ISSN 2278-3091

Volume 9, No.1.3, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse7191.32020.pdf

https://doi.org/10.30534/ijatcse/2020/7191.32020

Analysis of Performance of Classification Algorithms in Mushroom Poisonous Detection using Confusion Matrix Analysis

John Heland Jasper C. Ortega[1], Ace C. Lagman[2], Lizel Rose Q. Natividad[3],
Emilsa T. Bantug[4], Michael R. Resureccion[5], LanzJimuel O. Manalo[6]

[1]FEU Institute of Technology, Philippines, jcortega@feutech.edu.ph
[2]FEU Institute of Technology, Philippines, aclagman@ue.edu.ph
[3]San Beda University, Philippines, lnatividad@sanbeda.edu.ph
[4]Nueva Ecija University of Science and Technology, Philippines, emilsa.bantug@neust.edu.ph
[5]Univesity of the East, Philippines, resurreccion.michael@ue.edu.ph
[6]FEU Institute of Technology, Philippines, lomanalo@fit.edu.ph

ABSTRACT

Mushroom possesses potential benefits in terms of antioxidants, which are essential to the body. There are various health benefits one can get on consuming mushrooms, but not all types are said to be editable. There are types of mushrooms that are poisonous. The research aims to classify which among the mushroom are edible and poisonous based on given attributes such as odor, shape, texture, and color. The data set is originally donated from the Machine Learning Repository Department of the University of California in Irvine. This research focuses on analyzing the performance of the classification algorithms in mushroom poisonous detection. The researchers utilized the Knowledge Discovery in Databases processes in performing pattern extraction and performance analysis of algorithms. The decision tree method that is based on the entropy and information gain calculations has the highest computed accuracy result compare to other classification algorithms.

Key words: Classification algorithm, confusion matrix, data mining, mushroom

1. BACKGROUND OF THE STUDY

A mushroom is a form of plant-like that is known as fungus. Fungi bear large fruit bodies to be considered as a mushroom. Mushrooms can't produce their food in the presence of sunlight because of insufficient chlorophyll as compared to other plants. Mushrooms differ from their color, forms, shapes, and sizes. They occur in a wide variety of habitats on rotting logs of the woodland ranging from cold polar to hot extreme tropics down to the above-ground to below ground of the forest.

Mushrooms are special due to their edibility countries treat and consumed mushrooms as a kind of high nutritional value on food. The mushrooms that aren’t toxic happen to be utterly healthy and delicious as well. For the previous years, mushrooms have been used for their ability to add a distinctive taste to lots of different cuisines [1].

Although mushrooms are fungi, mushrooms belong to the vegetable category for cooking purposes that allows adding extra flavor without sodium and fat. However, only a few of them are edible. Distinguishing edible and toxic mushroom species need to be extra cautious to feature as there is no single characteristic by these dangerous mushrooms can be recognized, and not even one by which all consumable mushrooms can be recognized [2].

Several species of mushrooms are considered poisonous. Poisonous mushrooms can be hard to identify in the wild; despite that, some bear a resemblance to edible mushrooms, consuming them could mean danger [3].Addressing this problem is crucial as it can lead to harmful effects from ingestion of toxic substances present in mushrooms in which can eventually leads to gastro intestinal problems.

Consuming mushrooms that are seen in the wild is risky and ought to only be take on by people who are well-informed in recognizing mushrooms. The finest practice for the wild mushroom foragers is to focus on combining a small number of visually distinct edible mushroom species that cannot be easily confused with poisonous ones [4].

The researchers aimed to provide a way by analyzing the dataset and extract useful pattern to determine whether a mushroom can be poisonous or not based on physical and certain attributes.

1.1 Research Questions

a. How to evaluate the performance of classification algorithms in mushroom poisonous detection?

b. How to evaluate the developed system using a modified questionnaire based form ISO 9126 metrics?

2. LITERATURE REVIEW

Data Mining focuses on development of knowledge form datasets using machine learning techniques. It is also the application of a specific algorithm to extract patterns from data and transform it into information which can be used in different domains. [5].

Ottomet al. in [6] explored the use of data mining techniques in predicting whether the mushrooms are poisonous or not. In their study, the researchers have employed the use of neural network (NN), Support Vector Machines (SVM), Decision Tree, and l Nearest Neighbor (kNN).

