CLASSIFICATION OF MUSHROOM IN MYANMAR USING NAIVE BAYESIAN CLASSIFIER

KHAING EI EI ZAW

M.C.Sc. **JUNE, 2022**

CLASSIFICATION OF MUSHROOM IN MYANMAR USING NAIVE BAYESIAN CLASSIFIER

BY

KHAING EI EI ZAW

B.C.Sc.(Honous)

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

JUNE 2022

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin**, Rector, the University of Computer Studies, Yangon, for her kind permission to submit this thesis.

My thanks and regards go to my supervisor, **Dr Thin Lai Lai Thein,** Professor, Faculty of Information Science, University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw,** Professors, Faculty of Computer Science, the University of Computer Studies, Yangon, for her superior suggestion, administrative supports and encouragement during my academic study.

I also wish to express my deepest gratitude to **Daw Mya Hnin Mon**, Associate Professor, English Department, the University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

I especially thank to my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

ABSTRACT

Mushrooms are the most recognizable scrumptious food which is cholesterol free as well as plentiful in nutrients and minerals. Numerous types of mushrooms have been figured out all through the earth. Distinguishing palatable or harmful mushrooms through the unaided eye is very difficult, so mushroom species should have to arrange eatable and noxious. This framework will be arranged the sort of mushroom by utilizing Naive Bayesian classifier and K-Nearest Neighbor Method to foster helpful subset of mushroom highlights for characterization task. This system can classify the edible and poisonous mushrooms from mushroom dataset by using Naive Bayes Classifier. In this system, performance comparison of the two algorithms are used Naïve Bayesian classifiers and K-Nearest neighbor (KNN) by using confusion matrix. The Naive Bayesian classifiers have been perhaps the most loved approaches as premise of numerous grouping technique both hypothetically and basically. K-closest neighbor (KNN) is a regulated learning calculation where the consequence of new case inquiry is ordered in light of greater part of K-closest neighbor class.

This system is implemented by using C# programming language with Microsoft Visual 2013 and Microsoft SQL Server as the system database engine.

Keys: Naive Bayesian (NB), K-nearest neighbor (KNN), Mushroom Classification, supervised learning

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Mushroom is beefy and consumable natural product groups of a few types of parasites individuals from Basidiomycetes that normally fill in ground surface or substrate of different plants like straw and wood. Myanmar is ordered as one of the agrarian nations and has known as the stockroom of unmistakable mushroom in the agricultural nation. The quantity of types of mushroom that has been known as of recently is under 69.000 out of the assessment of 1.500.000 species on the planet and in Myanmar, there are under 200 species. This million types of mushroom, by and large, can be partitioned into two kinds, specifically consumable and harmful mushrooms. The Family of Agaricus and Lepiota ridiculously live in the open spaces; both with different shapes, varieties, and qualities which are not realized by many individuals are toxic. The Family of harmful Agaricus and Lepiota can cause ailment for one who consumes and furthermore can cause passing. The Family of Agaricus and Lepiota that are living fiercely can be consumed and, surprisingly, utilized as meds.

Mushroom hunting is acquiring prominence as a relaxation action for the most recent few years. Current examinations propose that a few mushrooms can be valuable to treat pallor, further develop body resistance, battle diabetes and a couple is even successful to treat malignant growth. In any case, not every one of the mushrooms end up being valuable. A few mushrooms are toxic too and utilization of these may bring about serious sicknesses in people and might in fact cause demise.

Mushrooms are being the most reasonably delivered food varieties, besides the fact that great taste yet in addition hold have an extraordinary healthy benefit. They have proteins, nutrients, minerals, and cell reinforcements. This can have different medical advantages. Utilization of mushrooms assists with battling various kinds of illnesses, for example, disease, assists with controlling blood cholesterol levels, and accordingly assists with battling diabetes. Mushrooms help in reinforcing our resistant framework and furthermore assist us with shedding pounds. They are a boggling combination of worthwhile as well as theoretical highlights. However, beside the sound mushrooms, there additionally exists harmful and wild mushrooms whose utilization might bring about serious diseases in people and could cause passing. It is difficult for a person to separate wild mushrooms from sound mushrooms.

Recognizing the edibility of mushroom physically is a too troublesome undertaking. Due to the greater part of the noxious mushrooms seem as though eatable mushroom attributable to variety and shape. Thus, computerization is vital in this field to diminish time and work. This framework arranges the kinds of mushroom by utilizing Naive Bayesian Classification. There are numerous arrangement approaches exist in AI. Different Authors are utilized characterization strategies, where Decision Tree ID3, CART and Neural Network classifier calculations have been utilized to order mushroom. The mushroom datasets were gathered from Ministry of Agriculture, Livestock and Irrigation and papers from on the web. 46 types of mushrooms are recorded. There are 16 credits. The properties use in datasets there are class, cap tone, cap shape, cap surface, cap umbonate, gills/pores tone, gills/pores connection, gills/pores dividing, stipe tone, stipe shape, stipe, annulus or ring, spore tone, spore shape, spore surface, spore size and last developing territory of mushroom. For the quantity of each class, comprising of 192 information remembered for the food mushroom class and 802 information remembered for the harmful mushroom classification, so the complete number of information utilized was 994 information. Each underlying of each property and class is a portrayal of the kind of characteristic concerned.

The primary reason for this framework is to arrange mushroom edible and poisonous. In this framework, mushroom datasets are divided into two classes, a training class and testing class. 70% of the information are assigned to the training set and 30% is dispensed to the testing set. When ascertain the precision, the framework compute the likelihood by utilizing Classifier and K-Nearest Neighbors calculation.

1.1 Motivation

In agricultural nations, the significance of palatable mushrooms inside customer inclinations and discernments has not been examined. Mushrooms are plants that are broadly consumed by the overall population, however not all mushrooms can be consumed straightforwardly, on the grounds that the kinds of mushrooms are practical and it is still too hard to even think about recognizing. This system can be classified the edible and poisonous mushrooms from mushroom dataset by using Naive Bayes Classifier. The best level of accuracy between the two algorithms can be determined by comparison. Naive Bayesian and K-Nearest Neighbors are compared by calculation accuracy, precision, recall and f-measure to get more reliable and good performance in classification. K-Nearest Neighbors algorithm takes longer time to process as compare to Naive Bayes. Naive Bayes can be provided high accuracy when large amount of data. Therefore, the consumers can support to know edible and poisonous for many species of mushrooms in this system.

1.2 Objectives of the Thesis

The objectives of the system are to study on determining edible or poisonous mushroom various species of mushroom in Myanmar. Consumers support to know edible and poisonous mushrooms. The proposed system to easily classify poisonous or not by machine learning using Naive Bayesian. The best level of accuracy between the two algorithms can be determined by comparison. Naive Bayesian and K-Nearest Neighbors are compared by calculation accuracy, precision, recall and f-measure to get more reliable and good performance in classification. And then, compute the accuracy and show evaluation result. This system can be proved the Naive Bayesian Result is better than K-Nearest Neighbors base on the experiment of Accuracy, Precision, Recall and F-measure.

1.3 Related Works

Neural Network classification on mushroom dataset with feature selection using evolutionary algorithm and auto–associative network by Yuhan Zhang. The result is comparison of prediction accuracy. Prediction accuracy of Neural Network with feature selection is 77%. Neural Network without feature selection is 70%. Neural network is difficult to know how many neurons and layers are necessary [1].

Comparative classification algorithm testing accuracy in previous data mining has not been done and based on the results of testing of the three best classification algorithms in the data mining. The C4.5 algorithm has the highest accuracy compared to the other two popular classification algorithms, and in terms of processing speed. The decision tree generated by this algorithm can be easily applied to application creation and this algorithm also cuts the number of variables required for identification. For further research, researchers can develop the results of this research into a mobile application equipped with images that make it easier for people to recognize the edible wild mushrooms. Research on the identification of edible mushrooms also can be developed using image processing or compared to other classification algorithm [2].

Decision tree for the classification of mushroom dataset, B.Lavanya and G.R.Preethi [3]. The result is comparison of prediction accuracy. The accuracy for ID3, CART and Hoeffding Tree (HT) are 69%, 90% and 100% respectively. ID3 does not handle numeric attribute and missing values.

1.4 Overview of the System

.

A mushroom is one of the growths types' food that has the most powerful supplements on the plant. Mushrooms enjoy significant benefits like by killing malignant growth cells, infections and upgrading the human safe framework. Right now, the mushroom alludes to the cycle that performed by robot in food industry. This method used to restrict the highlights like tone. Later, mushroom framework utilized explicit qualities that further develop the determination interaction of mushrooms. Such framework relies upon examining and exploring the highlights to get better grouping in view of the notable elements.

In this review case, the examination will be completed to track down the best exactness in deciding the arrangement of mushroom utilizing two characterization calculations. The proposed calculation is Naive Bayes calculation. For the performance comparison of accuracy, the two algorithms are used Naive Bayesian classifiers and K-Nearest neighbor (KNN) by using confusion matrix. And then, assesses and approves the outcomes by searching for the best exactness consequences of these two calculations. The following stage is looked at the consequences of the exactness of every calculation, to get a model grouping calculation that acquires the most elevated precision and time intricacy. The most elevated exactness results from this computation can be supposed to be the best calculation in deciding the grouping of noxious or consumable mushrooms. And afterward this framework proceeds with estimation exactness to test once more the consequence of the precision.

1.5 Organization of the Thesis

The thesis is organized in five chapters. They are as follows:

In Chapter 1, introduction of the system, objectives of the thesis, related works and thesis organization are described. **Chapter 2** presents the background theory of classification. **Chapter 3** discusses the detail methods of proposed system **Chapter 4** expresses the design and implementation of the proposed system. Finally, **Chapter 5** presents the conclusions, benefits of system and further extensions of the system.

CHAPTER 2
THEORETICAL BACKGROUND

Growths have a place with the contagious realm, consequently mushrooms don't have genuine leaves and roots, and don't have chlorophyll so they cannot do photosynthesis like plants overall. Parasites are arranged or characterized independently on the grounds that they can't be grouped in plants or creatures. There are growths that should be visible straightforwardly or are perceptible and some should be noticed utilizing a magnifying lens or tiny shape. As a general rule, parasites have numerous cells (multicellular) like consumable mushrooms and tempeh mushrooms, however some are single-celled (unicellular) like yeast or yeast (Saccharomyces). Multicellular organisms are made out of strings called hyphae. When seen with a magnifying lens, hyphae have an isolating structure (septa) and some are not parceled [1].

A mushroom is one of the growths types' food that has the most intense supplements on the plant. Mushrooms enjoy significant benefits like kill disease cells, infections and improving the human invulnerable framework. Right now, the mushroom alludes to the cycle that performed by robot in food industry. This method used to restrict the elements like tone. Later, mushroom framework utilized explicit qualities that further develop the choice course of mushrooms. Such framework relies upon breaking down and examining the highlights to get better characterization in light of the notable elements [2].

To recognize which mushrooms are palatable and harmful, there are multiple ways that can be utilized. One of the viewpoints that can be utilized as benchmarks in recognizing an organism is its morphological qualities. The morphological elements alluded to are the state of the umbrella, variety, living space and different highlights apparent to our eyes. We got these morphological qualities from the datasets we took from Ministry of Agriculture, Livestock and Irrigation and papers from online [3].

Datasets is an assortment of information. On account of plain information, an informational index relates to at least one data set tables, where each section of a table addresses a specific variable, and each column compares to a given record of the informational index being referred to. The informational index records values for every one of the factors, like level and weight of an item, for every individual from the informational index [4].

For this situation, there is utilized two techniques to decide the grouping of mushrooms, specifically the Naive Bayes strategy and furthermore the K-Nearest Neighbor technique as the classifier [5]. There is utilized these two techniques since they have different exactness and we can contrast it and the strategy which it is gotten better precision of the two techniques that have been tried. Extraction of morphological elements is utilized to assist with recognizing growths, so later it will be known including the kinds of eatable or harmful mushrooms.

2.1 Knowledge, Data and Uncertainty

Not all information is made equivalent. It is a continuum of portrayals with changing degrees of significant worth and activity capacity. These levels or states structure a movement from the most reduced level, where ease of use is negligible or potential to more significant levels where convenience is more clear and more prompt. Through different sorts of information handling one might advance from lower to higher states, expanding the importance of information concerning achieving some substantial assignment. The most noteworthy express, a choice, is information showing a promise to make a few move and results from the handling of information at different levels. Figure 2.1 shows a potential arrangement of information states and potential tasks to hop starting with one state then onto the next. The quantity of states or the substantial tasks used to move between various explicit states are not significant for the fact being made, only that a bunch of states with changing levels of convenience or activity capacity exist and that it is feasible to advance to a higher state by executing a few procedure on the information at lower states.

These thoughts make an interpretation of well to a characterization issue. Perceptions are information, a low state with potential however no prompt activity capacity. Characterization calculations, at an extremely undeniable level, basically apply a bunch of handling steps to these named perceptions and, ideally, produce a model equipped for settling on conclusions about the class of beforehand concealed examples. This model is then at the most elevated information express, its activity capacity is clear and quick.

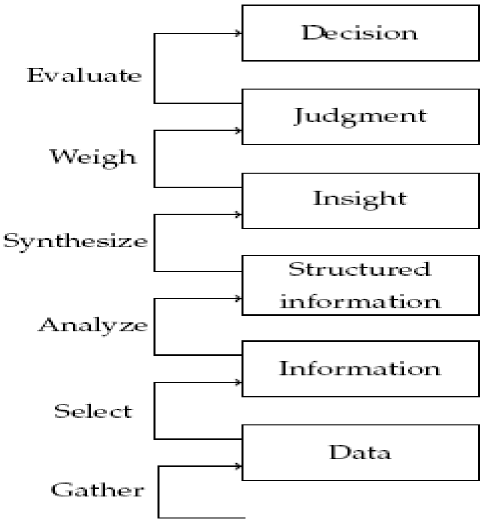


Figure 2.1: The progression from lower knowledge states with marginal usability to higher knowledge states with immediate usability.

Where does area information sit in this movement? It does not have the quick activity capacity of a choice, in any case everything expected to characterize new occasions would currently be known and no educational experience would occur. Then again, as it is a formalization of information is given by space specialists it is sensible to expect that it is more organized and has preferable convenience over simple perceptions. That is on the grounds that specialists currently to some degree handled these information by social occasion, choosing and breaking down information from various sources and encounters in the area. That is the way they become specialists. ILP frameworks do not, customarily, make this differentiation and as such the two perceptions and area information contribute similarly to the speculation being created, that is to say, the theory needs to fulfill the space information, every one of the positive perceptions and none of the negative. This expects that are certain beyond a shadow of a doubt about the name, everything being equal, which is only sometimes the situation, and about the importance of each and every assertion in the space information for the main pressing concern, which does not necessarily occur. The methodologies endeavor to catch in their construction that marked examples and space information are at various information states and ought to contribute in various ways to the model being produced. In view of this utilization programmed sensible deduction on space information and the new suggestions that are created return into the collection of area information. This is sensible in light of the fact that this sort of information, by its tendency, was at that point chose and broke down by a space master and isn't supposed to be uproarious or misleading. It can, in any case, be unimportant to the main pressing issue.

To manage the likelihood that a few recommendations in the space information are unimportant to the characterization issue viable there can be stayed away from the utilization of coherent surmising to develop the model from the area information, i.e., there can be permitted and utilized sensible deduction inside the current space information yet keep away from this sort areas of strength for of while building the model. Consider for instance that the accompanying suggestions are important for the current space information: "Lepiota have white gills, white spores and have rings on the stems", "MushroomX has white gills and white spores", "MushroomX has rings on the stems". From the later two attestations about "MushroomX" and the primary recommendation about "Lepiota" one can intelligently surmise that "MushroomX" has a place with the class "Lepiota". This new suggestion will be added to the area information however could conceivably be utilized in the model being fabricated.

Space information can be added additional aspects to the current named occurrences, similar to the types of a mushroom, yet whether this aspect will be important for the model really relies on how it makes sense of the basic relations among highlights and the worth of the objective trait. Basically this intends that albeit the choice to add another aspect is driven by sensible derivation, the choice to consolidate that additional aspect in the model is driven by measurable deduction.

2.2 Discussion of Mushroom Classification with Different Classifiers

Commonly the distinction between learner, model and classifier is somewhat nebulous. Once the model is built the nodes in the tree correspond directly to the attributes of every new instance that one might have to classify and it is common to call the model itself, a classifier.

2.2.1. Results and Discussion with Different Methods

To execute a few investigations and look at the presentation of the proposed calculation with the standard ID3 and C4.5 choice tree calculations a Java execution was created, as a component of the D2PM structure [Antunes, 2011]. The standard ID3 variant used to analyze was composed by the creators and results were thought about against Weka's execution [Hall et al., 2009] to guarantee no mix-ups were made. The C4.5 execution involved was the J48 execution in the Weka library with pruning and subtree raising empowered.

Disregarding information with values determined at various degrees of deliberation will be being normal in numerous spaces of utilization, there are not many standard benchmark informational indexes with these qualities and with a related philosophy. We chose the Mushroom and the Nursery informational collection from the UCI Machine Learning Repository [Bache and Lichman, 2013].

The nursery informational collection compares to 12960 perceptions with 8 credits and an objective quality with 5 potential qualities. Three of these five classes rule the informational index, with each having around 32% of the universe of occasions. The two excess classes are addressed by less than 3% of all occasions. The mushroom informational index incorporates depictions of 8124 examples relating to 23 types of gilled mushrooms in the Agaricus and Lepiota family (albeit no data is available about the types of every perception). There are 22 credits and the objective quality has two potential qualities: noxious or consumable. The perceptions are almost equitably dispersed between these two classes.

Area information got from the booklet "The Mushroom Hunter's Field Guide" and from [Zhang et al., 2002] was built unequivocal in an OWL 2 cosmology. Three arrangements of trials were then executed. The primary analyzes the precision of the proposed Hierarchy Decision Tree with the standard ID3 and C4.5 calculation on the first information, where all values are accurate. Take a gander at the intricacy of the created choice trees. Fig 2.2 explains an illustration of a choice tree created by HDT for haphazardly chose little subsets (around 50 cases) of the informational index as preparing sets. This straightforward tree has an exactness over the whole informational index of 0.914 while the standard ID3 calculation, for a similar preparation set, creates a tree that has a precision of just 0.549, nearly as terrible as haphazardly picking a class. The second arrangement of tests shows how the precision of all calculations advance with the size of the preparation sets. A subset with 1000 occurrences was haphazardly chosen from the first informational collection to act as a test set. Six subsets were haphazardly chosen from the leftover occurrences of the first informational index, with amounts of 700, 300, 70, 50, 20 and 15 occasions to be utilized as preparing sets. To survey the precision of the three calculations, there is utilized cross-approval by rehashed irregular subsampling.

Five disjoint subsets were haphazardly chosen and each was isolated in two disjoint subsets, a preparation set and a test set. It can be shown the mean correct nesses.

