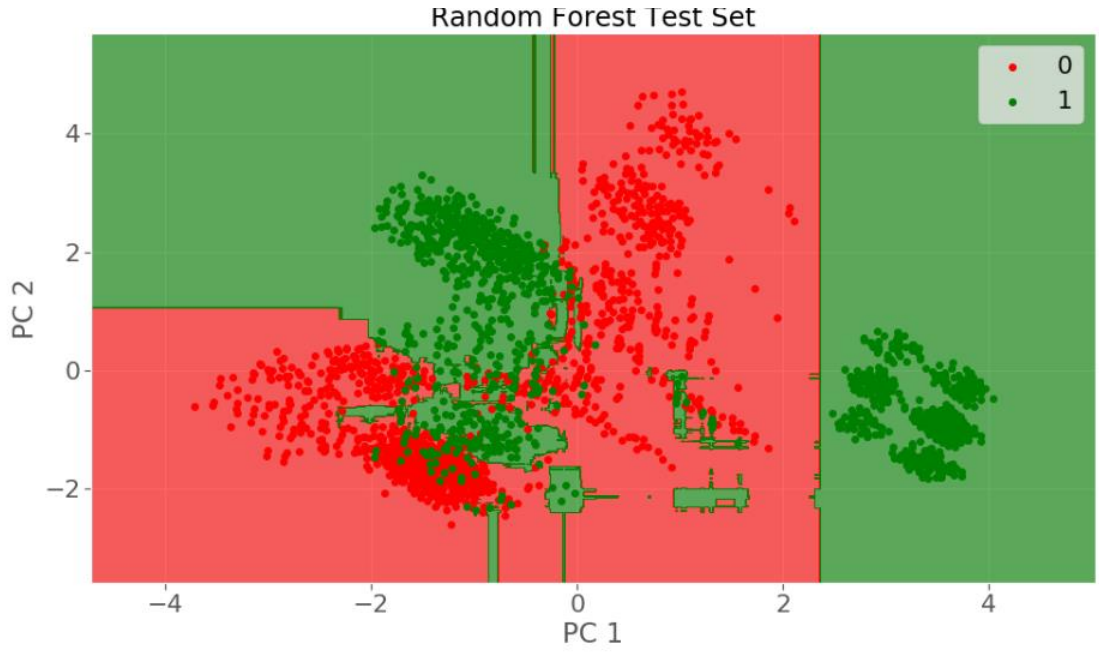

**Testing Dataset**

**Random Forest**

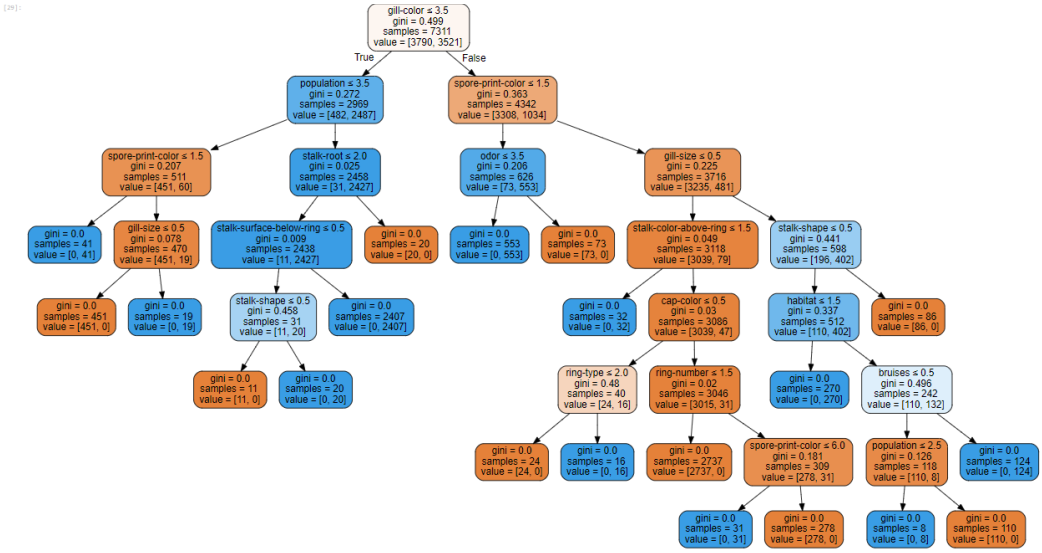
**Training Dataset**




**Testing Dataset**



**Decision Tree**




## VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

In this section the results of the proposed methodology are compared with the existing watermarking scheme with Bhatnagar *et al.* [15]. Bhatnagar proposes dual watermarking scheme where gray scale secondary image is embedded in the primary watermark image. This new watermark image embedding the primary image is embedded in the cover image using zigzag sequencing. This can be inferred from the results that the extracted primary and secondary watermark image seriously degrade.