Knowledge Discovery in Databases (KDD) refers as a structured way to develop predictive and cluster models that can profile and predict future instances. The main steps of KDD involve data preprocessing, modeling and evaluation. [7].

Decision tree algorithm has been applied in multiple disciplines, which include medicines, education, and business. It provides a tree based structure that presents nodes and stems connected in the root node. The root node serves as the highest information gain meaning it has the significant relationship towards the target variable The extraction of rule sets can be converted into decision rules. [8].

Naïve Bayes is a probabilistic analysis, which follow the Bayes theorem which determines the probability of an event to occur given sets of evidences[9].

Neural Network is a powerful algorithm that can train the system to produce the desired output. It is also called the learning algorithm where it has the ability to update weights and improve and optimize its prediction accuracy using feed forward and back propagation techniques [10].

3. METHODOLOGY

In order to solve the research questions, the researchers used Knowledge Discovery in Databases to extract hidden patterns form mushroom dataset and developed modified instrument used to evaluate system in terms of ISO 9126 metrics.

3.1 Knowledge Detection in Databases

The researchers used the adapted steps of Knowledge Discovery in Databases indicated in the figure below.

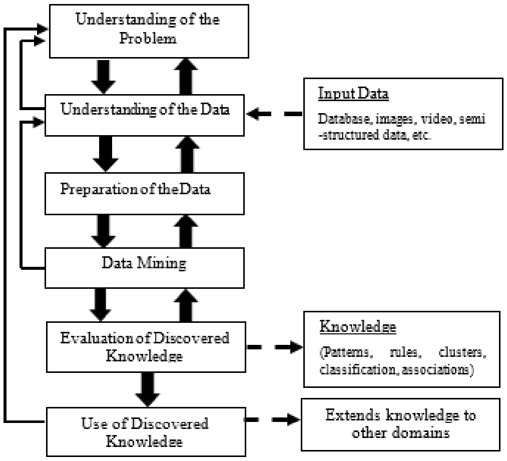

Understanding of the Problem
Understanding of the Data
Input Data
Database, images, video, semi -structured data, etc.
Preparation of the Data
Data Mining
Evaluation of Discovered Knowledge
Knowledge
(Patterns, rules, clusters, classification, associations)
Use of Discovered Knowledge
Extends knowledge to other domains

Figure 1:Six (6) KDD model.

The adapted version of the KDD consists include six (6) unique steps. The KDD is a data mining methodology which presents step by step process in extraction of data models used for prediction and cluster analysis.

3.2 Problem Understanding

This specific section involves the researchers to provide a deeper understanding the problem and its constraints. This leads to potential and probable solutions that may be created to solve the problem.