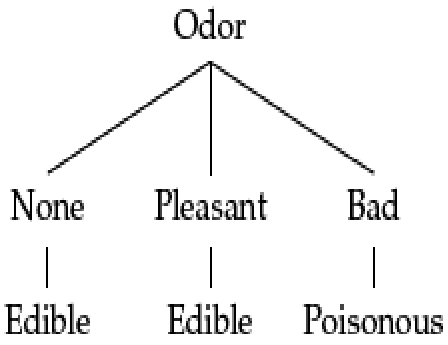


Figure 2.2: Example of a decision tree generated by HDT from a small training set (< 50 instances)

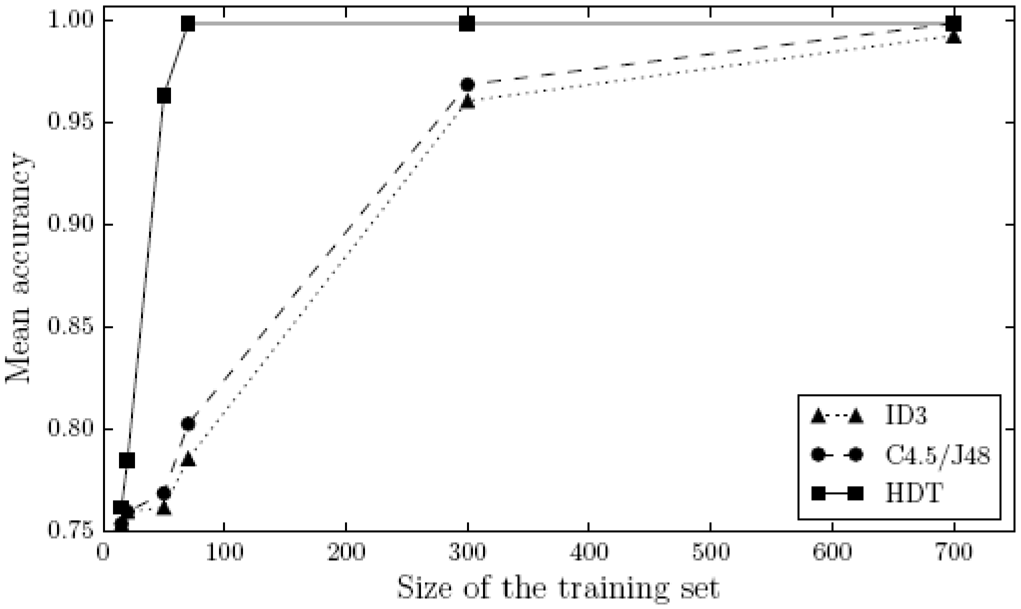


Figure 2.3: The power of training set size on the accuracy of ID3, C4.5 and HDT in the Mushrooms data set.

The outcomes acquired (showed in Figure 2.3 and Figure 2.4) show that our methodology beats both ID3 and C4.5 in every one of the tried subsets of the two informational indexes. The thing that matters is more articulated in the more modest preparation sets, turning out to be less perceptible as the size of the preparation set increments.

On more modest preparation sets all things considered, not all conceivable quality qualities are available. As HDT attempts to construct a more broad model, for certain hubs relating to digest credits, it is as yet ready to anticipate the class of examples containing highlights that were absent in the preparation set, while ID3 and C4.5 fall flat. At the point when the size of the preparation set develops and all trait values become present, ID3 and C4.5 get up to speed.

The last arrangement of tests concentrates on how the precision of the calculations changes with a rising number of prices being dynamic, e.g., not recognizing the specific scent of a mushroom but rather having the option by knowing if it has a wonderful or terrible stench. Beginning from an informational collection with no theoretical qualities, six informational collections were then created with a surmised level of conceptual characteristic upsides of 5%, 10%, 15%, 20%, 25% and half.

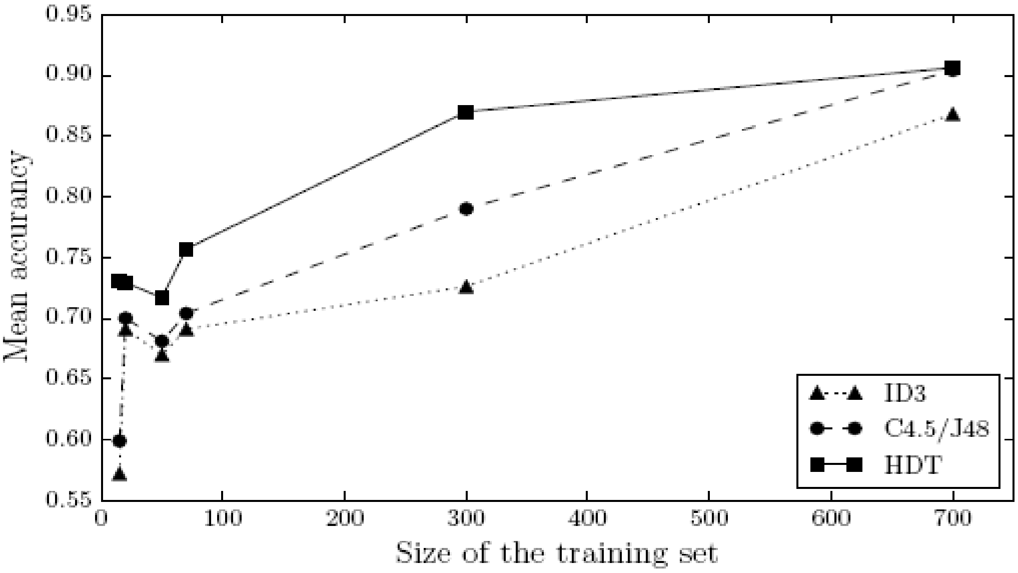


Figure 2.4: The power of training set size on the accuracy of ID3, C4.5 and HDT in the Nursery data set.

The after effects of these tests show that HDT can keep up with it's precision better than ID3 and C4.5 when, rather than the specific quality qualities, just more dynamic adaptations of the highlights are accessible. This fills in as proof that HDT can be utilized highlight ordered progressions to fabricate more hearty models that keep up with great prescient power when some data exists about the trait esteem, however the data is deficient to decide it's worth precisely. Figure 2.5 shows these outcomes.

HDT's capacity to keep up with it's prescient power notwithstanding less exact trait values relies upon the nature of the accessible component ordered progressions and, in some degree, on the idea of the arrangement issue. For some characterization assignments it might just be the situation that for certain characteristics, the specific worth is expected to anticipate the right class. In these cases HDT will be picked the substantial variant of the characteristic while building the model yet the heartiness of such a model is adversely impacted, albeit still better compared to conventional methodologies. Figure 2.6 shows this.

These outcomes are in accordance with our assumptions. In the first place, even on information were all prices are accurate, space information can be assisted with building models that proceed as great or better while being significantly less complex. This distinction in exactness is more articulated with more modest preparation sets. Next, when the particular substantial qualities are obscured however a more dynamic variant is accessible, HDT keeps up with its presentation surprisingly well while the exhibition of conventional ID3 and C4.5 diminishes as additional qualities are communicated at more elevated levels of reflection.

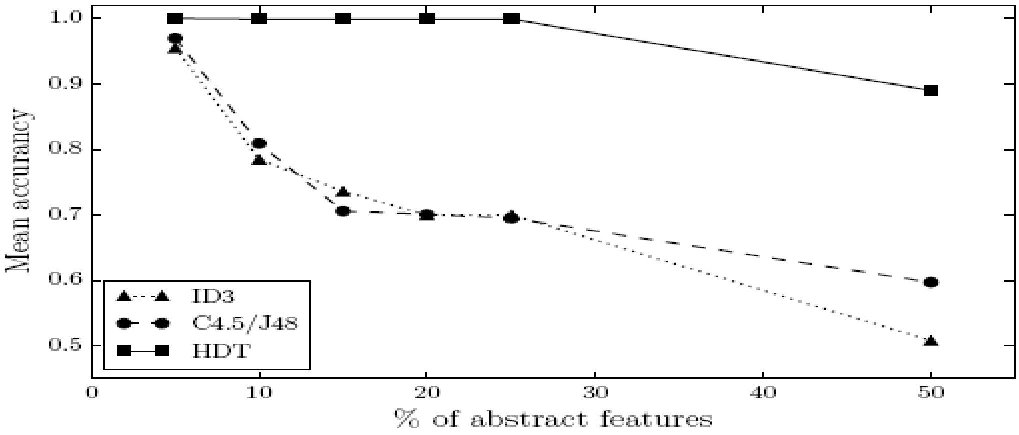


Figure 2.5: Accuracy of ID3, C4.5 and HDT in the Mushrooms data set with a raising number of abstract prices.

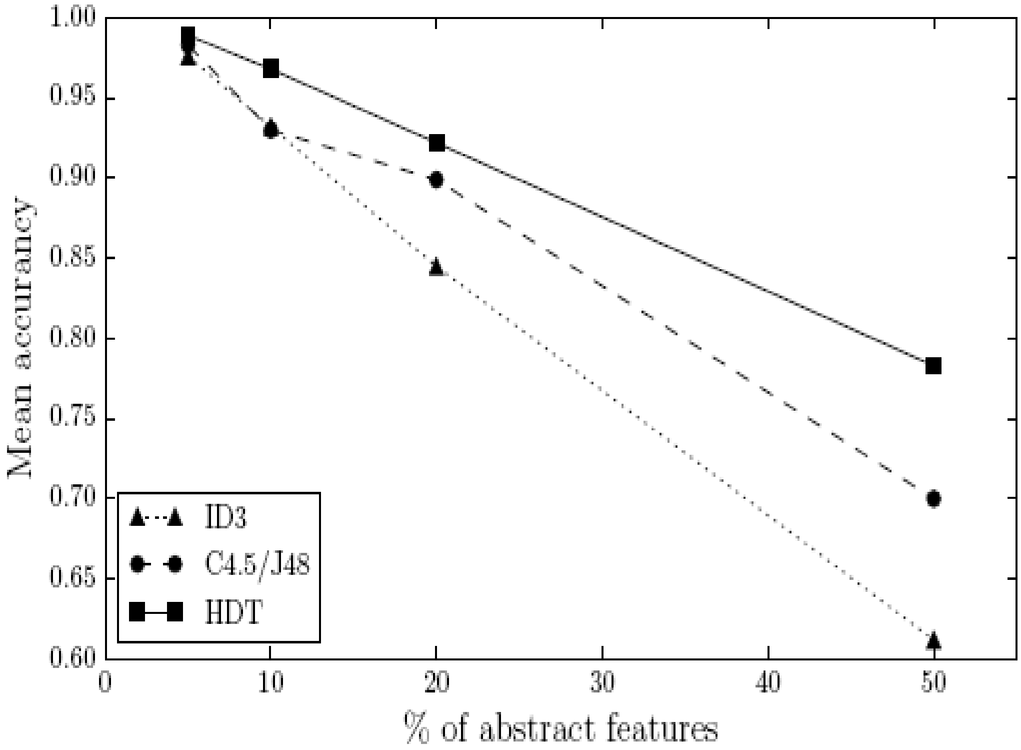


Figure 2.6: Accuracy of ID3, C4.5 and HDT in the Nursery data set with a raising number of abstract prices.

2.2.2. Summary of Different Approaches for Mushroom Classification

The outcomes show that the technique we propose can perform significantly better compared to conventional classifiers even with little preparation sets, accomplishing levels of execution that require more customary ways to deal with be prepared with a lot bigger sets. They additionally show that as we decline the level of substantial elements, supplanting them with less exact (more conceptual) ones our strategy can keep up with its exhibition while the precision of conventional methodologies diminishes fundamentally.

Regardless of these qualities this technique actually experiences the exceptionally restricted scope of aphorisms (SubClassOf and SuperClassOf) that are upheld. Albeit these are sufficient to fabricate progressive systems, they are plainly sufficiently not to characterize additional fascinating principles that would permit us to characterize new aspects from existing properties, e.g., all mushrooms have a terrible stench and white spores are of the species bogus parasol.

2.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) is a non-parametric order calculation, for example , it makes no assumptions on the rudimentary dataset. It is identified for its effortlessness and viability. It is a regulated training calculation. A marked preparation set is given where the information focuses are ordered into different groups, so that group of the unlabeled information can be anticipated.

In Classification, various attributes decide the group to which the unlabeled information has a place. KNN is generally utilized as a classifier. Grouping information in light of nearest or adjoining preparing models in a set region is utilized. This technique is utilized for its straightforwardness of execution and low calculation time. For ceaseless information, it utilizes the Euclidean distance to compute its closest neighbor. For another info the K closest neighbor are determined and the greater part among the adjoining information chooses the arrangement for the new information. Despite the fact that this classifier is basic, the worth of 'K' assumes a significant part in ordering the unlabeled information. There are numerous ways of choosing the qualities for 'K', however it can basically run the classifier on various occasions with various qualities to observe which worth provides the best outcome. The calculation cost is somewhat high since every one of the estimations are being made while the preparation information is being arranged, not when it is experienced in the data.

It is a lethargic training calculation as not much is done when the data is being prepared with the exception of putting away the preparation information and retaining the dataset all things considered. It does not perform speculation on the preparation set. So the whole central data being prepared is needed when in the testing phase. In relapse, K - closest neighbor predicts consistent qualities. This worth is the normal of the upsides of its K - closest neighbor.

2.3.1. Development of KNN

K-closest neighbor arrangement was created to execute trademark examination when clear parametric approximations of likelihood densities were obscure or hard to decide. In an unpublished US Air Force School of Aviation Medicine report in 1951, Fix and Hodges presented a non-parametric calculation for design grouping that has since become known the K-closest neighbor rule.

2.3.2. Tasks of KNN

KNN is a supervised learning classifier. Mainly there are two phases in classification:

1. Learning Phase: Using the train data a classifier is created.

2. Assessment of the classifier.

As indicated by the closest neighbors method, the new unlabeled information is arranged by figuring out which groups its neighbors have a place with. KNN calculation uses this idea in its computation. In the event K-Nearest Neighbor calculation, a specific worth of K is fixed which helps us in ordering the obscure tuple. When a new unlabeled tuple is experienced in the data, K-Nearest Neighbor performs two tasks:

To start with, it breaks down the K focuses nearest to the new data of interest, i.e., the K closest neighbor.

Next, utilizing the neighbors classes, K closest neighbor decides concerning which class should the new information be arranged into.

At the point when a few new information is added, it characterizes the information in like manner. It is more valuable in a dataset which is generally partitioned into groups and has a place with a particular district of the information plot. Consequently this calculation gets more exactness isolating the information inputs into various classes in a more clear manner. K closest neighbor sorts out the class having the greatest number of focuses sharing minimal separation from the information guide that requirements toward be ordered. Thus, the Euclidean distance should be determined between the test and the predetermined preparation tests.

After we assemble K closest neighbor, it is essentially take most of them to anticipate the class of the preparation model. The variables that influence the presentation of K closest neighbor are: the worth of K, the Euclidean distance and the standardization of the boundaries. To comprehend the definite working of the calculation, the means are as per the following:

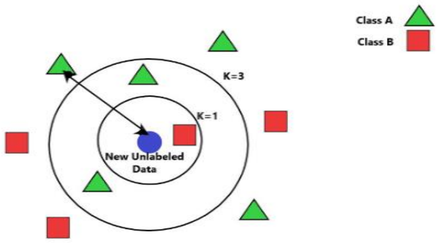


Figure 2.7: Sample Classification of KNN

Set the training data : { (x(1), y(1)) , (x(2), y(2)), ... , (x(m), y(m)) }

Step1: Put the train data

Step2: For each new unlabeled data,

A. Compute Euclidean distance with all train data points using the formulary

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

B. Find the k- nearest neighbors

C. Select group having the maximum amount of nearest neighbors.

In the wake of putting away the preparation, set all boundaries should be standardized, with the goal that the estimations become more straightforward. The consequence of the grouping is delicate to the worth of 'K'. The info changeable 'K' concludes the quantity of neighbors that should be thought of. The worth of 'K' impacts the calculation as utilizing the 'K' esteem we can fabricate the limits of each class.

TO DETERMINE K: The greatest worth of K is picked by first inspecting the information Larger upsides of K are more exact as they lessens the net commotion however this isn't ensured. A decent worth of K can likewise be resolved utilizing cross approval. On the off chance that K=1, the information is basically assigned to the group of its closest neighbor. At K=1, the mistake rate is reliably zero for the preparation information. This occurs in light of the fact that the closest highlight any preparation information viewpoint is itself. Consequently the greatest outcomes are acquired if the worth of K=1. Yet, with K=1, the limits are over fitted.

In the event of tiny upsides of 'k' the calculation is too delicate to even consider noising. To obtain a good worth of K, the preparation and approval set should be isolated from the underlying data. Assuming the two Nearest neighbors (K=2) have a place with two unique classes, the result is obscure. In this way, we increment the quantity of closest neighbors to a bigger worth (say 5-closest neighbors). This will characterize an earliest neighbor area and will give the lucidity.

Bigger upsides of 'K' make the group limits smoother, which probably won't be alluring as then the marks of different groups might obtain remembered for the area.

While the preparation information focuses are available in a dissipated way, the worth of K is challenging to decide.

2.3.3. Advantages

K-nearest neighbor is known for its effortlessness, intelligibility and adaptability. It is not difficult to decipher. The estimation time is less. Likewise, the prescient power is exceptionally high which makes it compelling and effective. K-nearest neighbor is exceptionally compelling for huge preparation set. The means continued in the grouping done by this calculation are moderately less perplexing than that followed by different calculations.

The numerical calculations are not difficult to fathom and comprehend. They do not include computations that appear to be troublesome. Fundamental ideas like that of Euclidean distance estimation are utilized which upgrade the effortlessness of the calculation as opposed to selecting other composite strategies like that of incorporation or separation. It is valuable for non-direct information. K-nearest neighbor is compelling for characterization as well as relapse.

2.3.4. Disadvantages

K-nearest neighbor can be costly in assurance of K if the data is huge. It needs a more prominent stockpiling than a powerful classifier. In K-nearest neighbor the expectation stage is delayed for a bigger dataset. Likewise, calculation of exact distances assumes a major part in the assurance of the calculation's precision. One of the significant stages in KNN is deciding the boundary K. Now and again it isn't clear which kind of distance to utilize and which component will give the best outcome. The calculation cost is very high as the distance of each preparing model is to be determined. K-nearest neighbor is a sluggish gaining calculation as it does not gain from the preparation information, it’s suggestion retains it and afterward utilizes that information to characterize the new information.

2.3.5. K-nearest neighbor and its Variants

As examined before, the effectiveness of the calculation can be improved by causing changes in the variables that to oversee it. There are numerous variations of K-nearest neighbor that have been concentrated before to make this calculation more successful, some of them are:

1. Locally Adaptive K-nearest neighbor:
   Locally versatile K-nearest neighbor calculations proposed by[1]. It picks the worth of k that ought to be utilized to order a contribution by looking at the aftereffects of cross-approval calculations in the nearby neighborhood of the unlabeled information.

1. Weight Adjusted K-nearest neighbor:
   The calculation by [2] recommends that the distances, on which the quest for the closest neighbors is situated in the initial step, must be changed into comparable measures, which can be utilized as loads. The relegated loads conclude how much a quality impacts the characterization activity. This classifier is especially valuable for the situation where a dataset has many elements, some of which can be viewed as un-essential, however it has high computational expense.

2. Improved K-nearest neighbor for Text Categorization:
   [3]proposes a refined K-nearest neighbor calculation for text order, which builds the characterization model by combining K-nearest neighbor text classification and confined one pass grouping calculation. On the off chance that a steady worth of K is utilized for every one of the classes, the class with bigger number of properties will enjoy a benefit. In better KNN, a reasonable number of closest neighbors are utilized by the conveyance of information in preparing sets, to foresee the class of an unlabeled information.