## VIII. CONCLUSION AND FUTURE WORK

In this study, We aimed to learn how well can we predict whether the mushroom is edible or poisonous. We intended to implement two classification algorithms to build models for prediction .

Two classification models were tested for prediction accuracy and sensitivity and we determined that the Decision Tree gives the highest accuracy and sensitivity among the two. As a future work, we will extend this study to include feature engineering methods, to measure if the predictive power of the models could be increased or not.

## IX. REFERENCES


1. Wibowo, A., Rahayu, Y., Riyanto, A., & Hidayatulloh, T. (2018, March). Classification algorithm for edible mushroom identification. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 250-253). IEEE.
2. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1584-1589). IEEE.
3. Chen, N., & Marques, N. C. (2009, January). Extending learning vector quantization for classifying data with categorical values. In *International Conference on Agents and Artificial Intelligence* (pp. 124-136). Springer, Berlin, Heidelberg.
4. Ismail, S., Zainal, A. R., & Mustapha, A. (2018, April). Behavioural features for mushroom classification. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 412-415). IEEE.
5. Yang, X., Skidmore, A. K., Melick, D. R., Zhou, Z., & Xu, J. (2006). Mapping non-wood forest product (matsutake mushrooms) using logistic regression and a GIS expert system. *ecological modelling*, *198*(1-2), 208-218.
6. Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349-391.
7. Bojarski, M., Choromanska, A., Choromanski, K., & LeCun, Y. (2014). Differentially-and non-differentially-private random decision trees. *arXiv preprint arXiv:1410.6973*.
8. Markov, Z., & Russell, I. (2006). An introduction to the WEKA data mining system. *ACM SIGCSE Bulletin*, *38*(3), 367-368.
9. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009). Weka-a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook* (pp. 1269-1277). Springer, Boston, MA.
10. Hamid, N. A., Nawi, N. M., & Ghazali, R. (2011). The effect of adaptive gain and adaptive momentum in improving training time of gradient descent back propagation algorithm on classification problems. *International Journal on Advanced Science, Engineering and Information Technology*, *1*(2), 178-184.
11. Al-Mejibli, I. S., & Abd, D. H. (2017). Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices. *Journal of Al-Qadisiyah for computer science and mathematics*, *9*(2), Page-103.
12. Marmelstein, R. E. (1997). Application of genetic algorithms to data mining. In *Proceedings of 8th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS-97), edited by E. Santos Jr., AAAI Press* (pp. 58-65).
13. He, Z., Xu, X., Deng, S., & Dong, B. (2005). K-histograms: An efficient clustering algorithm for categorical dataset. *arXiv preprint cs/0509033*.
14. Ramesh, V., & Ramar, K. (2011). Classification of agricultural land soils: a data mining approach. *Agricultural Journal*, *6*(3), 82-86.
15. Andrews, S., & Orphanides, C. (2010). Analysis of large data sets using formal concept lattices.

16. Huang, Y. (2009). Advances in artificial neural networks–methodological development and application. *algorithms*, *2*(3), 973-1007.

17. Gorzałczany, M. B., & Rudziński, F. (2004, June). Application of genetic algorithms and Kohonen networks to cluster analysis. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 556-561). Springer, Berlin, Heidelberg.

18. Chowdhury, D. R., & Ojha, S. (2017). An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach. *International Research Journal of Engineering and Technology (IRJET)*, *4*(01), 529-534.

19. Chen, W., Panahi, M., & Pourghasemi, H. R. (2017). Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena*, *157*, 310-324.

20. Sharifian, S., Motamedi, S. A., & Akbari, M. K. (2011). A predictive and probabilistic load-balancing algorithm for cluster-based web servers. *Applied soft computing*, *11*(1), 970-981.