3.3 Data Understanding

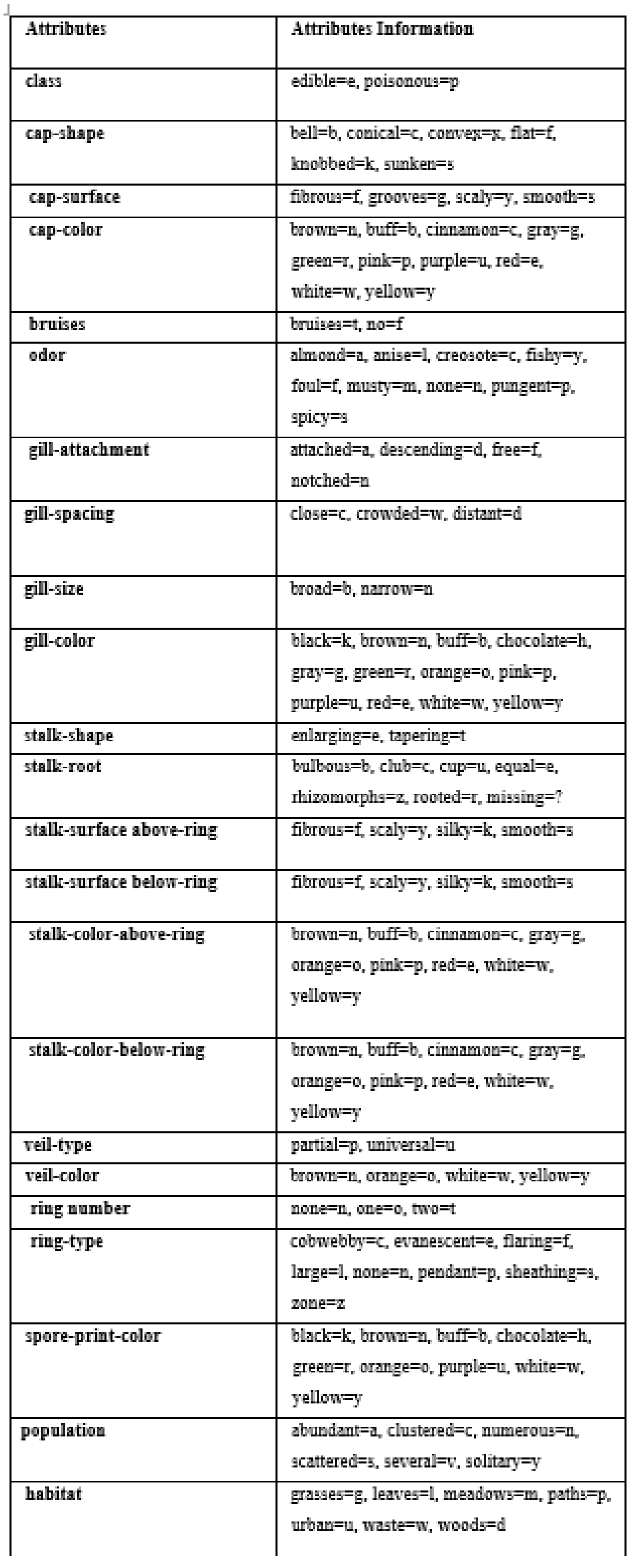
The researchers used the data from the repository of the UCI machine learning. This section determines the rational of the research and potential of the data to achieve researchers' goals. The data is lifted from Gary Lincoff, a mycologist whose enthusiasm to mushrooms led him to become the author, editor, and teacher at New York Botanical Garden [11].

3.4 Data Preparation

Data preparation provides a cleanup stage in data mining process. It solves data quality issues and problems by providing several data preparation techniques which include data imputation, data dimension reduction and feature selection.

The data set composes of 8124 number of rows of data records and total of 22 attributes. Each mushroom species is identified as class of edible and poisonous. Table 1 summarizes the attribute which are used for classifying mushrooms.

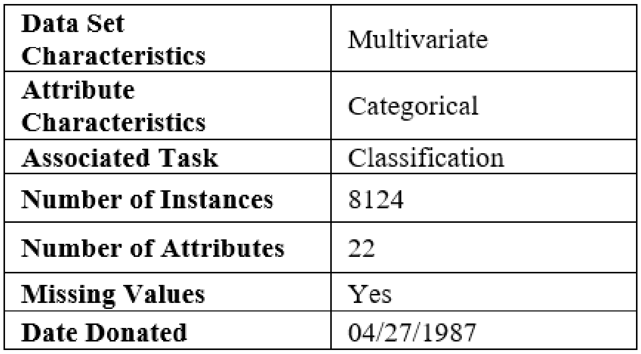
Table 1: Attribute Information



Attributes	Attributes Information
class	edible=e, poisonous=p
cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
cap-color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
bruises	bruises=t, no=f
odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
gill-attachment	attached=a, descending=d, free=f, notched=n
gill-spacing	close=c, crowded=w, distant=d
gill-size	broad=b, narrow=n
gill-color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
stalk-shape	enlarging=e, tapering=t
stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
stalk-surface above-ring	fibrous=f, scaly=y, silky=k, smooth=s
stalk-surface below-ring	fibrous=f, scaly=y, silky=k, smooth=s
stalk-color-above-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
stalk-color-below-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
veil-type	partial=p, universal=u
veil-color	brown=n, orange=o, white=w, yellow=y
ring number	none=n, one=o, two=t
ring-type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

The mushroom data set includes the conclusion on the result, which corresponds to whether the species is under the category of edible or poisonous. The analysis shows that the more the number of records and attributes better is the result. But the variation is too less. The motive is to find out whether these given attributes are necessary or not.