3. Adaptive K-nearest neighbor:
   K-nearest neighbor recognizes same number of closest neighbors for each new information. Versatile K-nearest neighbor by [4] figures out a fit worth of K for each test. Initial an ideal worth of K is found. Then, at that point, to anticipate the arrangement of the unlabeled information, the worth of K is set equivalent to the ideal worth of K of it's closest neighbor in the preparation

data. The implementation of the proposed calculation is then tried on various datasets.

4. K-nearest neighbor with Shared Nearest Neighbors

   A better K-closest neighbor calculation is introduced by [5] utilizing divided closest neighbor comparability which can figure likeness among test tests with closest neighbor tests. It utilizes Similarity judgment calculation and works out the closest neighbor similitude an incentive for each preparing test. Then it ascertains the most extreme between these qualities.

5. KNN with K-Means:

   One more ad libbed way to deal with the calculation is portrayed by [6]. This calculation attempts to isolate a bunch of focuses into K sets or groups so the focuses in each bunch are near one another. The focuses of these newly made groups are taken as the new preparation tests. To foresee the characterization of an unlabeled information, its separation from the recently found preparing focus is determined, and the middle what shares the base separation from the information is allotted to that group. Dissimilar to standard K-nearest neighbor, there is the information boundary K isn't passed. This records to being one of its benefits.

6. SVM K-nearest neighbor

   Support Vector Machine (SVM) is an order strategy that can be applied on straight as well as non-direct information. It is a composite variant of K-nearest neighbor blended in with SVM for visual classification acknowledgment, and is expanded in [7]. In this calculation, the preparation is finished with the assistance of K closest neighbors to the un-named data of interest. To begin with, the K-closest information not set in stone. Then, at that point, pairwise distance between these K information focuses is processed. Subsequently we get a distance framework from the determined distances. A Kernel network is then planned from the got distance framework. This bit framework is taken care of as contribution to SVM classifier. The outcome acquired is the group of the obscure piece of information. Then again, one could utilize SVMs yet time utilization is one of its downsides. Additionally, it includes estimation of pairwise distances.

7. K-nearest neighbor with Mahalanobis Metric

   The measurement distance is critical in grouping of another data of interest. Mahalanobis is another distance metric, approach of which is canvassed in [8].The metric guarantees that the K-closest neighbors are included in similar class and the examples having a place with various classes are isolated by an enormous level of contrast.

8. Generalized K-nearest neighbor

   KNN can likewise be utilized for constant - esteemed class credits. For the this arrangement, the typical qualities determined among neighbors is allotted to the group property of the unlabeled information. [9]implements this calculation to foresee the constant - esteemed group characteristic.

9. Informative K-nearest neighbor

   Typically the worth of K depends on the information, making it hard to pick the boundary as per various applications. [10] presented another metric that actions the enlightening ness of objects to be grouped. Educational ness estimates the significance of focuses. In this strategy, there are two information boundaries K and I. The greater part class of most educational coming down models will be the class of the new test.

10. Bayes K-nearest neighbor

    The information values encompassing the objective are created by a similar likelihood conveyance, extending outwards over the reasonable number of neighbors. [11] recursively figured the likelihood of the last change-point and moved towards the objective, and registered the back likelihood dispersion over K.

Roshna Chettri and Shrijana Pradhan described the comparison result of prediction accuracy: accuracy for K-NN, Naïve Bayes and Case based Reasoning are 72%, 85% and 92 % respectively for the analysis of “Internet of things: Comparative Study on Classification Algorithms (KNN, Naïve Bayes and Case Based Reasoning). Following figure shows that analysis by graph.

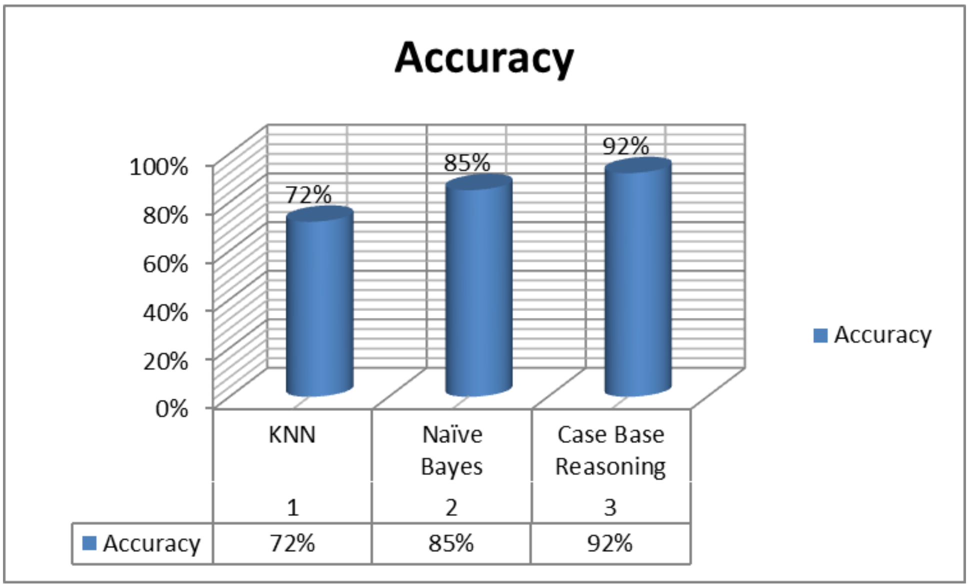


Figure 2.8: Accuracy Comparison

CHAPTER 3

NAIVE BAYESIAN CLASSIFIER

The characterization calculation is a supervised learning method that is utilized to recognize the classification of novel perceptions based on preparing information. In this classification, a calculate gains from the certain dataset or perceptions and afterward characterizes novel perception into various classes or gatherings. Such as "**Yes or No", "0 or 1", "Spam or Not Spam", "cat or dog",** etc. Classes can be called as targets/labels or categories. Dissimilar to relapse, the result variable of Classification is a class, not a worth, for example, "Green or Blue", "natural product or creature", and so on. Because the characterization calculation is a supervised learning strategy, subsequently it obtains named word information, and that implies it includes word with the relating yield. During characterization calculation, a discrete result function(y) is planned to enter variable(x).

3.1. Naive Bayesian Classifier

The Naive Bayesian classification depends on the Bayes hypothesis, and is especially fit when the dimensionality of the information sources is big-priced. Regardless of its effortlessness, Naïve Bayesian classification can frequently accomplish tantamount execution with some refined order techniques, for example, choice tree and chose brain net classifier. Gullible Bayesian classifiers have likewise displayed high-priced exactness and quickness when practiced to enormous datasets. Here part, we will momentarily survey Bayes' hypothesis, then give an outline of Naive Bayesian classification and its utilization in AI, particularly record characterization.

3.1.1. Bayes Theory and Preparation

A generally involved system in grouping is given by the straightforward hypothesis of likelihood well-known as Bayes hypothesis or standard. Fore there will be presented Bayesian Theory, let first survey two principal laws of likelihood hypothesis in the accompanying structure:

$$p(X) = \sum_{Y} p(X,Y) \qquad (3.1)$$

$$p(X,Y) = p(Y|X)p(X). \qquad (3.2)$$

where the first mathematical statement is the *sum law*, and the next mathematical statement is the *produce law*. This *p*(*X*, *Y*) is a join probability, the amount *p* (*Y*|*X*) is a conditional probability, and the amount *p*(*X*) is a marginal probability. These twice easy laws form the base for all of the probabilistic theorem.

Depended on the result law, all together with the similarity proprietary *p*(*X*,*Y*) = *p*(*Y*,*X*), it is simple to obtain the next Bayesian theory,

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}, \qquad (3.3)$$

which assumes a focal part in AI, particularly order. Utilizing the total rule, the denominator in Bayes' hypothesis can be communicated as far as the amounts showing up in the numerator.

The denominator in Bayes' hypothesis can be viewed just like the standardization consistent expected to guarantee that the amount of the contingent likelihood on the left-wing part of mathematical statement (3.3) over all upsides of Y approaches one.

Allow to think the straightforward guide toward all the more likely figure out the essential ideas of likelihood hypothesis and the Bayes' hypothesis. Assume us have two packs. They are one pink and one yellow, and in the pink pack there have two oranges, four apples and six lemons, and in the yellow pack have three oranges, six apples and one lemon. Presently guess us haphazardly pick one of the containers and from that case we arbitrarily collect a thing, and have seen which kind of thing it is. All the while, we supplant the thing in the case from which it came, and it could envision rehashing this cycle commonly. Allow us to assume that there is taken the pink pack 40% and the yellow pack 60% of the time, and that when there is collected a thing from a container we are similarly prone to choose some things in the crate.

Let us specify randomly variable *Y* to represent the pack us make choice, then it has

p(*Y* = *p*) = 4/10 and

p(*Y* = *y*) = 6/10,

where $p(Y = p)$ is the marginal probability that we make choice the pink pack, and $p(Y = y)$ is the minor likelihood that pick the yellow pack. Assume that we take a crate indiscriminately, and afterward the likelihood of choosing a thing is the negligible part of that thing given the chose pack, which can be composed as the accompanying contingent probabilities

$$p(X = o|Y = p) = 2/12 \qquad (3.4)$$

$$p(X = a|Y = p) = 4/12 \qquad (3.5)$$

$$p(X = l|Y = p) = 6/12 \qquad (3.6)$$

$$p(X = o|Y = y) = 3/10 \qquad (3.7)$$

$$p(X = a|Y = y) = 6/10 \qquad (3.8)$$

$$p(X = l|Y = y) = 1/10. \qquad (3.9)$$

Note that these probabilities are normalized so that

$$p(X = o|Y = p)+ p(X = a|Y = p)+ p(X = l|Y = p) = 1$$

and

$$p(X = o|Y = y)+ p(X = a|Y = y)+ p(X = l|Y = y) = 1.$$

Presently guess a thing has been chosen and it is an orange, and it might want to realize which box it came from. This expects that we assess the likelihood conveyance over packs adapted on character of the thing, though the probabilities in mathematical statement (3.4) - (3.9) represent the circulation of thing molded on personality of the case. In view of Bayes' hypothesis, it can compute the back likelihood by turning around the restrictive likelihood.

$$\begin{aligned} \mathrm{P(Y{=}p|X{=}o)} &= \frac{\mathrm{p(X{=}o|Y{=}p)p(Y{=}p)}}{\mathrm{P(X{=}o)}} \\ &= \frac{2/12\times 4/10}{37/150} \\ &= 10/37 \end{aligned}$$

where the total probability of deciding an orange $p(X = o)$ can be computed by applying the sum and produce laws.

$$\begin{aligned} P(X{=}o) &= p(X{=}o|Y{=}p)p(Y{=}p)+p(X{=}o|Y{=}y)p(Y{=}y) \\ &= 2/12\times 4/10 +3/10\times 6/10 \\ &=37/150 \end{aligned}$$

From the aggregate rule, it then, at that point, sees that p(Y = p|X = o) = 1−10/37= 27/37. Overall reasons, we are keen on the probabilities of the groups known the information tests.

Assume, it can be utilized irregular variable Y to signify the group name for information tests, and arbitrary variable X to address the component of information tests. It can decipher p(Y = Ck) as the earlier likelihood for the group Ck, which addresses the likelihood that the group name of an information test is Ck before there will be noticed the information test. When there is noticed the component X of an information test, it can then utilize Bayes hypothesis to process the comparing back likelihood p(Y|X). The amount p(X|Y) can be communicated as how plausible the noticed information X is for various groups, which is known as the probability.

Memo that the probability is not a likelihood dispersion over Y, and its essential regarding Y does not be guaranteed to rise to one. Considering this meaning of probability, we can express Bayes hypothesis as back ∝ probability × earlier. Since there is presented Bayes hypothesis, in the following subsection, there will see the way Bayes hypothesis is utilized in the Naive Bayesian classification.

3.1.2 Naive Bayesian Classification

Naive Bayesian classification is known to be the least complex Bayesian classification, and it has turned into a significant probabilistic model and has been strikingly fruitful practically speaking regardless of areas of strength for its suspicion. Gullible Bayes has demonstrated compelling in text characterization, clinical analysis, and PC execution the board, among different applications. In the accompanying subsections, it will depict the type of Naive Bayesian classification, and greatest probability gauges as well as its applications.

Issue set: Let the user initially characterize the issue set as follows: Assume, there have a bunch of preparing set {(x(i),y(i))} comprising of N models, each x(i) is a d-layered highlight path, and each y(i) means the group mark for the model. There will be accepted arbitrary factors Y and X with parts X1, ...,Xd comparing to the name y and the component path x = _x1,x2, ...,xd_. Note that the superscript is utilized to record preparing models for I = 1, ...,N, and the addendum is utilized to allude to each element or irregular changeable of a path. By and large, Y is a discrete changeable that falls into precisely one of K potential groups {Ck} for k ∈ {1, ...,K}, and the elements of X1, ...,Xd can be any discrete or ceaseless properties.

The user job is to prepare a classifier that will output the posterior probability *p*(*Y*|*X*) for possible values of *Y*. Equal Bayesian theory, *p*(*Y* =*Ck*|*X* = *x*) can be described as

$$\begin{aligned} p(Y = C_k | X = x) &= \frac{p(X = x | Y = C_k) p(Y = C_k)}{p(X = x)} \\ &= \frac{p(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d | Y = C_k) p(Y = C_k)}{p(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d)} \end{aligned} \qquad (3.10)$$

One way to study *p*(*Y*|*X*) is to use the training information to analysis *p*(*X*|*Y*) and *p*(*Y*). We can then use these estimation, together with Bayesian theory, to take a decision *p*(*Y*|*X* = *x*(*i*)) for any another instance *x*(*i*).

Learning precise Bayesian classifiers is commonly immovable. Taking into account the case that Y is boolean and X is a path of d boolean highlights, the user really want to gauge roughly 2d boundaries p(X1 = x1, X2 = x2, ..., Xd = xd | Y =Ck). That's what the explanation is, for a specific worth Ck, there are 2d potential upsides of x, which require to register 2d −1 free boundaries. Given twice potential qualities for Y, we want to gauge a sum of 2(2d −1) such boundaries. Also, to get solid evaluations of every one of these boundaries, we should notice every one of these particular occasions on numerous occasions, which is plainly ridiculous in most commonsense order spaces. For instance, on the off chance that X is a path with 20 boolean highlights, the user should gauge more than 1 million boundaries.

To deal with the unmanageable example intricacy for studying the Bayesian classification, the Naive Bayesian classification diminishes this intricacy by making a restrictive freedom presumption that the highlights X1, ...,Xd are restrictively free of each other, given Y. For the past case, this restrictive freedom supposition serves to emphatically diminish the quantity of boundaries to be assessed for demonstrating p(X|Y) from the first 2(2d −1) to simply 2d. Think about the probability p(X = x| Y = Ck) of mathematical statement (3.10), the user have

$$
\begin{aligned}
& p(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d | Y = C_k) \\
= & \prod_{j=1}^{d} p(X_j = x_j | X_1 = x_1, X_2 = x_2, \ldots, X_{j-1} = x_{j-1}, Y = C_k) \\
= & \prod_{j=1}^{d} p(X_j = x_j | Y = C_k). \qquad (3.11)
\end{aligned}
$$

The two row follows from the chain law, the general property of probabilities, and the third line follows direct from the above conditional independence, that the rate for the random variable *Xj* is independent of all other character rates, *Xj_* for *j_* = *j*, when conditioned on the identity of the label *Y*. This is the *Naive Bayesian* appropriation. It is a relation well-made and especially useful for appropriation. When *Y* and *Xj* are boolean changeables, we only require 2*d* parameters to specify *p*(*Xj*|*Y* =*Ck*).

After acting mathematical statement (3.11) in mathematical statement (3.10), we can take the basic equation for Naïve Bayesian classification

$$p(Y = C_k | X_1 \ldots X_d) = \frac{p(Y = C_k) \prod_j p(X_j | Y = C_k)}{\sum_i p(Y = y_i) \prod_j p(X_j | Y = y_i)}. \qquad (3.12)$$

If there are interested only in the most probable cost of *Y*, then there have the Naive Bayesian classifier law:

$$Y \leftarrow \arg\max_{C_k} \frac{p(Y = C_k) \prod_j p(X_j | Y = C_k)}{\sum_i p(Y = y_i) \prod_i p(X_j | Y = y_i)}, \qquad (3.13)$$

Because the denominator does not be based on *Ck*, the first class formulation can be make easy to the following

$$Y \leftarrow \arg\max_{C_k} p(Y = C_k) \prod_j p(X_j | Y = C_k). \qquad (3.14)$$

3.1.3 Maximum Likelihood Estimation for Naive Bayesian Methods

In numerous viable functions, boundary assessment for Naive Bayesian methods utilizes the strategy for greatest probability gauges. To sum up, the Naive Bayesian method has two kinds of boundaries that should be assessed. Firstly,

$$\pi k \equiv p(Y = Ck)$$

for any of the possible rates *Ck* of *Y*. The parameter can be interpreted as the probability of seeing the label *Ck*, and there have the constraints $\pi k \geq 0$ and $\Sigma^{K}_{k=1} \pi k = 1$. Memo there are K of these parameters, $(K-1)$ of which are unconnected.

For the *d* input characters *Xi*, assume each can take on *J* possible discrete rates, and there will be used $Xi = xi\ j$ to signify that. Secondly,

$$\theta i\ jk \equiv p(Xi = xi\ j|Y = Ck)$$

for each information include Xi, every one of its potential qualities xi j, and every one of the conceivable valuesCk of Y. The incentive for θi jk can be deciphered as the likelihood of component Xi taking worth xi j, molded on the hidden mark being Ck. Memo that they should fulfill $\Sigma j\ \theta i\ jk = 1$ for each sets of I, k qualities, and there will be dJK such boundaries, and memo that main d(J−1)K of these are autonomous. These boundaries can be assessed utilizing most extreme probability gauges in light of working out the general frequencies of the various occasions in the information. Greatest probability gauges for θi jk given a bunch of preparing models are

$$\hat{\theta}_{ijk} = \hat{p}(X_i = x_{ij}|Y = C_k) = \frac{count(X_i = x_{ij} \wedge Y = C_k)}{count(Y = C_k)} \tag{3.15}$$

where *count*(*x*) return the numeral of examples in the train sets that make the grade property *x*, e.g., $calculate(Xi = xi\ j\ \wedge Y = Ck) = \Sigma^{N}_{n=1}\{X^{(n)}i = xi\ j\ \wedge Y^{(n)} = Ck\}$, and $calculate(Y = Ck) = \Sigma^{N}_{n=1}\{Y^{(n)} = Ck\}$. This is an exceptionally regular gauge: We basic count the times name Ck is found related to Xi taking worth xi j, and calculate the time the mark Ck is found altogether, and afterward take the proportion of these two terms.

To keep away from the case that the information doesn't end up containing any preparation models fulfilling the condition in the numerator, it is normal to adjust a smoothed gauge that successfully includes some of extra daydreamed models similarly over the potential upsides of Xi. The smoothed gauge is given by

$$\hat{\theta}_{ijk} = \hat{p}(X_i = x_{ij}|Y = C_k) = \frac{count(X_i = x_{ij} \wedge Y = C_k) + l}{count(Y = C_k) + lJ}, \tag{3.16}$$

where *J* is the numeral of distinct rates that *Xi* can get on, and *l* decides the strength of this smoothing. If *l* is set to 1, this procedure is called Laplace smoothing.

Maximum likelihood predicts for *πk* get the following set

$$\hat{\pi}_k = \hat{p}(Y = C_k) = \frac{count(Y = C_k)}{N}, \qquad (3.17)$$

where $N = \Sigma^{K}_{k=1} count(Y = C_k)$ is the numeral of examples in the train sets. Alike, it can obtain a smoothed predict by using the following set

$$\hat{\pi}_k = \hat{p}(Y = C_k) = \frac{count(Y = C_k) + l}{N + lK}, \qquad (3.18)$$

where *K* is the numeral of distinct rates that *Y* can get on, and *l* again decides the strength of the prior presumptions related to the observed data.

3.2 Probabilistic and Naive Bayesian Classification

Probabilistic classifiers are intended to utilize an implied combination type for age of the fundamental records. This combination type normally expects that each class is a part of the blend. Every combination part is basically a generative model, which gives the likelihood of testing a specific term for that part or class. To this end this sort of classifiers are in many cases likewise called generative classifiers. The guileless Bayesian classifier is maybe the least difficult and furthermore the most usually utilized generative classifier. It demonstrates the dispersion of the records in each class utilizing a probabilistic type with freedom presumptions about the conveyances of various terms. Two classes of types are usually utilized for innocent Bayesian grouping. The two types basically process the back likelihood of a class, in view of the dispersion of the words in this report. These types overlook the real place of the orders in the report, and work with the "sack of orders" presumption. The significant distinction stuck between these two types is the presumption regarding taking (or not taking) word frequencies into account, and the comparing way for examining the likelihood space:

- Multivariate Bernoulli Type: In this type, it can be utilized the presence or nonappearance of the words in a message report as elements to address a record. Consequently, the frequencies of the words are not utilized for the displaying a report, and the word highlights in the text are thought to be double, with the two qualities showing presence or nonattendance of a word in text. Since the elements to be demonstrated are paired, the type for reports in each class is a multivariate Bernoulli type.
- Multinomial Type: In this type, it catches the frequencies of terms in a record by addressing a report with a pack of words. The records in each class can then be displayed as tests drawn from a multinomial word conveyance. Thus, the restrictive likelihood of a record given a class is essentially a result of the likelihood of each noticed word in the relating class.

Regardless of how it model the records in each class (be it a multivariate Bernoulli model or a multinomial model), the part class models (i.e., generative models for reports in each class) can be utilized related to the Bayesian rule to figure the back likelihood of the group for a given record, and the group with the most elevated back likelihood can then be doled out to the report.

There has been significant disarray in the writing on the distinctions between the multivariate Bernoulli model and the multinomial model. A decent piece of the distinctions between these two models might be found. The accompanying will portray these two models in more detail.

3.2.1. Bernoulli Multivariate Type

1. The class of strategies regards a record as a bunch of unmistakable words with no recurrence data, in which a component (term) might be either present or missing. Allow it will be accepted that the dictionary from which the terms are drawn are signified byV ={t1 . . .tn}. Let
2. The user expect that the sack of-words (or text report) being referred to contains the terms Q = {ti1 . . .tim}, and the class is drawn from {1. . .k}. Then, the user want to show the back likelihood that the archive (which is

thought to be produced from the term disseminations of one of the classes) has a place with class I, considering that it contains the terms Q = {ti1 . . .tim}. The most effective way to comprehend the Bayes technique is by grasping it as an examining/ generative interaction from the hidden combination model of classes. The Bayes likelihood of class I can be displayed by examining a bunch of terms T from the term dissemination of the classes:

3. Assuming that It is examined a term set T of any size from the term conveyance of one of the haphazardly picked classes, and the ultimate result is the set Q, then what is the back likelihood that we had initially picked class I for inspecting? The deduced likelihood of picking class I is equivalent to its partial presence in the assortment.
4. The user signify the class of the examined set T by CT and the relating back likelihood by P(CT = i|T = Q). This is basically the very thing we are attempting to find. It is critical to take note of that since the user do not permit substitution, the users are basically picking a subset of terms from V without any frequencies connected to the picked terms. In this manner, the set Q may not contain copy components. Under the innocent Bayes suspicion of freedom between terms, this is basically identical to either choosing or not choosing each term with a likelihood that relies on the fundamental term dissemination. Moreover, it is likewise vital to take note of that this model has no limitation based on the quantity of conditions picked. As the user will see later, these presumptions are the vital contrasts with the multinomial Bayes model. The Bayes approach characterizes a given set Q in light of the back likelihood that Q is an example from the information conveyance of class I, i.e., P(CT = i|T = Q), and it expects the user to figure the accompanying two probabilities to accomplish this:
5. What is the prior probability that a put T is a sample from the term distribution of class i? This probability is denoted by $P(C^T = i)$.
6. If there is sampled a set T of any size *from the term distribution of class i*, then what is the probability that our sample is the set Q? This probability is denoted by $P(T = Q|C^T = i)$.

The user will now provide a more mathematical description of Bayes modeling. In other words, the user wish to model $P(C^T = i|Q$ is sampled). The user can use the Bayes rule in order to write this conditional probability in a way that can be *estimated* more easily from the underlying corpus. In other words, it can simplify as follows:

$$
\begin{aligned}
P(C^T = i|T = Q) &= \frac{P(C^T = i) \cdot P(T = Q|C^T = i)}{P(T = Q)} \\
&= \frac{P(C^T = i) \cdot \prod_{t_j \in Q} P(t_j \in T|C^T = i) \cdot \prod_{t_j \notin Q} (1 - P(t_j \in T|C^T = i))}{P(T = Q)}
\end{aligned}
$$

The last state of the above succession utilizes the innocent autonomy suspicion, since we are expecting that the probabilities of event of the various terms are free of each other. This is basically vital, to change the likelihood conditions to a structure that can be assessed from the fundamental information.

The class appointed to Q is the one with the most elevated back likelihood given Q. It is not difficult to see that this choice isn't impacted by the denominator, which is the negligible likelihood of noticing Q. That is, the user will relegate the accompanying class to *Q*:

$$
\begin{aligned}
\hat{i} &= \arg\max_i P(C^T = i|T = Q) \\
&= \arg\max_i P(C^T = i) \cdot \\
&\quad \prod_{t_j \in Q} P(t_j \in T|C^T = i) \cdot \prod_{t_j \notin Q} (1 - P(t_j \in T|C^T = i))
\end{aligned}
$$

It is vital to take memo of that all terms in the right hand-side of the last condition can be assessed from the preparation corpus. The worth of P(CT = I) is assessed as the worldwide part of the reports having a place with group I, the worth of P(t j ∈ T|CT = I) is the small portion of records in the ith group that contain term t j. We memo that the above are every one greatest probability evaluations of the comparing probabilities. By and by, Laplacian smoothing is utilized, in which little qualities are added to the frequencies of terms to keep away from no probabilities of meagerly present terms. In many uses of the Bayesian classifier, we just consideration about the character of the class with the most elevated likelihood esteem, as opposed to the genuine likelihood esteem related with it, which is the reason we don't have to figure the normalizer P(T = Q). Truth be told, on account of paired classes, various improvements are conceivable in processing these Bayes "likelihood" values by utilizing the logarithm of the Bayes articulation, and eliminating various terms that don't influence the requesting of class probabilities.

Despite the fact that for arrangement, we don't have to register P(T = Q), a few applications require the specific calculation of the back likelihood P(CT = i|T = Q). For instance, on account of directed peculiarity identification (or uncommon class location), the specific back likelihood esteem P(CT = i|T = Q) is required to reasonably think about the likelihood esteem over various test occasions, and rank them for their strange nature. In such cases, the user would have to figure P(T = Q). One method for accomplishing this is just to take a total over all the classes:

$$P(T=Q)=\sum_i P(T=Q|C^T=i)P(C^T=i)$$

This depends on the contingent freedom of elements for each group. Since the boundary values are assessed for each group independently, the user might deal with the issue of information scantiness. An elective approach to processing it, which might lighten the information inadequacy issue, is to additional make the suspicion of (worldwide) freedom of terms, and register it as:

$$P(T=Q)=\prod_{j\in Q} P(t_j\in T)\cdot\prod_{t_j\notin Q}(1-P(t_j\in T))$$

where the term probabilities depend on worldwide term conveyances in every one of the classes. A characteristic inquiry emerges, concerning whether it is feasible to plan a Bayesian classification that does not utilize the innocent presumption, and types the conditions between the terms during the characterization cycle. Techniques that sum up the gullible Bayesian classification by not utilizing the autonomy supposition do not function admirably in light of the greater computational expenses and the powerlessness to gauge the boundaries precisely and heartily within the sight of restricted information. On the one limit, a supposition of complete reliance brings about a Bayesian organization model that ends up being computationally pricey. Then again, it has been shown that permitting restricted degrees of reliance can give great tradeoffs among precision and computational expenses.

While the freedom supposition that is a down to earth guess, it has been showing that the methodology has some hypothetical legitimacy. To be sure, broad exploratory tests have would in general show that the credulous classifier functions admirably practically speaking.

The Bayes strategy gives a characteristic method for integrating such extra data into the grouping system, by making new elements for every one of these qualities. The standard Bayes procedure is then utilized related to this expanded portrayal for order. The Bayes strategy has additionally been utilized related to the joining of different sorts of space information, for example, the consolidation of hyperlink data into the order interaction.

The Bayes strategy is likewise fit to various leveled order, while the preparation information is organized in a scientific categorization of points. For instance, the Open Directory Project (ODP), Yahoo! Scientific classification, and an assortment of information locales have tremendous assortments of records that are organized into progressive gatherings. The progressive construction of the points can be taken advantage of to perform more viable characterization, since it has been seen that setting delicate element determination can give more helpful grouping results. In progressive order, a Bayes classifier is worked at every hub, which then, at that point, furnishes us with the following branch to follow for characterization purposes. Two such strategies are proposed, in which hub explicit elements are utilized for the grouping system. Obviously, many less highlights are expected at a specific hub in the order, on the grounds that the elements that are picked are pertinent to that branch.

3.2.2 Multinomial Spreading

This group of methods regards a record as a bunch of words with frequencies joined to each word. Consequently, the arrangement of words is permitted to have copy components. As in the past case, the user accept that the arrangement of words in archive is signified by Q, drawn from the jargon set V. The set Q contains the unmistakable terms {ti1 . . .tim} with related frequencies F = {Fi1 . . .Fim}. There will be indicate the terms and their frequencies by [Q,F]. The complete number of terms in the report (or archive length) is signified by L = Σmj =1 F(i j). Then, at that point, there will be want to display the back likelihood that the report T has a place with group I, considering that it contains the terms in Q with the related frequencies F. The Bayes likelihood of group I can be displayed by utilizing the accompanying testing process:

Assuming, there is examined L terms consecutively from the term conveyance of one of the arbitrarily picked groups (permitting reiterations) to make the term set T, and the ultimate result for tested set T is the set Q with the comparing frequencies F, then, at that point, what is the back likelihood that we had initially picked class I for inspecting? The deduced likelihood of picking group I is equivalent to its fragmentary presence in the assortment.

The previously mentioned likelihood is indicated by P(CT = i|T = [Q,F]). A supposition that is regularly utilized in these models is that the length of the archive is free of the group name. While it is effectively conceivable to sum up the technique, with the goal that the record length is utilized as an earlier, freedom is generally expected for effortlessness. As in the past case, it is really want to appraise two qualities to process the Bayes back.

1. What is the prior probability that a put T is a example from the phrase spreading of group i? This probability is represented by $P(C^T = i)$.
2. If we exampled L phrases *from the phrase spreading of group i* (with repetitions), then what is the probability that we exampled put T is the put Q with related frequencies F? This probability is represented by $P(T = [Q,F]|C^T = i)$.

$$P(C^T = i|T = [Q,F]) = \frac{P(C^T = i) \cdot P(T = [Q,F]|C^T = i)}{P(T = [Q,F])}$$

$$\propto P(C^T = i) \cdot P(T = [Q,F]|C^T = i). \tag{3.19}$$

As in the early state, it is not necessary to calculate the denominator, $P(T = [Q,F])$, for the purpose of determining the group label for Q. The rate of the probability $P(C^T = i)$ can be predicted as the fraction of documents belonging to group i. The calculation of $P([Q,F]|C^T = i)$ is more complicated. When the user examine the serial order of the L other samples, the number of possible ways to example the other phrases so as to result in the outcome [Q, F] is given by L! Π*mi* =1 *Fi*! . The probability of *each* of these series is given by Π*t jЄQP*(*t j Є T*)^{Fj} , by using the naive independence assumption. So, the users have:

$$P(T = [Q,F]|C^T = i) = \frac{L!}{\prod_{i=1}^{m} F_i!} \cdot \prod_{t_j \in Q} P(t_j \in T|C^T = i)^{F_j}. \tag{3.20}$$

Substitute Equation 3.20 in Equation 3.19 to acquire the class with the most elevated Bayes back likelihood, where the group priors are processed as in the past case, and the probabilities P(t j ∈ T|CT = I) can likewise be effectively assessed as already with Laplacian smoothing. Note that to pick the class with the most elevated back likelihood, we don't actually need to figure L! Πmi =1 Fi!, as it is a consistent not relying upon the class mark (i.e., the equivalent for every one of the classes). We likewise memo that the probabilities of class nonattendance are absent in the above conditions as a result of the manner by which the examining is performed.

Various varieties of the multinomial type have been purposed. In the work, it is demonstrated the way that a classification pecking order can be utilized to work on the gauge of multinomial boundaries in the credulous Bayes classifier to further develop grouping exactness essentially. The key thought is to apply shrinkage methods to smooth the boundaries for information inadequate youngster classifications with their normal parent hubs. Subsequently, the preparation information of related classifications are basically "shared" with one another in a weighted way, which works on the power and exactness of boundary assessment when there are lacking preparation information for every individual kid class. The work has played out a broad examination between the Bernoulli and the multinomial models on various corpora, and the accompanying ends were introduced:

- The multi-variate Bernoulli type can once in a while perform better compared to the multinomial type at little jargon sizes.
- Multinomial type outflanks the multi variate Bernoulli type for enormous jargon sizes, and quite often beats the multi variate Bernoulli when jargon size is decided ideally for both. On the normal a 27% decrease in blunder.

The advance of referenced results strongly imply that the two types might have various qualities and could thusly be valuable in various situations.

3.3 Calculation Accuracy

The method used to determine the final accuracy of tests performed is the confusion matrix for the multi-class method. This technique is utilized to perform framework estimations with numerous forecast classes. The distinction from the multi-class disarray network with the normal disarray framework is that the end-product are determined combined precision of the general exactness of all test information. Boundaries of the precision are introduced in Table 3.1

Classification	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3.1. Confusion matrix

where:

True Negative(TN) : if the prediction and actual results are negative
False Negative (FN) : if the positive prediction results, and the actual results negative
False Positive (FP) : if the negative prediction results, and the actual results positive
True Positive (TP) : if the predictive and actual results are positive.

Calculation of the total accuracy of the tests performed using the following formula.

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

$$\text{precision} = \frac{TP}{TP+F\text{P}}$$

$$\text{recall} = \frac{TP}{TP+F\text{N}}$$

$$\text{f\_Measure} = \frac{2\times\ (\text{precision x recall})}{(\text{precision + recall})}$$

CHAPTER 4

SYSTEM DESIGN AND IMPLEMENTATION

Mushrooms are the most natural delightful food which is cholesterol free as well as plentiful in nutrients and minerals. Numerous types of mushrooms have been known all through the world. Recognizing eatable or noxious mushroom through the unaided eye is very troublesome, so mushroom species should have to group consumable and harmful. This framework will be characterized the sort of mushroom by utilizing Naive Bayesian classifier and K-Nearest Neighbor Method to develop valuable subset of mushroom highlights for arrangement task.

4.1. Myanmar mushroom species

Data Source: Ministry of Agriculture, Livestock and Irrigation, Department of Agriculture, Mandalay Division and papers from on the Web.

There are many species of mushroom in Myanmar. Mushrooms were collected, preserved, organized, separated and explained. 46 species of mushrooms are collected. The number of mushroom species are shown in Table 4.1.

No.	Species
1.	Zaung - Pya – Hmo (Edible)
2.	Kywet - na - ywet - Hmo (ear fungus) (Edible)
3.	Wa Yaung - Hmo (Lentinus squarrosulus) (Edible)
4.	Tha - Yet - Hmo (Clitocybe caespitosa Pk.) (Edible)
5.	Nya - Hmo (Corpinus disseminatus (Curt) Fr.) (Edible)
6.	In - U (Astraeus hygrometricus) (Edible)
7.	Hmo - Chin - Taung (Dicytophora indusiata (Pers.) Fish) (Edible)
8.	Hmo - thanguin sut (Lepiota morgani Pk.) (Poisonous)
9.	Myet - Kya - Hmo - U (Lycoperdon wrightii Berk. and Curt.)(Edible)
10.	Hmo - Chay – To (Russula delica (Pres.) Fr.) (Edible)
11.	Kun - Tatawe – Hmo (Russula emetica (Schaeff.) Pers.) (Poisonous)
12.	Earth ball (Scleroderma citrinum Pers.) (Edible)
13.	Taung - Bo – Hmo (Termitomyces schimperi (Pat) Hein) (Edible)
14.	Hmo - Ohn – nat (Termitomyces cartilagineus) (Edible)

15.	Straw Mushroom (Volvariella volvacea Bull. Fr.) (Edible)
16.	Button Mushroom (Edible)
17.	Monkey Head Mushroom (Edible)
18.	Black Forest Mushroom (Shiitake) (Edible)
19.	Milky Mushroom (Edible)
20.	Narmeko Mushroom (Pholiota microspora) (Edible)
21.	Silver Ear Mushroom (Edible)
22.	Lingzhi Mushroom (Ganoderma lucidum) (Edible)
23.	Coprinus disseminatus (pers) Gray (Edible)
24.	Macrollepiota Konradii (Hujjsman ex. P.D.Orton) (Edible)
25.	Amanita caesarea (Scop.) Per (Edible)
26.	Amanitopsis vaginata (Bull.) Roze. (Edible)
27.	Hygrocybe ceracea (Sowerby) P. Kumm. (Edible)
28.	Boletus pulverulentus Opat. (Edible)
29.	Hmo Seinn Sarr (Lactarius clarkeae Cleland.) (Edible)
30.	Milkcap (Lactarius volemus (Fr.) Fr.) (Edible)
31.	Russula virescens (Schaeff.) Fr. (Edible)
32.	Death Cap (Amanita phalloides) (Poisonous)
33.	Webcaps (Cortinarius sp) (Poisonous)
34.	Oyster Mushroom (Edible)
35.	Orange Jelly (Dacryopinax spathularia) (Edible)
36.	Grey Oyster mushroom (Edible)
37.	King Oyster mushroom (Edible)
38.	White Oyster Mushroom (Edible)
39.	Golden Mushroom (Edible)
40.	Pink Oyster Mushroom (Pleurotus salmoneo stramineus) (Edible)
41.	Inn Tine Ni (Tricholoma sp) (Edible)
42.	Inn Tine Sein (Tricholoma sp) (Edible)
43.	Inn Tine War (Tricholoma sp) (Edible)
44.	Hmo Thin Gan (Canthrellus sp) (Edible)
45.	Hmo War Tar (Cantharellus sp) (Edible)
46.	Mho Auu (Calvatia sp) (Edible)

Table 4.1: Lists of collected Myanmar mushroom species

4.2. Myanmar Mushroom Attributes Information

Data Source: Ministry of Agriculture, Livestock and Irrigation, Department of Agriculture, Mandalay Division and papers from on the Web.