Table 2: Data Set Information



Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical
Associated Task	Classification
Number of Instances	8124
Number of Attributes	22
Missing Values	Yes
Date Donated	04/27/1987

There are missing values on the mushroom classification data set, the proponents use the symbol '?" on classifying the missing data on the given set of attributes.

To improve the quality of the dataset, the researchers used imputation technique. Data imputation is the representation of missing values in a data set. Missing values can increase the chances of producing errors and inaccuracy that limit the reliability of confidence importance.

3.5 Modeling

The KDD procedure's main focus is modeling. This process refers to the extraction of data models, which can be for predicting and clustering. The researchers utilized classification algorithms since the target variable is defined. Classification algorithms are classified as supervised learning.

3.6 Neural Network

The algorithm of the neural network is effective in utmost cases of sorting problems because of its learning algorithm. Figure 2 below indicates the basic neural network algorithm structure. It consists of hidden, input and output layers.

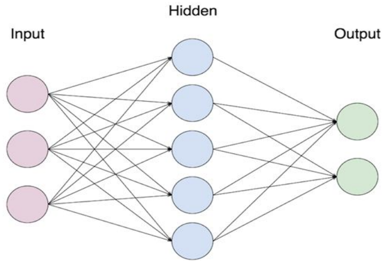


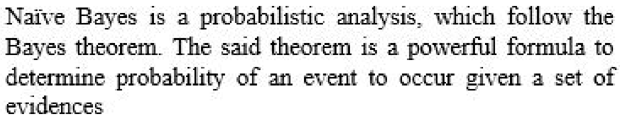
Figure 2: Neural Network Basic Architecture

The input layer presents all the necessary inputs used in the model. The hidden layers are results of the summation of all inputs and its corresponding weights and transformed it into a continuous or nominal values using activation functions. The basic neural network functions presents into two main processes, which include feed forward and back propagation methods. The main gist of feed forward method is by providing a randomized weight of all edges, which is being used for all computations in providing prediction results. Error rates will be minimized using back propagation method. This is part of the optimization technique of the neural network algorithm where edges will be updated to minimize the prediction errors from the actual and predicted instances

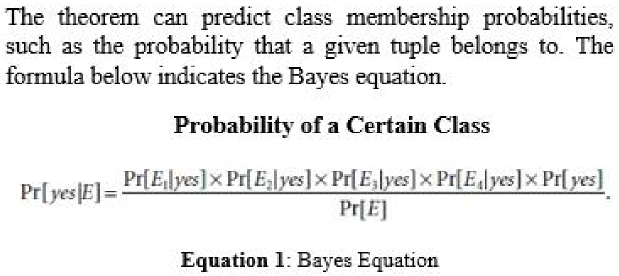
3.6.1 Rudimentary Neural Network Algorithm

1. The weights of the edge will be initialized at arbitrary ways
2. Use feed forward technique to get the actual predicted values.
3. Calculate the prediction error from the predicted and actual values
4. Update the edge weights using back propagation method.
5. Use feed forward technique again to get the actual predicted values and calculate predicted error.
6. The algorithm dismisses when the network blunder is already optimized.

3.7 Naïve Bayes Algorithm



Naïve Bayes is a probabilistic analysis, which follow the Bayes theorem. The said theorem is a powerful formula to determine probability of an event to occur given a set of evidences

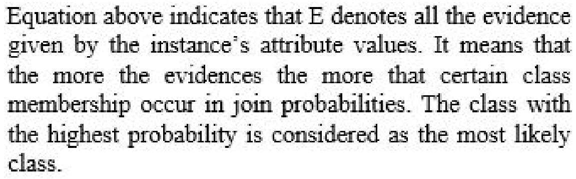


The theorem can predict class membership probabilities, such as the probability that a given tuple belongs to. The formula below indicates the Bayes equation.

Probability of a Certain Class

$$\Pr[yes|E] = \frac{\Pr[E_1|yes] \times \Pr[E_2|yes] \times \Pr[E_3|yes] \times \Pr[E_4|yes] \times \Pr[yes]}{\Pr[E]}.$$

Equation 1: Bayes Equation



Equation above indicates that E denotes all the evidence given by the instance's attribute values. It means that the more the evidences the more that certain class membership occur in join probabilities. The class with the highest probability is considered as the most likely class.