The classification of this datasets was conducted to classify the mushrooms whether edible or poisonous based on its behavioural features. The dataset contained 16 numbers of attributes (features). These attributes are shown in Table 4.2.

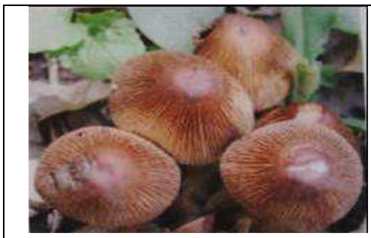
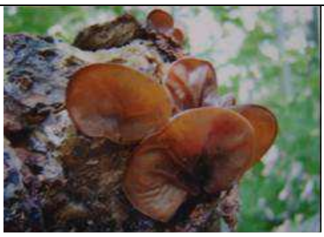
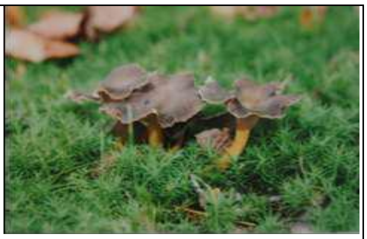
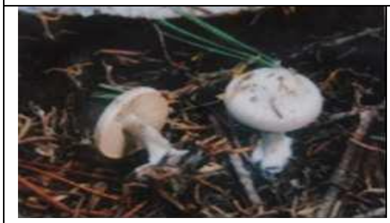
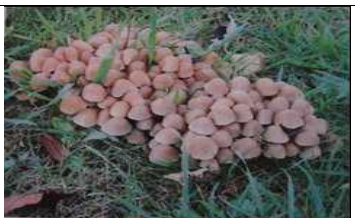
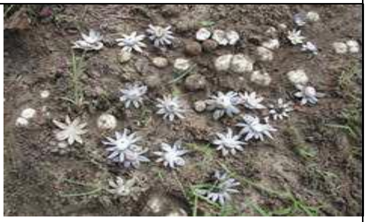
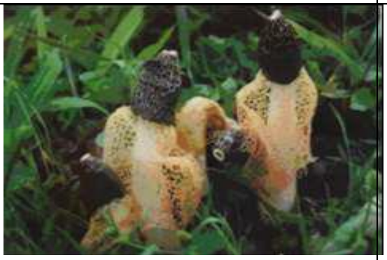
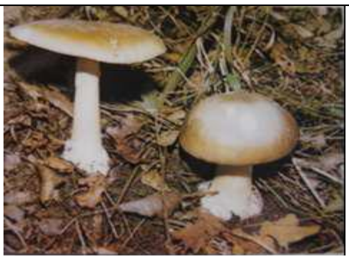
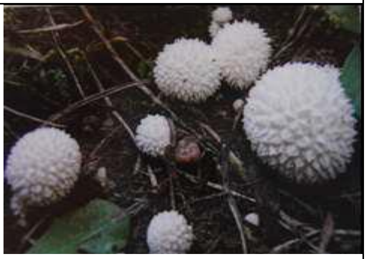
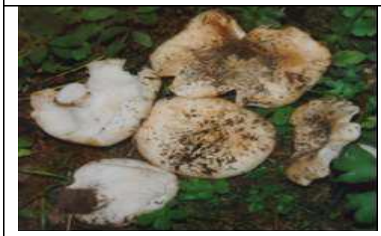
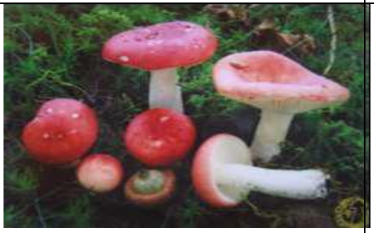
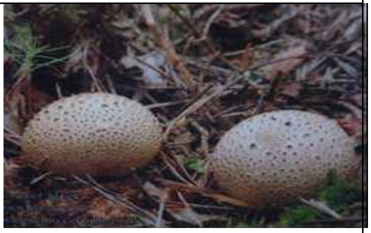
Attribute No.	Attribute Name	Descriptions
1	cap-color	white to pale gray, white, orange-red, leaden-brown, pale-orange, dark-brown, orange-brown, grey-brown, dull green, red, gray, silver, brown, buff yellow, brownish yellow, greenish yellow, golden, golden brown, grayish brown, pink, orange, pale, brick-red, yellow, bright yellow, purple
2	cap-shape	campanulate, expanded, convex, convex to depressed, convex with depression, globose, ear-shaped, bell, flat, depressed, puffball, round, umbrella-shaped, funnel-shaped, lobed, kidney-shaped, skirt, pattern, conical, ball, shell, fan, irregular
3	cap-surface	Fertile, flat scales, smooth, waxy, powdery, velvety, fibrous, rough, dry, hard, silky
4	cap-umbonate	present, Slightly present, Slightly, Absent
5	gill-color	grayish-brown, white, yellow, pale-yellow, creamy-white, golden-yellow brown, dark- brown, black, pale pink, cinnamon brown, pink, red, chocolate, purple-gray, cream, absent
6	gill-attachment	free, adnate, adnate to decurrent, decurrent, attached, adnexed, absent
7	gill-spacing	close , crowded , distant, absent
8	stipe-color	white, yellow, reddish brown, orange-brown, red, brown, gray, cream, pink, black, absent
9	stipe-shape	slender, equal, unequal, conical, fan, cup, curved, club, cylindrical, fusiform, rhizoids, fibrous, asymmetrical, flat, bulbous, elliptic, tubular, straight, absent

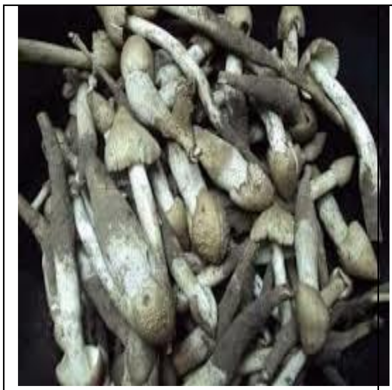
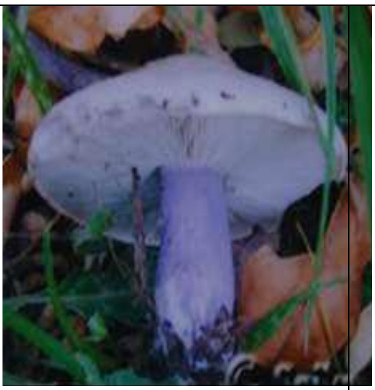
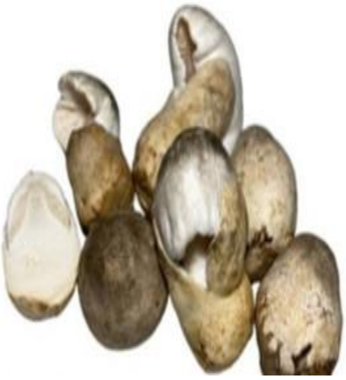
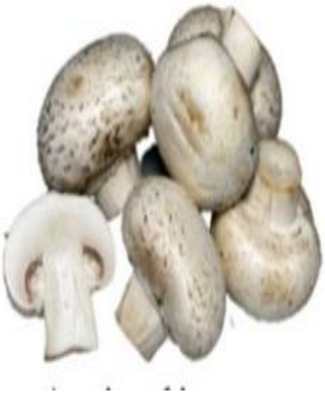
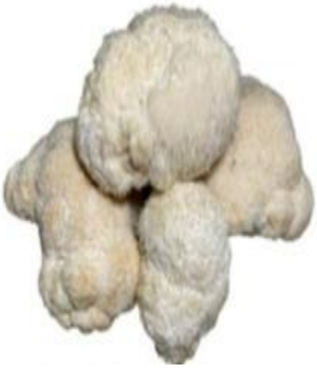
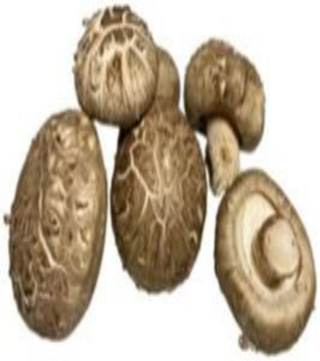
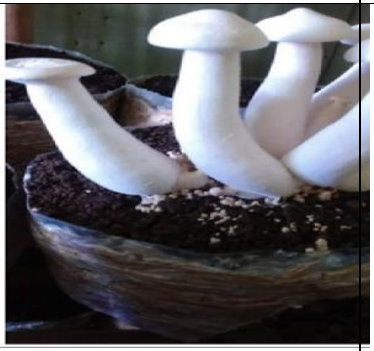
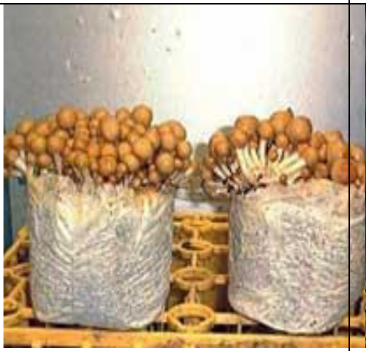
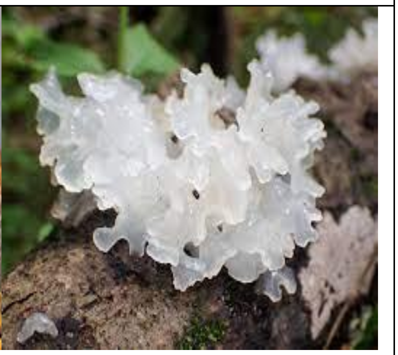
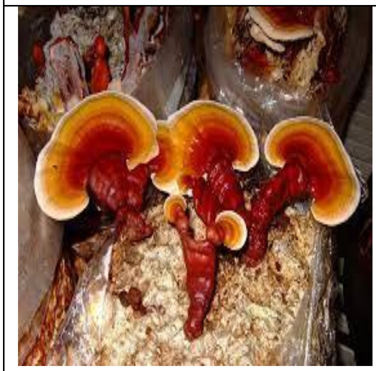
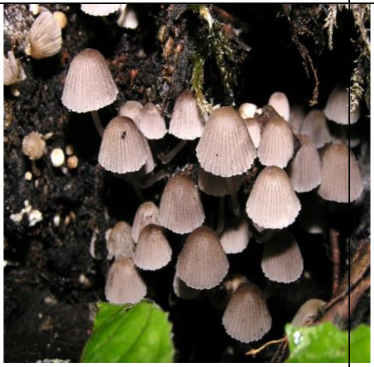
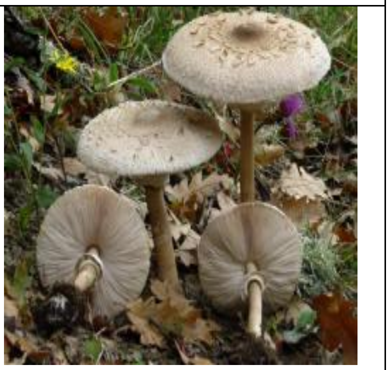
10	stipe	hollow, solid, short, long, fleshy, truncated, thin, thick, dry, absent
11	annulus or ring	absent, present, double
12	spore-color	dark-brown, white, pink, olive-brown, brown, brownish black, rosy, purple-brown, red, yellow, cinnamon, pale, gray
13	spore-shape	elliptic, globose, fusiform, oblongoid, cylindrical, round, subelliptic, tellipsoid, broadly elliptic, angular, club, curved, tender, tropical, conical, bean, amygdaloid,
14	spore-texture	smooth, rough, smooth apical germ pore, spring with faint reticulum, sordid, ovate, fibrous, amyloid, meaty, Jelly-like
15	spore-size	6-7.2×4.8-4.8 µm, 8.4-11.4×6-7.2 µm, 6-7.2×6-6 µm, 8.4-12×6-8.4 µm, 8.4-10.8×4.8-6 µm, 6-7.2×3.6-4.8 µm, 7.2-9.6×4.8-6 µm, 6-8.4×6-7.2 µm, 7.2-7.2×7.2-7.2 µm, 6-7.2×4.8-6 µm, 10 - 12µm, 10 - 15 ×4. 0 - 6.0 µm, 6.0-7.5 ×3.0 - 5.0 µm, 4 -6 × 2 - 3 µm, 9.0 - 10.0 x 6 - 7µm, 12-13µm×3-5µm, 4-6 x 3-4 cm, 12-15x10-12µm, 4-6 × 1-1.5 µm, 8-12 × 7-9 µm, 10-12 × 9-10µm, 7-8 µm, 3 - 6×3 - 5µm, 7-9× 4-5µm, 8.9 ×4.6 µm, 6.8-5-6 µm, 5.9–6.8 µm by 4.2–5.1 µm , 4-6x2.5-3µm, 7.5–11 × 3–4 µm, 6-9 by 2–3.5 micrometers, 5-8 / 4-6 µm, 9 –13 ×6 –8 µm, 8-9× 6-7µm, 8–10 µm, 12-14× 5.5-6.5µm, 8.8-11.0× 5.5-8.0µm, 9-13 by 6.5-9 µm, 7–10 µm, 6.5-8 x 3.5-4.5 µm, 4-5.5 x 2-4.5µm, 10-13 x 5.5-7µm, 6-7× 3-4µm, 6-7× 3-4µ, 14-15× 3-3.5µm, 3.5-5.5× 0.75-1µm, 17-25× 6-8µm, 3.2-4.3µm, 9-13 to 5-7µm, 25-35/3-5µm, 5-10µm, 8-12.5 by 3.5-5µm, 6.8-9.3µm, 8-10 by 5.5-7µm, 6.5-9× 2.8-3.5µm, 5-10× 2-.5µm, 3.5µm, no

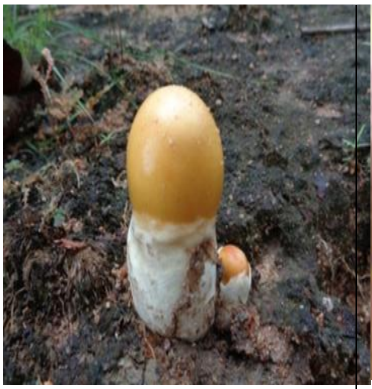
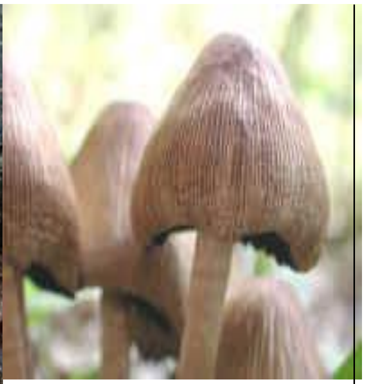
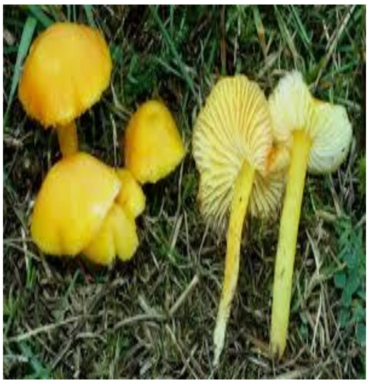
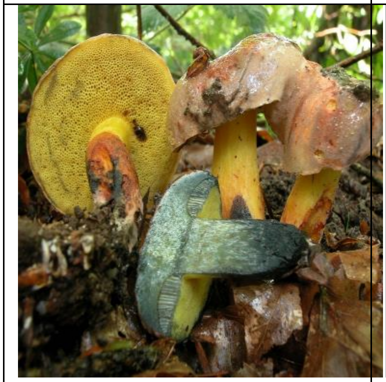
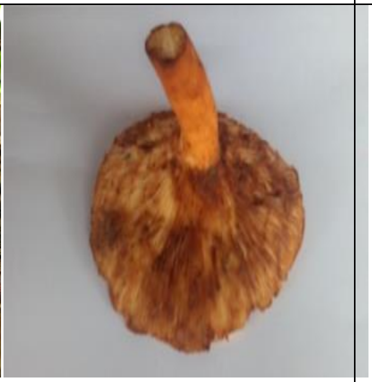
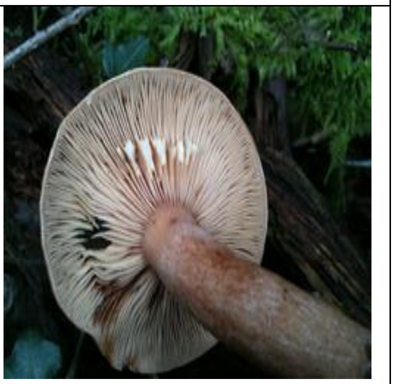
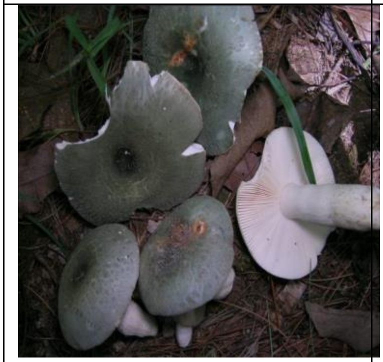
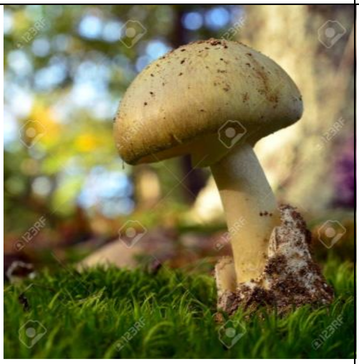
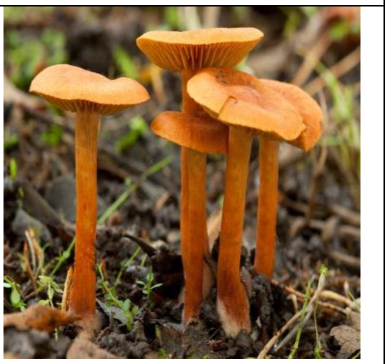
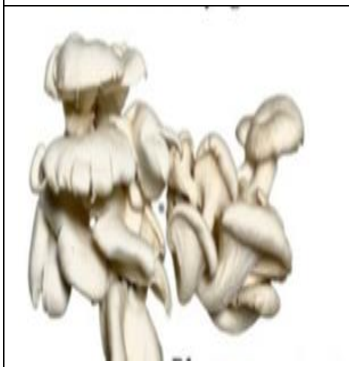
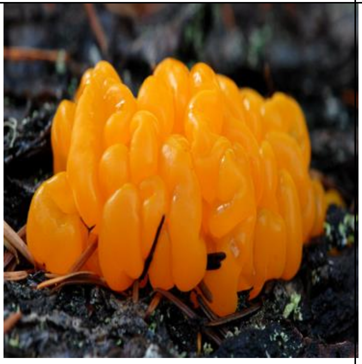
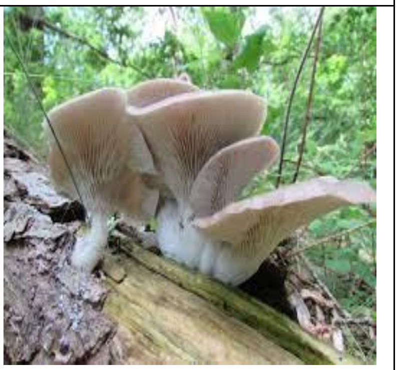
16	growing habitat	Decay woods, woods of deciduous trees, soil, grasses, bamboo, bush, under small tree, paddy straw, logs, houses, oak, woodlands, tree stumps, fields, broad-leaf trees, softwoods, hardwoods, dry trees, underneath the soil
17	class	edible, poisonous

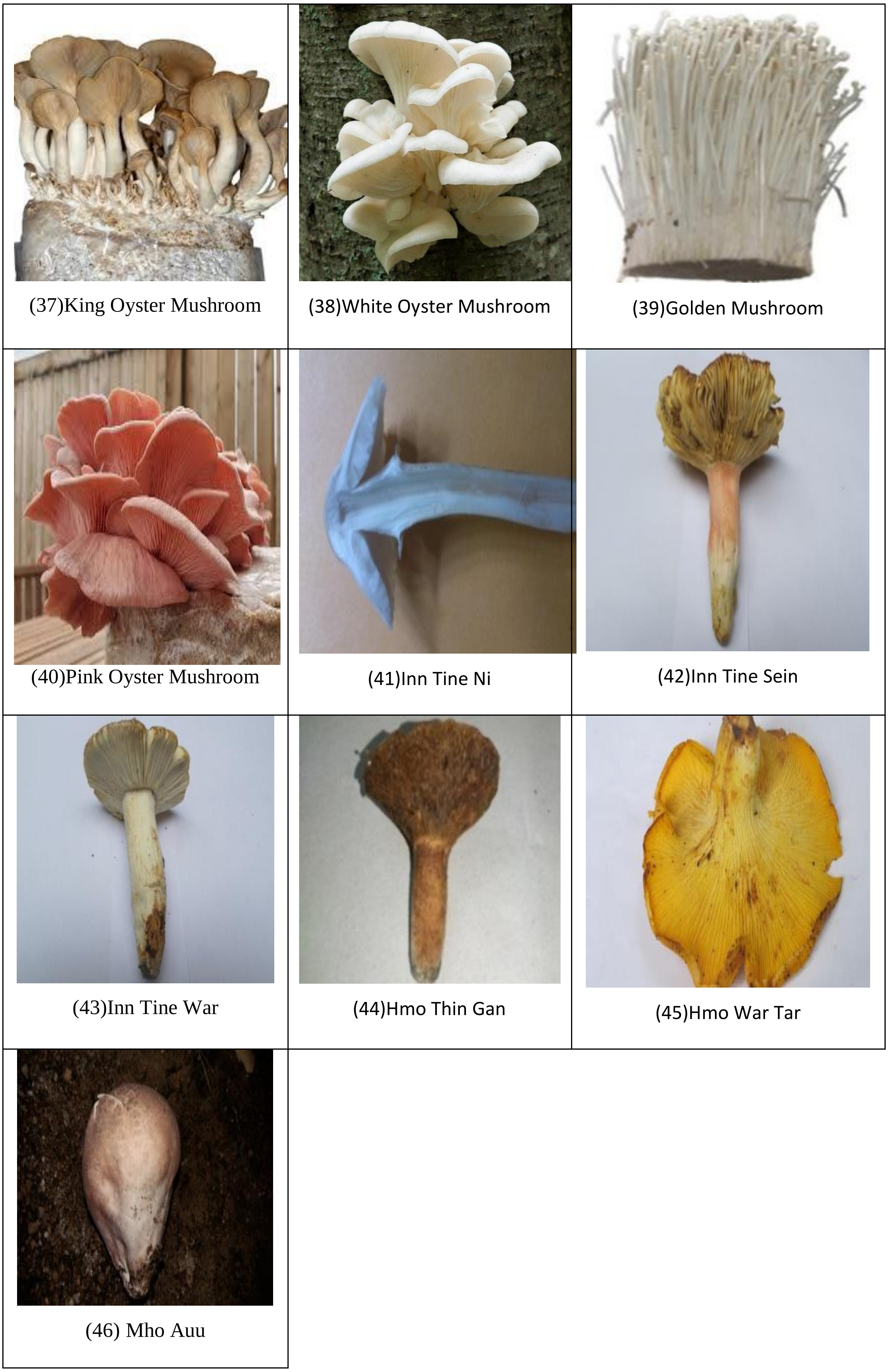
Table 4.2: Lists of collected Mushroom Attributes

4.3. Mushroom Species Photo

 (1)Zaung - Pya- Hmo	 (2)Kywet - na - ywet - Hmo	 (3)Wa Yaung - Hmo
 (4)Tha - Yet - Hmo	 (5)Nya-Hmo	 (6)In- U
 (7)Hmo-Chin-Taung	 (8)Hmo-thanguin sut	 (9)Myet-Kya-Hmo -U
 (10)Hmo - Chay – To	 (11)Kun - Tatawe – Hmo	 (12)Earth ball

 (13)Taung - Bo – Hmo	 (14)Hmo - Ohn – nat	 (15)Straw Mushroom
 (16)Button Mushroom	 (17)Monkey Head Mushroom	 (18)Black Forest Mushroom
 (19)Milky Mushroom	 (20)Narmeko Mushroom	 (21)Silver Ear Mushroom
 (22)Lingzhi Mushroom	 (23)Coprinus disseminatus	 (24)Macrolepiota konradii

 (25)Amanita caesarea	 (26)Amanitopsis vaginata	 (27)Hygrocybe ceracea
 (28)Boletus pulverulentus	 (29)Hmo Seinn Sarr	 (30)Milkcap
 (31)Russula virescens	 (32)Death Cap	 (33)Webcaps
 (34)Oyster Mushroom	 (35)Orange Jelly	 (36)Grey Oyster Mushroom



(37)King Oyster Mushroom

(38)White Oyster Mushroom

(39)Golden Mushroom

(40)Pink Oyster Mushroom

(41)Inn Tine Ni

(42)Inn Tine Sein

(43)Inn Tine War

(44)Hmo Thin Gan

(45)Hmo War Tar

(46) Mho Auu

Figure 4.1: Mushroom Species photo

4.4. System Flow Diagram

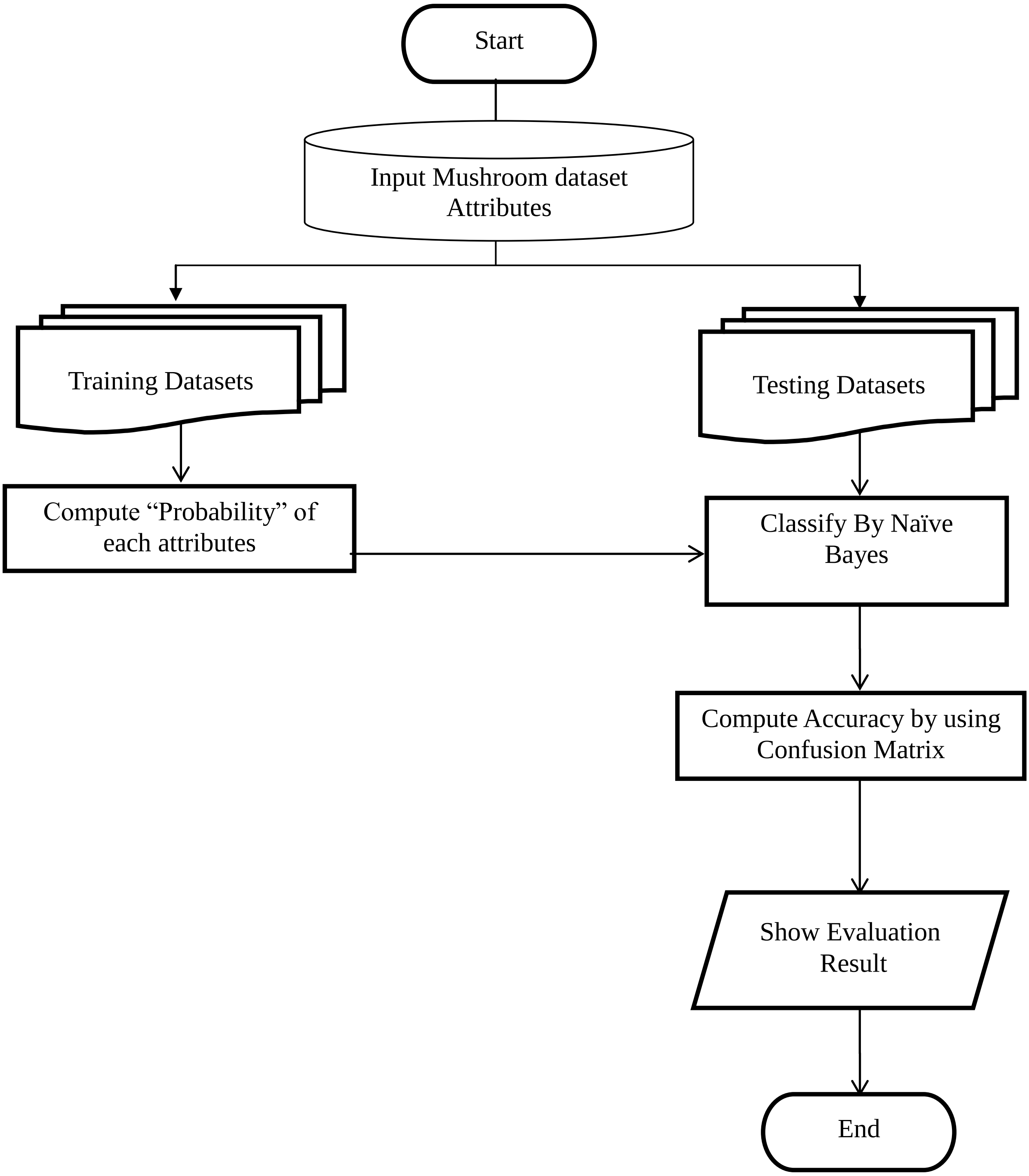
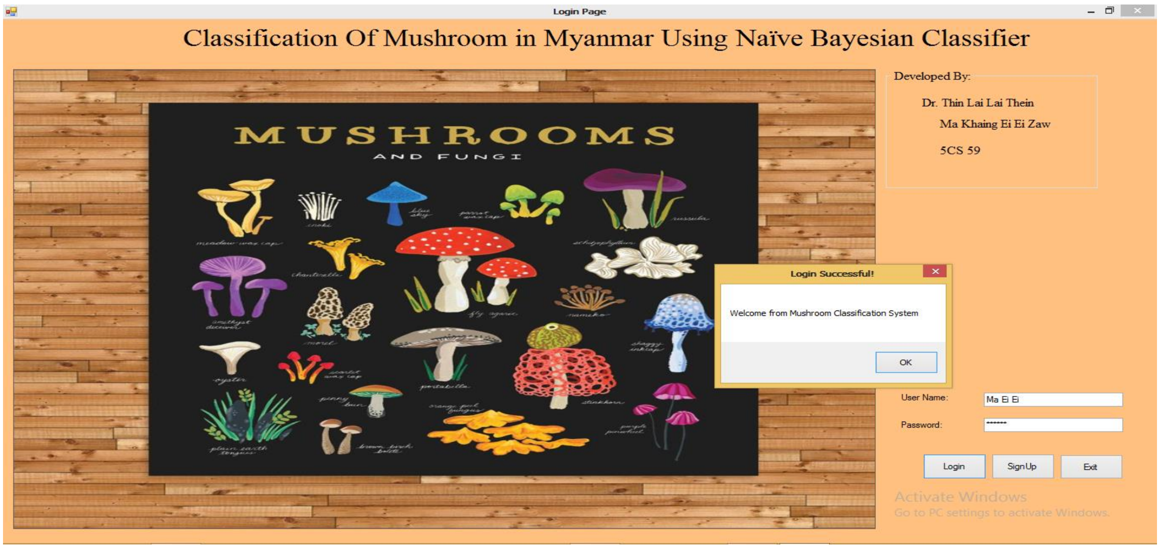

Start
Input Mushroom dataset Attributes
Training Datasets
Testing Datasets
Compute “Probability” of each attributes
Classify By Naïve Bayes
Compute Accuracy by using Confusion Matrix
Show Evaluation Result
End

Figure 4.2: System Flow Diagram

4.5. Implementation of the System



Classification Of Mushroom in Myanmar Using Naïve Bayesian Classifier

Developed By:
Dr. Thin Lai Lai Thein
Ma Khaing Ei Ei Zaw
5CS 59

Login Successful!

Welcome from Mushroom Classification System

User Name: Ma Ei Ei
Password: *******

Activate Windows
Go to PC settings to activate Windows.

Figure 4.3: Login Page of System

The proposed mushroom classification system can only be used for registered user. So, the user who want to use this system must be registered (Sign Up) first and can be enter by registered user information. After the authentication is successful, the user can get the main page of the system as shown in figure 4.4.

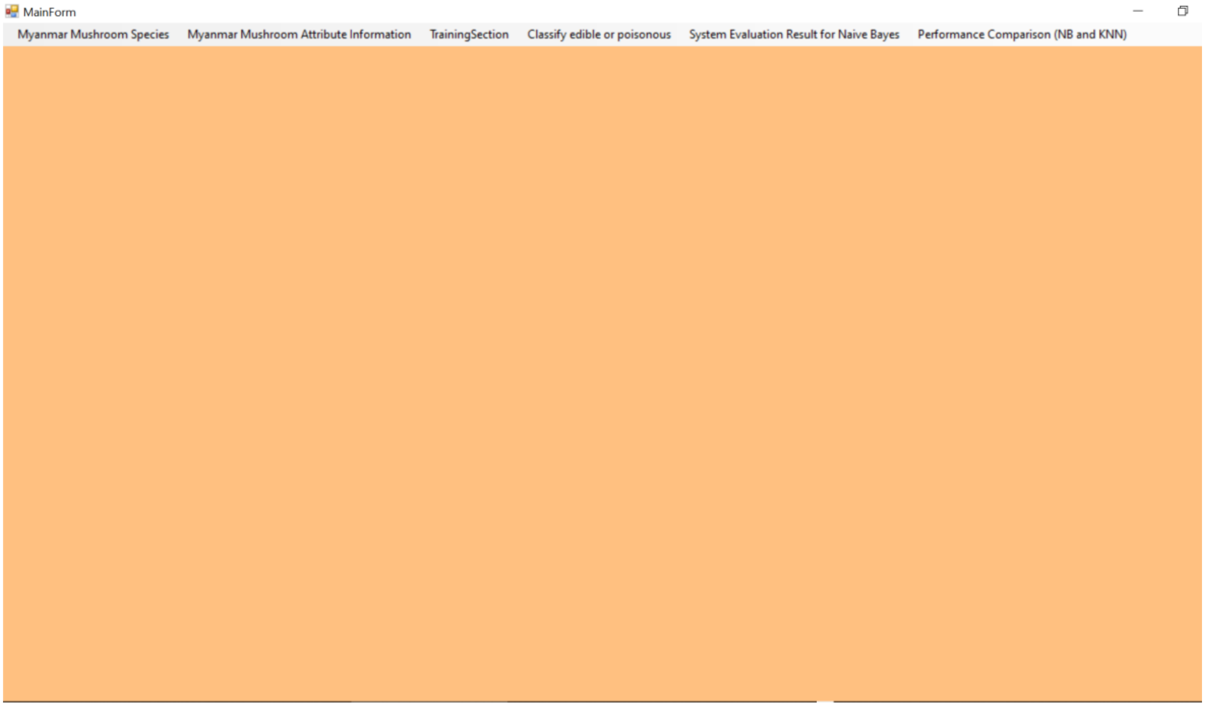
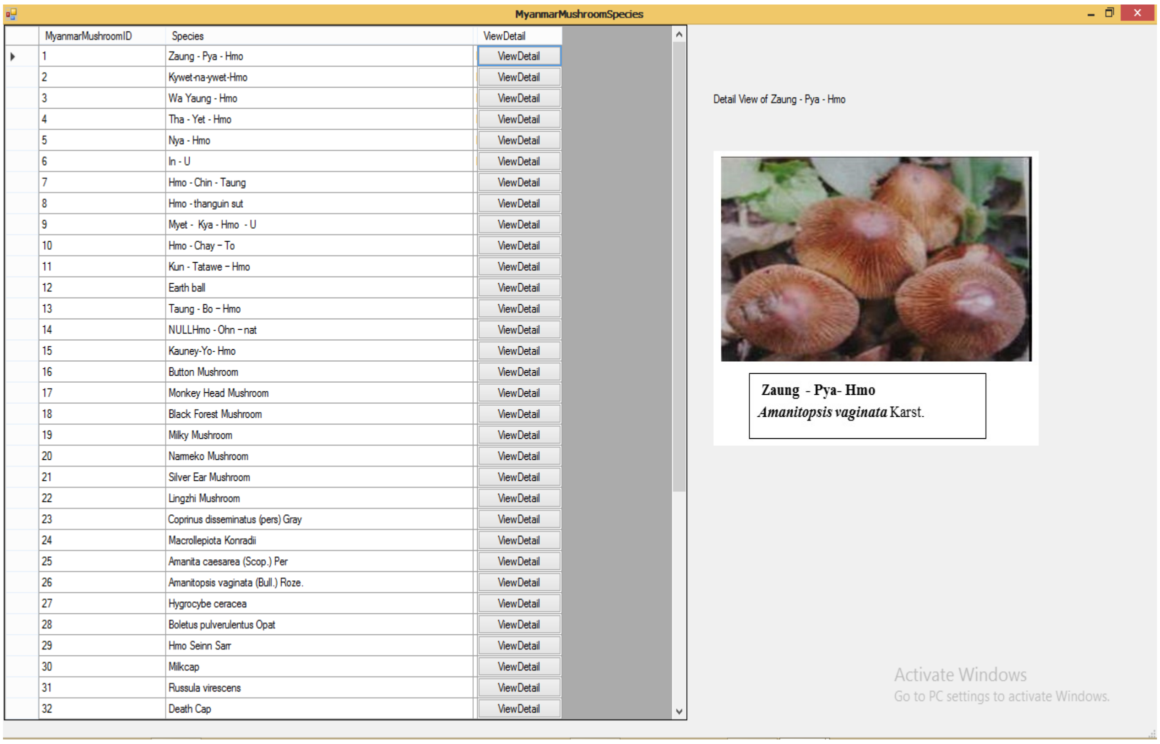

MainForm
Myanmar Mushroom Species
Myanmar Mushroom Attribute Information
TrainingSection
Classify edible or poisonous
System Evaluation Result for Naive Bayes
Performance Comparison (NB and KNN)

Figure 4.4: Main Page of System

The main page of system has six main menus: they are Myanmar mushroom species, Myanmar mushroom attribute information, Training Section, Classify edible and poisonous, System Evaluation Result for Naive bayes and Performance Comparison (NB and KNN).

Firstly, “Myanmar Mushroom Species menu” which will tabular presentation page of various Myanmar mushroom species with photo description for visual support as shown in figure 4.5.



MyanmarMushroomSpecies

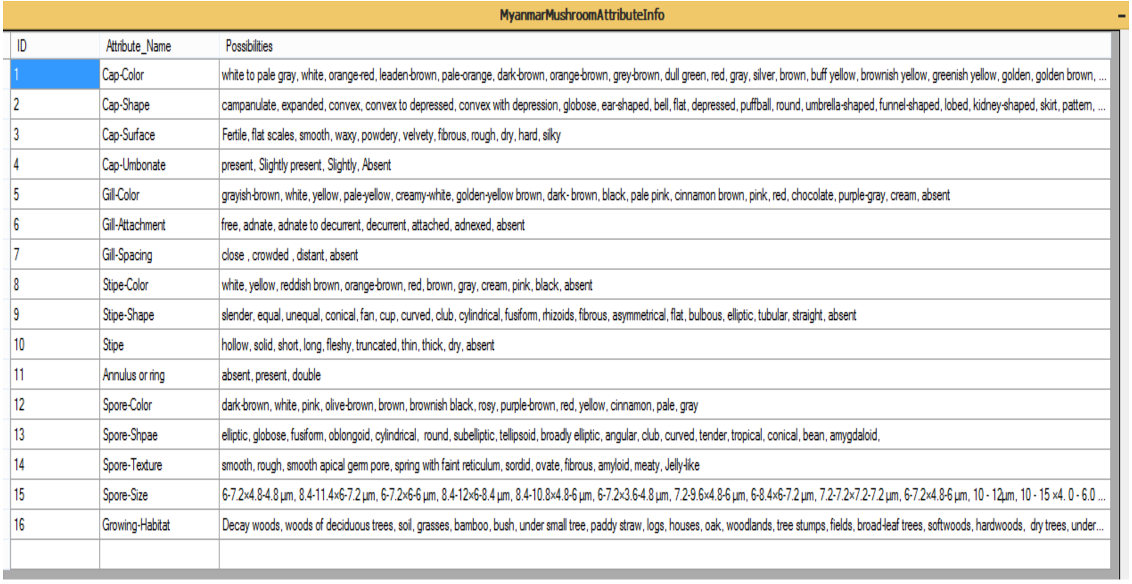
MyanmarMushroomID	Species	ViewDetail
1	Zaung - Pya - Hmo	ViewDetail
2	Kywet-na-ywet-Hmo	ViewDetail
3	Wa Yaung - Hmo	ViewDetail
4	Tha - Yet - Hmo	ViewDetail
5	Nya - Hmo	ViewDetail
6	In - U	ViewDetail
7	Hmo - Chin - Taung	ViewDetail
8	Hmo - thanguin sut	ViewDetail
9	Myet - Kya - Hmo - U	ViewDetail
10	Hmo - Chay – To	ViewDetail
11	Kun - Tatawe – Hmo	ViewDetail
12	Earth ball	ViewDetail
13	Taung - Bo – Hmo	ViewDetail
14	NULLHmo - Ohn – nat	ViewDetail
15	Kauney-Yo- Hmo	ViewDetail
16	Button Mushroom	ViewDetail
17	Monkey Head Mushroom	ViewDetail
18	Black Forest Mushroom	ViewDetail
19	Milky Mushroom	ViewDetail
20	Nameko Mushroom	ViewDetail
21	Silver Ear Mushroom	ViewDetail
22	Lingzhi Mushroom	ViewDetail
23	Coprinus disseminatus (pers) Gray	ViewDetail
24	Macrollepiota Konradii	ViewDetail
25	Amanita caesarea (Scop.) Per	ViewDetail
26	Amanitopsis vaginata (Bull.) Roze.	ViewDetail
27	Hygrocybe ceracea	ViewDetail
28	Boletus pulverulentus Opat	ViewDetail
29	Hmo Seinn Sarr	ViewDetail
30	Milkcap	ViewDetail
31	Russula virescens	ViewDetail
32	Death Cap	ViewDetail

Detail View of Zaung - Pya - Hmo

Zaung - Pya- Hmo
Amanitopsis vaginata Karst.

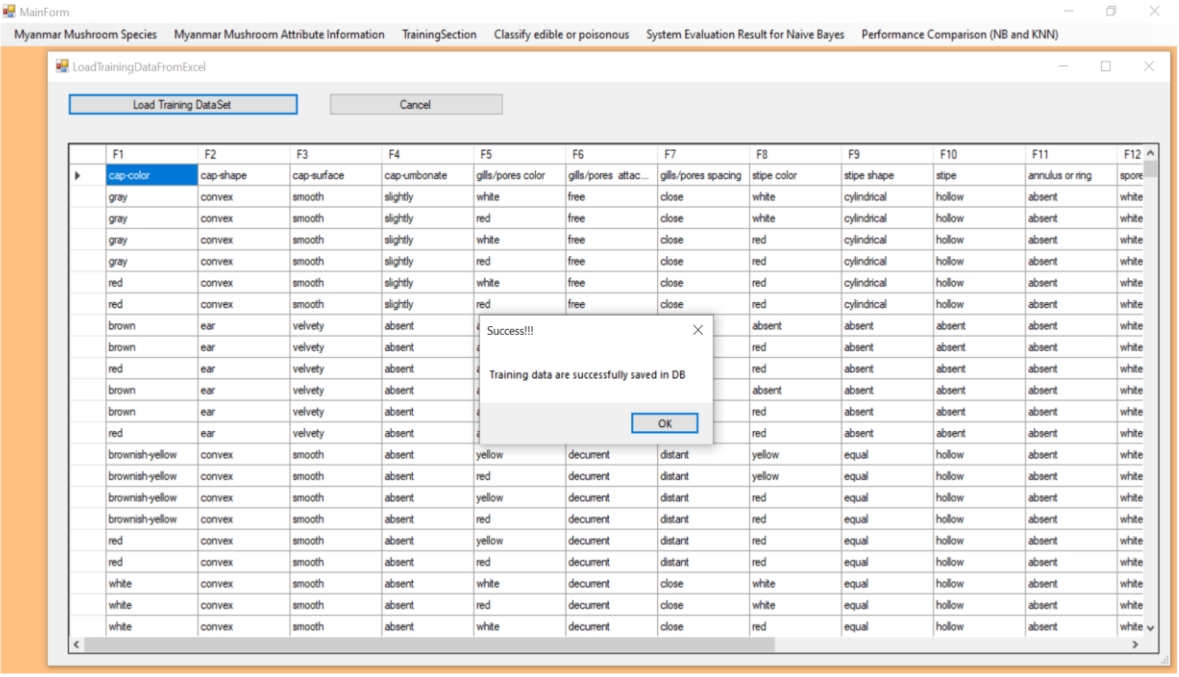
Figure 4.5: Mushroom Species List Page

Secondly, “Myanmar Mushroom Attributes Information” menu is to explain and describe various characteristic attributes of each edible and poisonous species mushroom as shown in figure 4.6. This is main supported knowledge for the classification of edible or poisonous mushroom of Myanmar.



ID	Attribute_Name	Possibilities
1	Cap-Color	white to pale gray, white, orange-red, leaden-brown, pale-orange, dark-brown, orange-brown, grey-brown, dull green, red, gray, silver, brown, buff yellow, brownish yellow, greenish yellow, golden, golden brown, ...
2	Cap-Shape	campanulate, expanded, convex, convex to depressed, convex with depression, globose, ear-shaped, bell, flat, depressed, puffball, round, umbrella-shaped, funnel-shaped, lobed, kidney-shaped, skirt, pattern, ...
3	Cap-Surface	Fertile, flat scales, smooth, waxy, powdery, velvety, fibrous, rough, dry, hard, silky
4	Cap-Umbonate	present, Slightly present, Slightly, Absent
5	Gill-Color	grayish-brown, white, yellow, pale-yellow, creamy-white, golden-yellow brown, dark- brown, black, pale pink, cinnamon brown, pink, red, chocolate, purple-gray, cream, absent
6	Gill-Attachment	free, adnate, adnate to decurrent, decurrent, attached, adnexed, absent
7	Gill-Spacing	close , crowded , distant, absent
8	Stipe-Color	white, yellow, reddish brown, orange-brown, red, brown, gray, cream, pink, black, absent
9	Stipe-Shape	slender, equal, unequal, conical, fan, cup, curved, club, cylindrical, fusiform, rhizoids, fibrous, asymmetrical, flat, bulbous, elliptic, tubular, straight, absent
10	Stipe	hollow, solid, short, long, fleshy, truncated, thin, thick, dry, absent
11	Annulus or ring	absent, present, double
12	Spore-Color	dark-brown, white, pink, olive-brown, brown, brownish black, rosy, purple-brown, red, yellow, cinnamon, pale, gray
13	Spore-Shpae	elliptic, globose, fusiform, oblongoid, cylindrical, round, subelliptic, tellipsoid, broadly elliptic, angular, club, curved, tender, tropical, conical, bean, amygdaloid,
14	Spore-Texture	smooth, rough, smooth apical gem pore, spring with faint reticulum, sordid, ovate, fibrous, amyloid, meaty, Jelly-like
15	Spore-Size	6-7.2×4.8-4.8 µm, 8.4-11.4×6-7.2 µm, 6-7.2×6-6 µm, 8.4-12×6-8.4 µm, 8.4-10.8×4.8-6 µm, 6-7.2×3.6-4.8 µm, 7.2-9.6×4.8-6 µm, 6-8.4×6-7.2 µm, 7.2-7.2×7.2-7.2 µm, 6-7.2×4.8-6 µm, 10 - 12µm, 10 - 15 ×4. 0 - 6.0 ...
16	Growing-Habitat	Decay woods, woods of deciduous trees, soil, grasses, bamboo, bush, under small tree, paddy straw, logs, houses, oak, woodlands, tree stumps, fields, broad-leaf trees, softwoods, hardwoods, dry trees, under...

Figure 4.6: Myanmar Mushroom Attributes Information



MainForm
Myanmar Mushroom Species
Myanmar Mushroom Attribute Information
TrainingSection
Classify edible or poisonous
System Evaluation Result for Naive Bayes
Performance Comparison (NB and KNN)
LoadTrainingDataFromExcel
Load Training DataSet
Cancel
OK

Success!!!

Training data are successfully saved in DB

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
cap-color	cap-shape	cap-surface	cap-umbonate	gills/pores color	gills/pores attac...	gills/pores spacing	stipe color	stipe shape	stipe	annulus or ring	spore
gray	convex	smooth	slightly	white	free	close	white	cylindrical	hollow	absent	white
gray	convex	smooth	slightly	red	free	close	white	cylindrical	hollow	absent	white
gray	convex	smooth	slightly	white	free	close	red	cylindrical	hollow	absent	white
gray	convex	smooth	slightly	red	free	close	red	cylindrical	hollow	absent	white
red	convex	smooth	slightly	white	free	close	red	cylindrical	hollow	absent	white
red	convex	smooth	slightly	red	free	close	red	cylindrical	hollow	absent	white
brown	ear	velvety	absent				absent	absent	absent	absent	white
brown	ear	velvety	absent				red	absent	absent	absent	white
red	ear	velvety	absent				red	absent	absent	absent	white
brown	ear	velvety	absent				absent	absent	absent	absent	white
brown	ear	velvety	absent				red	absent	absent	absent	white
red	ear	velvety	absent				red	absent	absent	absent	white
brownish-yellow	convex	smooth	absent	yellow	decurrent	distant	yellow	equal	hollow	absent	white
brownish-yellow	convex	smooth	absent	red	decurrent	distant	yellow	equal	hollow	absent	white
brownish-yellow	convex	smooth	absent	yellow	decurrent	distant	red	equal	hollow	absent	white
brownish-yellow	convex	smooth	absent	red	decurrent	distant	red	equal	hollow	absent	white
red	convex	smooth	absent	yellow	decurrent	distant	red	equal	hollow	absent	white
red	convex	smooth	absent	red	decurrent	distant	red	equal	hollow	absent	white
white	convex	smooth	absent	white	decurrent	close	white	equal	hollow	absent	white
white	convex	smooth	absent	red	decurrent	close	white	equal	hollow	absent	white
white	convex	smooth	absent	white	decurrent	close	red	equal	hollow	absent	white

Figure 4.7: Training Data of Mushroom Datasets

This is first phase for the classification process. The user must be loaded the training data excel sheet from the system supported open dialog box and data are loaded to the system and then stored in the system database. After successfully upload the training dataset, the message will be shown in Figure 4.7.

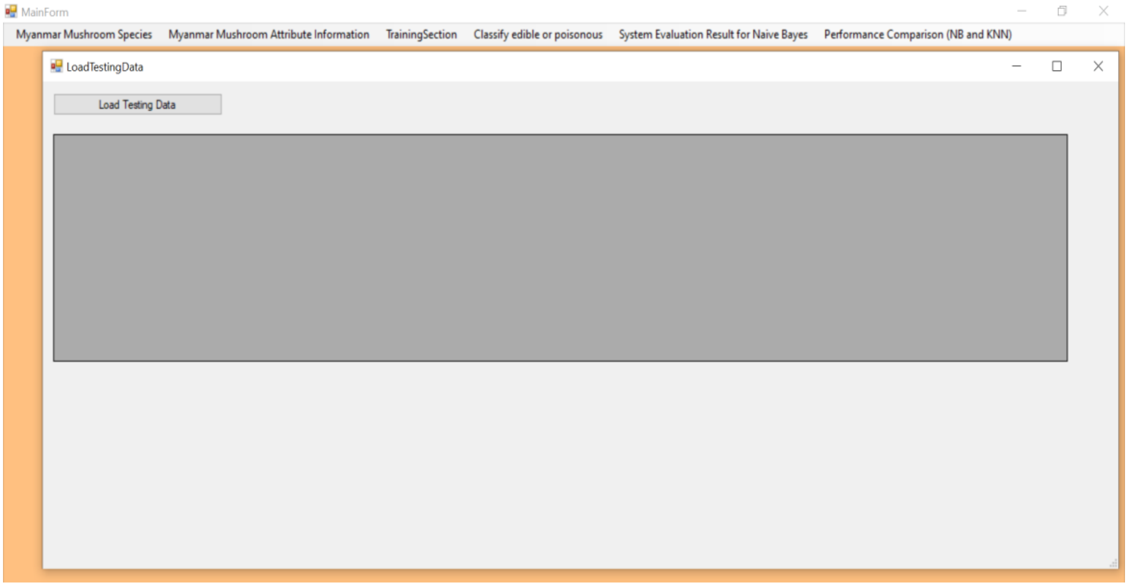
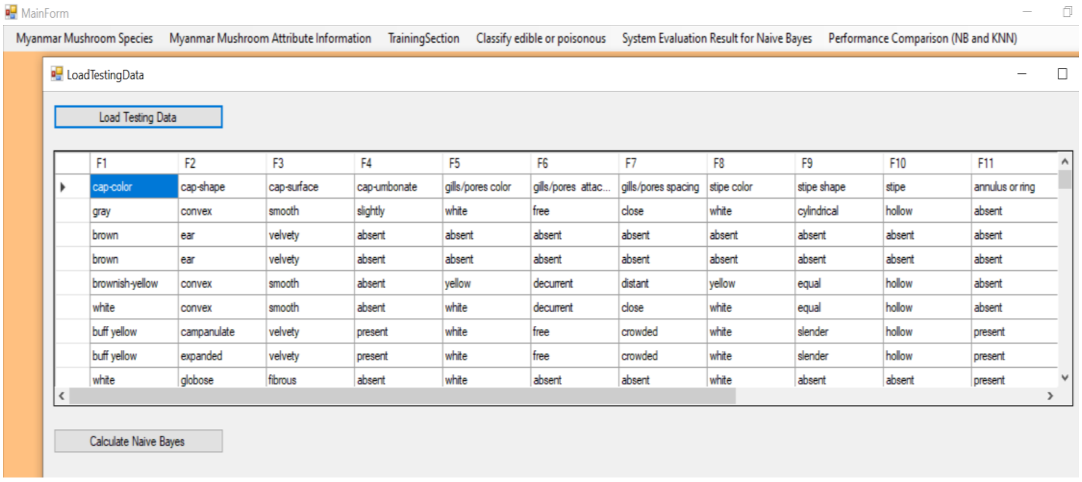

MainForm
Myanmar Mushroom Species
Myanmar Mushroom Attribute Information
TrainingSection
Classify edible or poisonous
System Evaluation Result for Naive Bayes
Performance Comparison (NB and KNN)
LoadTestingData
Load Testing Data

Figure 4.8: Load Testing Data

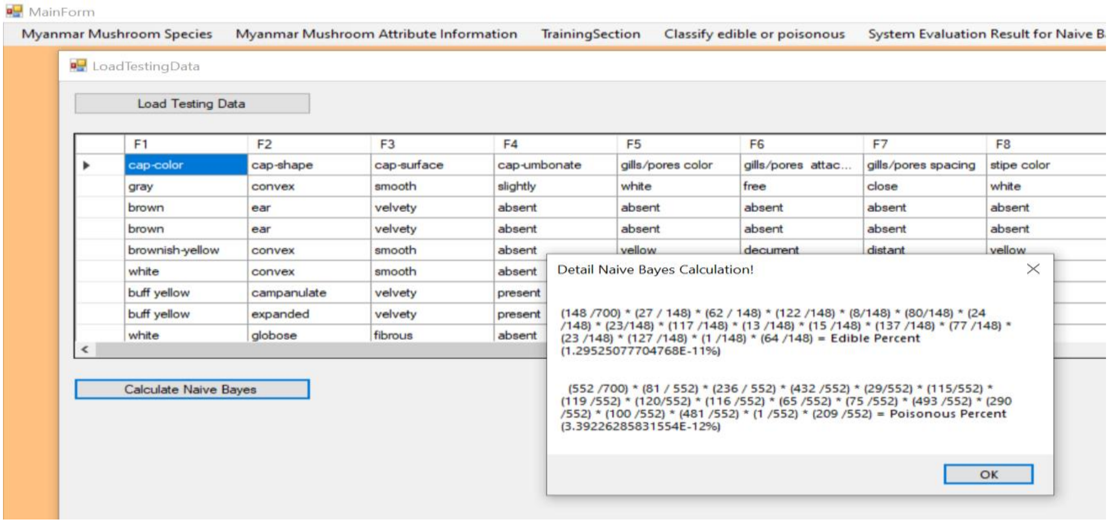
To classify the mushroom edible and poisonous, the user must be loaded testing data via the system support "Load Testing Data" button from the page shown in figure 4.8.


MainForm
Myanmar Mushroom Species
Myanmar Mushroom Attribute Information
TrainingSection
Classify edible or poisonous
System Evaluation Result for Naive Bayes
Performance Comparison (NB and KNN)
LoadTestingData
Load Testing Data
Calculate Naive Bayes

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
cap-color	cap-shape	cap-surface	cap-umbonate	gills/pores color	gills/pores attac...	gills/pores spacing	stipe color	stipe shape	stipe	annulus or ring
gray	convex	smooth	slightly	white	free	close	white	cylindrical	hollow	absent
brown	ear	velvety	absent	absent	absent	absent	absent	absent	absent	absent
brown	ear	velvety	absent	absent	absent	absent	absent	absent	absent	absent
brownish-yellow	convex	smooth	absent	yellow	decurrent	distant	yellow	equal	hollow	absent
white	convex	smooth	absent	white	decurrent	close	white	equal	hollow	absent
buff yellow	campanulate	velvety	present	white	free	crowded	white	slender	hollow	present
buff yellow	expanded	velvety	present	white	free	crowded	white	slender	hollow	present
white	globose	fibrous	absent	white	absent	absent	white	absent	absent	present

Figure 4.9: Testing Data of Mushroom Datasets

Then, "Calculate Naive Bayes" button is support to proceed the classification process by Naive Bayes Classifier. Classification observation of each testing records are as shown in figure 4.10.


MainForm
Myanmar Mushroom Species
Myanmar Mushroom Attribute Information
TrainingSection
Classify edible or poisonous
System Evaluation Result for Naive B
LoadTestingData
Load Testing Data
Calculate Naive Bayes
OK

	F1	F2	F3	F4	F5	F6	F7	F8
▶	cap-color	cap-shape	cap-surface	cap-umbonate	gills/pores color	gills/pores attac...	gills/pores spacing	stipe color
	gray	convex	smooth	slightly	white	free	close	white
	brown	ear	velvety	absent	absent	absent	absent	absent
	brown	ear	velvety	absent	absent	absent	absent	absent
	brownish-yellow	convex	smooth	absent	yellow	decurrent	distant	yellow
	white	convex	smooth	absent				
	buff yellow	campanulate	velvety	present				
	buff yellow	expanded	velvety	present				
	white	globose	fibrous	absent				

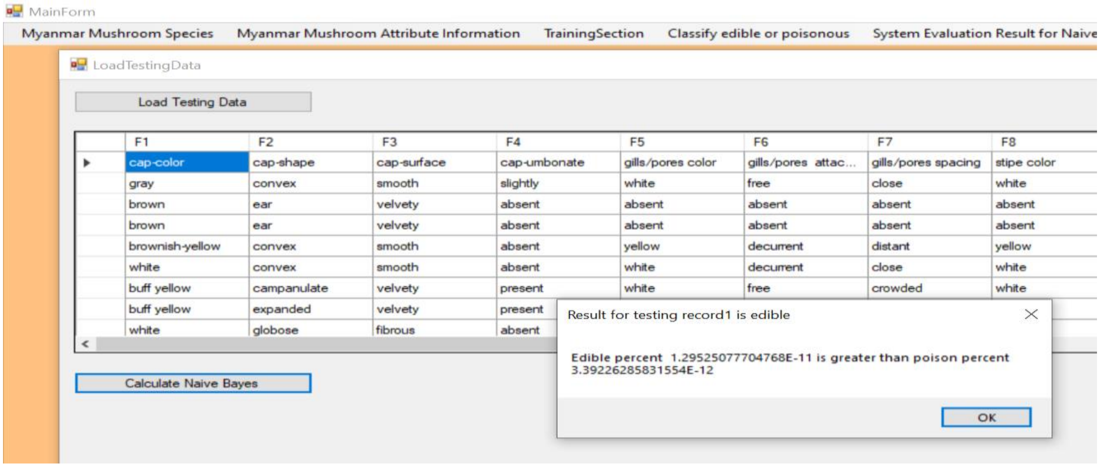
Detail Naive Bayes Calculation!

(148 /700) * (27 / 148) * (62 / 148) * (122 /148) * (8/148) * (80/148) * (24 /148) * (23/148) * (117 /148) * (13 /148) * (15 /148) * (137 /148) * (77 /148) * (23 /148) * (127 /148) * (1 /148) * (64 /148) = Edible Percent (1.29525077704768E-11%)

(552 /700) * (81 / 552) * (236 / 552) * (432 /552) * (29/552) * (115/552) * (119 /552) * (120/552) * (116 /552) * (65 /552) * (75 /552) * (493 /552) * (290 /552) * (100 /552) * (481 /552) * (1 /552) * (209 /552) = Poisonous Percent (3.39226285831554E-12%)

Figure 4.10: Classification Observation by Naive Bayes

And then, result for all testing datasets calculate by Naive Bayes as shown in figure 4.11.

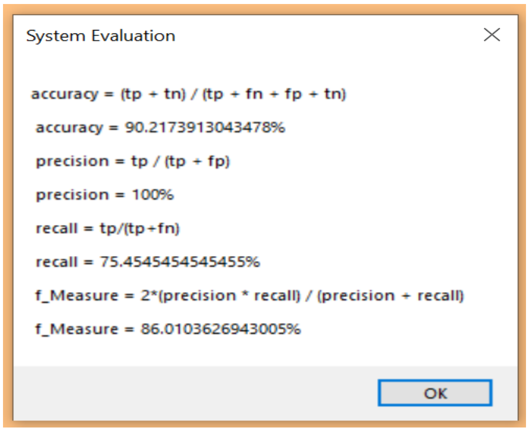


	F1	F2	F3	F4	F5	F6	F7	F8
▶	cap-color	cap-shape	cap-surface	cap-umbonate	gills/pores color	gills/pores attac...	gills/pores spacing	stipe color
	gray	convex	smooth	slightly	white	free	close	white
	brown	ear	velvety	absent	absent	absent	absent	absent
	brown	ear	velvety	absent	absent	absent	absent	absent
	brownish-yellow	convex	smooth	absent	yellow	decurrent	distant	yellow
	white	convex	smooth	absent	white	decurrent	close	white
	buff yellow	campanulate	velvety	present	white	free	crowded	white
	buff yellow	expanded	velvety	present				
	white	globose	fibrous	absent				

Result for testing record1 is edible

Edible percent 1.29525077704768E-11 is greater than poison percent 3.39226285831554E-12

Figure 4.11: System Result for Mushroom Datasets



accuracy = (tp + tn) / (tp + fn + fp + tn)

accuracy = 90.2173913043478%

precision = tp / (tp + fp)

precision = 100%

recall = tp/(tp+fn)

recall = 75.4545454545455%

f_Measure = 2*(precision * recall) / (precision + recall)

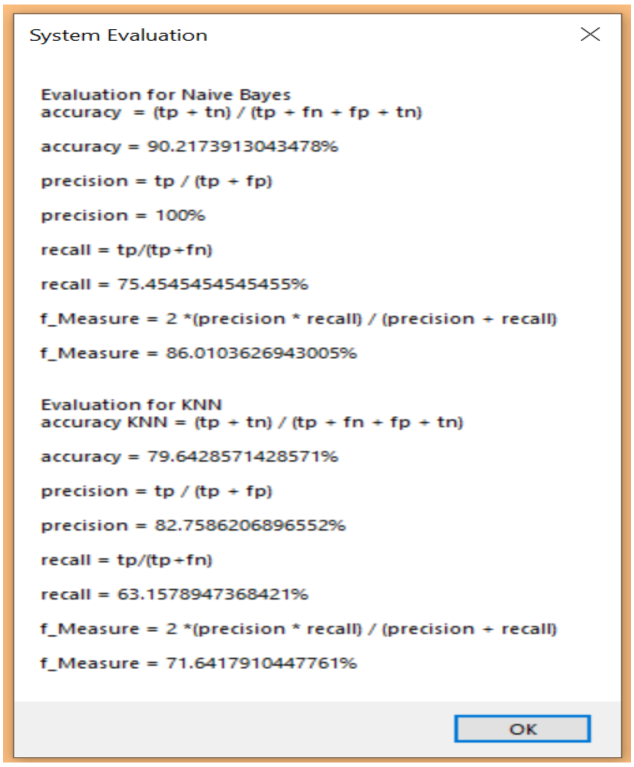
f_Measure = 86.0103626943005%

Figure 4.12: System Evaluation Result

The system evaluation result for Naive Bayesian Classifier, it can be seen that predicting accuracy, precision, recall and f-Measure as shown in figure 4.12.

Due to the comparative classification, the user can compare and can get more accurate classification results. The performance evaluation is determined by using confusion matrix: accuracy calculation, precision calculation, recall calculation and F-Measure calculation.

The system evaluation result for Naïve Bayesian and K- Nearest Neighbors, it can be seen that predicting accuracy, precision, recall and f-Measure as shown in figure 4.13.



```
Evaluation for Naive Bayes
accuracy  = (tp + tn) / (tp + fn + fp + tn)

accuracy = 90.2173913043478%

precision = tp / (tp + fp)

precision = 100%

recall = tp/(tp+fn)

recall = 75.4545454545455%

f_Measure = 2 *(precision * recall) / (precision + recall)

f_Measure = 86.0103626943005%

Evaluation for KNN
accuracy KNN = (tp + tn) / (tp + fn + fp + tn)

accuracy = 79.6428571428571%

precision = tp / (tp + fp)

precision = 82.7586206896552%

recall = tp/(tp+fn)

recall = 63.1578947368421%

f_Measure = 2 *(precision * recall) / (precision + recall)

f_Measure = 71.6417910447761%
```

Figure 4.13: Performance Comparison of Naive Bayesian and K-Nearest Neighbors

Some poisonous mushrooms can kill, so the user must be able to accurately name the fungus and be 100% sure of what it is before consumption. There are some apparent rules for picking safe mushrooms but these are just fanciful if not downright dangerous;

- 'It's ok if the user can peel the cap.' It is easy to peel a Death Cap.
- 'Mushrooms growing on wood are safe.' No not all of them are and some are deadly, like the Funeral Bell.

- ‘If you see other animals eating them they are ok.’ This rule is not true, many animals can eat poisonous fungi with no ill effects.

Some good rules apply for avoiding poisonous mushrooms if there are novices;

1. Avoid mushrooms with white gills, a skirt or ring on the stem and a bulbous or sack like base called a volva. The user can be missing out on some good edible fungi but it means the user will be avoiding the deadly members of the Amanita family.
2. Avoid mushrooms with red on the cap or stem. Again you will be missing out on some good mushrooms but more importantly the user can be picking poisonous ones.
3. Finally don’t consume any mushrooms unless the users are 100% sure of what they are. The user know it has already mentioned this but it is by far the most important rule.

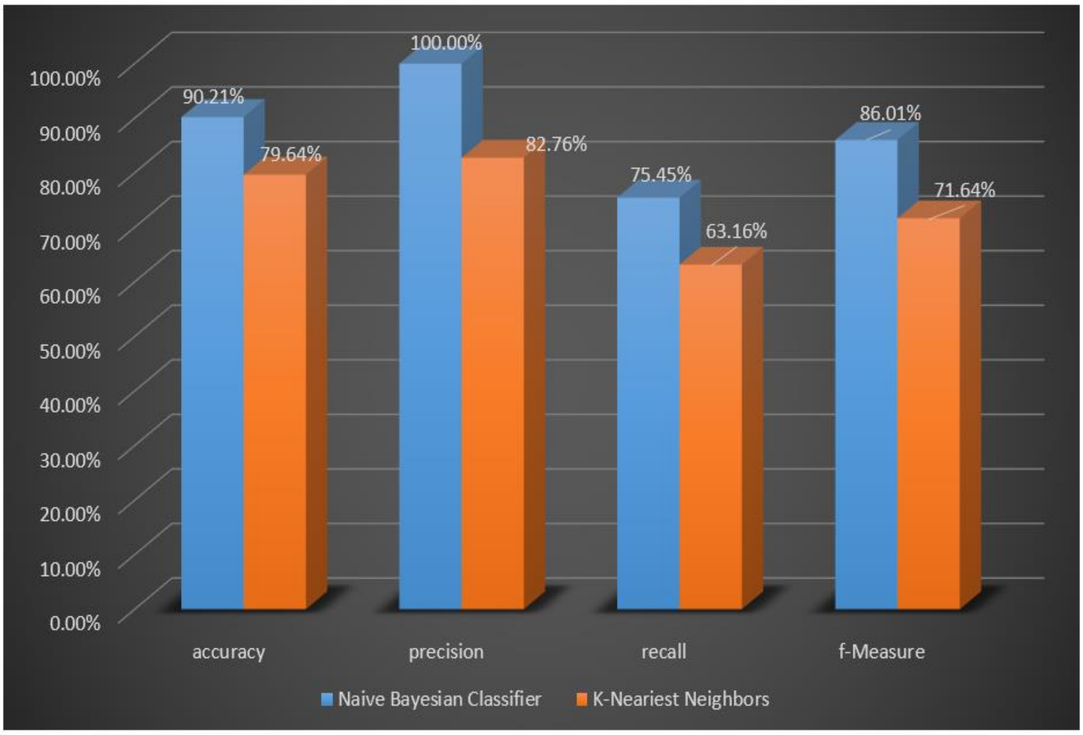
These rules do not mean all other mushrooms are safe but help rule out some of the nastier types.

This system aims to classify the various mushroom based on their various shape by using Naive Bayes and KNN Classifiers. The training data are collected according to the characteristic, shape and species as described in section 4.1 and 4.2. Then, the testing data are classified by two different classifier and each observations are compare by calculation accuracy, precision, recall and f-measure to get more reliable and good performance in classification. Each classification of two different methods are compared.

	Naive Bayesian Classifier	**K-Nearest Neighbors**
accuracy	**90.21%**	**79.64%**
precision	**100%**	**82.76%**
Recall	**75.45%**	**63.16%**
f-Measure	**86.01%**	**71.64%**

Table 4.3: Performance Comparison of Naive Bayesian and K-Nearest Neighbors

From the experimental results obtained, it can be seen that Naive Bayes gave the highest test accuracy better than K-Nearest Neighbors as shown in Table 4.3.



	accuracy	precision	recall	F-Measure
Naive Bayesian Classifier	90.21%	100%	75.45%	86.01%
K-Neariest Neighbors	79.64%	82.76%	63.16%	71.64%

Figure 4.14: Comparison Chart

CHAPTER 5
CONCLUSION AND FURTHER EXTENSIONS

5.1 Conclusion

Characterization is the method involved with finding a bunch of models that depict and recognize information classes or ideas. Bayesian grouping depends on Bayes hypothesis. The Naive Bayesian grouping use is centered around the Bayesian recipe to ascertain the likelihood of each class given the upsides, everything being equal. This framework can be applied to foresee the class mark of obscure example given the example information. The general presentation of the Naive Bayesian arrangement can act as a gauge of the restrictive freedom of properties. This framework has introduced creating of characterization from huge informational indexes. The two methodologies show in managing mushroom information for characterization which thinks about the arrangement issue of mushroom by utilizing Naive Bayesian order and KNN. Guileless Bayes is a direct classifier while KNN isn't; It will in general be quicker when applied to enormous information. In examination, KNN is typically more slow for a lot of information, as a result of the computations expected for each new advance simultaneously. Innocent Bayes give high precision when enormous measure of information.

In view of the google web search tool, it is figured out two examination distributions in the field of palatable and noxious mushroom recognizable proof and arrangement. The Naive Bayes classifier uses the naive Bayes formula to compute the probability of each class given the prices of all attributes. The Bayes Classifier is based on Bayes theorem of posterior probability. The system support users in classifying edible and poisonous mushroom based on the mushroom attributes. The classification based on mushroom datasets by using Naive Bayesian Classifier. For the performance comparison of accuracy, the two algorithms are used Naive Bayesian classifiers and K-Nearest neighbor (KNN) by using confusion matrix. The mushroom identification method that has been done used Naive Bayes and KNN algorithms with the prediction accuracy of 90.21% and 79.64%. Gullible Bayes is one of arrangement calculation and order is one of the significant examinations in information mining.

5.2. Benefits of the System

All species of mushrooms are not palatable. So prior to consuming, it can be checked for edibility. In any case, mushrooms additionally have high mycotoxin, which decide if the mushrooms are consumable or toxic. By and large, individuals do not realize that there is a major distinction between palatable mushrooms and noxious mushrooms. Each mushroom has its own unique attributes. These qualities are distinguished as elements that can be utilized to order the mushrooms into two classes; which are consumable and harmful. Exact assurance and legitimate ID of species are the main safe method for guaranteeing edibility, and defend against conceivable mishap of consuming toxic one. The proposed framework can characterize palatable and toxic mushrooms in light of specific ascribes like shape, size, variety, etc.

5.3 Further Extensions

Many mushrooms species are basically the same as one another. A parallel order cannot be dependable. The essential information can be utilized to recreate other randomized forms of the information with an inconsistent number of speculative mushrooms per species. To be sure, since the essential information likewise included the two multinomial classes name and family, mimicking new varieties of optional information for multinomial classification is additionally conceivable. This really intends that rather than just distinguishing a mushroom as noxious or consumable, there work can be stretched out to recognize a specific family or certain species.

Besides, multivariate grouping is additionally conceivable by recreating optional information (with two or every one of the three of the classes all the while).

AUTHOR'S PUBLICATION

[1] Khaing Ei Ei Zaw, Thein Lai Lai Thein, "Classification Of mushroom in Myanmar using Naive Bayesian Classifier", The National Journal of Parallel and Soft Computing (NJPSC 2022),2022.

REFERENCES

[1] Agung Wibowo, Yuri Rahayu, Andi Riyanto, Taufik Hidayatulloh; “Classification Algorithm for Edible Mushroom Identification”, International Conference on Information and Communications Technology (ICOIACT), March 2018.

[2] Balika J. Chelliah, S. Kalaiarasi, Apoorva Anand, Janakiram G, Bhaghi Rathi, Nakul K. Warrier; “Classification of Mushrooms using Supervised Learning Models”, Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India, April 2018.

[3] Bandana Garg, “DESIGN AND DEVELOPMENT OF NAÏVE BAYEs CLASSIFIER”, North Dakota State University of Agriculture and Applied Science, June 2013.

[4] Eyad Sameh Alkronz, Khaled A. Moghayer, Mohamad Meimeh, Mohannad Gazzaz; “Classification of Mushroom Using Artificial Neural Network”, Department of Information Technology, Faculty of Engineering and Information Technology, Al-Azhar University - Gaza, Palestine, February 2019.

[5] Kanchi Tank; “A Comparative Study on Mushroom Classification using Supervised Machine Learning Algorithms”, International Journal of Trend in Scientific Research and Development (IJTSRD) Volume 5 Issue 5, July-August 2021.

[6] Kasarapu Ramani, “ MACHINE LEARNING TOOLS FOR DATASET CLASSIFICATION ”, Department of IT, Sree VidyanikethanEngg College (Autonomous), Tirupati, India, January 2018.

[7] Mohammad Ashraf Ottom1, Noor Aldeen Alawad2, Khalid M. O. Nahar2; “Classification of Mushroom Fungi Using Machine Learning Techniques”, International Journal of Advanced Trends in Computer Science and Engineering, October 2019.

[8] Muhammad Husaini, “A Data Mining Based On Ensemble Classifier Classification Approach for Edible Mushroom Identification”, University of Science Malaysia, Pulau Pinang, Malaysia, July 2018.

[9] Roshna Chettri , Shrijana Pradhan and Lekhika Chettri Internet of Things: Comparative Study on Classification Algorithms (k-NN, Naive Bayes and Case based Reasoning), Sikkim Manipal University, November 2015.

[10] Ramazan Sener, “DETERMINATION OF POISON MUSHROOM USING NAIVE BAYES ALGORITHM”, Firat University, February 2020.

[11] Rakesh Kumar Y1 | Dr. V. Chandrasekhar2, “Machine Learning Methods to Classify Mushrooms for Edibility”, International Journal for Modern Trends in Science and Technology, 6(9): 54-58, 2020.

[12] Septian AriePrayoga 1, Ismasari Nawangsih 2, Tri NgudiWiyatno 3, “IMPLEMENTASI METODE NAÏVE BAYES CLASSIFIER”, Pelita Teknologi: Jurnal Ilmiah Informatika, Arsitektur dan Lingkungan 14 (2) 2019.

[13] Yuhan Zhang, Neural network classification on mushroom dataset with feature selection using evolutionary algorithm and auto-associative network, The ABCs 2018 - 1st ANU Bio-inspired Computing conference in Canberra, Australia, 20th July 2018.