3.7.1 Basic Naïve Bayes Algorithm

1. Calculate prior probability
2. Determine likelihood probability with each attribute for each class
3. Input these value in Bayes Formula and calculate posterior probability

3.8 Decision Tree Algorithm

One classification algorithm usually used for predictive modeling is the decision tree algorithm. The decision tree produces a tree-based classification which is one of the successful classification algorithms used by experts and converted them into powerful classification rules

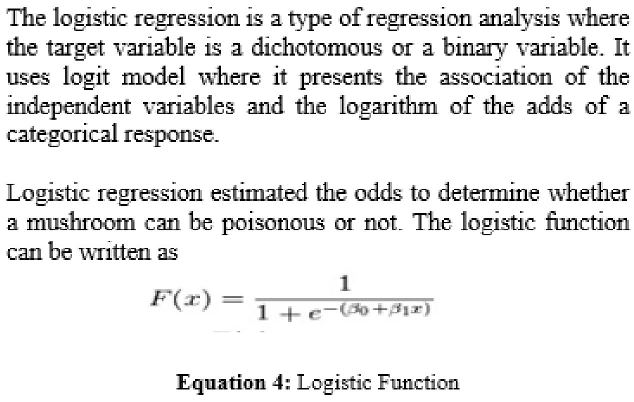
Decision tree algorithm such as ID3 splitting technique is based on the calculation of entropy and information gain. Information gain theory in machine learning refers to the amount of information gained about a random variable or signal from observing another variable. It consists of nodes and creates a tree-like structure which being called decision nodes. [8].

Decision tree is considered as one of the easily understandable classification technique. It utilizes both entropy and information gain in training and learning that can be useful in giving insights about attribute influence [12].

3.8.1 Basic Decision Tree Algorithm

1. Compute the entropy for the entire dataset
2. The following steps must be formed for every attribute
   a. Compute the entropy of each attribute
   b. Determine the average information entropy
   c. Compute the information gain
3. Pick the highest gain attribute that serves as root nod
4. Repeat until we get the decision tree structure.

3.9 Logistic Regression

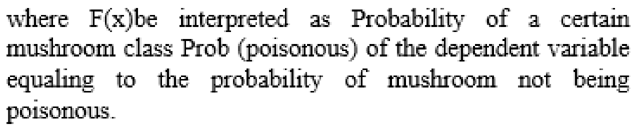


The logistic regression is a type of regression analysis where the target variable is a dichotomous or a binary variable. It uses logit model where it presents the association of the independent variables and the logarithm of the adds of a categorical response.

Logistic regression estimated the odds to determine whether a mushroom can be poisonous or not. The logistic function can be written as

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Equation 4: Logistic Function



where F(x)be interpreted as Probability of a certain mushroom class Prob (poisonous) of the dependent variable equaling to the probability of mushroom not being poisonous.

3.10Cross-Validation Technique

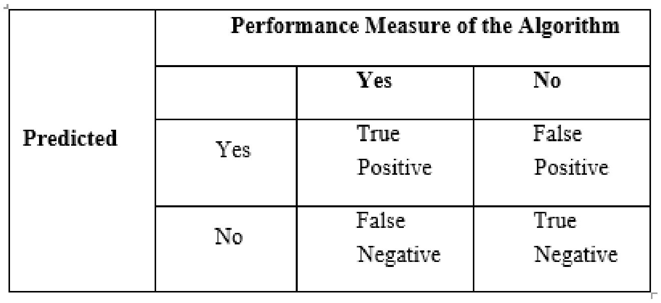
The cross-validation technique is a method to divide the entire data sets in a desired number of folds. This method is being used to avoid over fitting and under fitting. The number of folds indicates how many numbers or iteration test of classier must occur. One fold will be used as test set and the remaining folds will be used as training set. The steps reiterate the procedure until it finished the definite number of folds given.

3.11Evaluation of Discovered Knowledge

The confusion matrix is a useful technique to access the result of the performance of the algorithm[13].

In accuracy rate computation, it determines numbers of of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [14].

Table 3: Classification Rate Table



Predicted	Performance Measure of the Algorithm		
		Yes	No
	Yes	True Positive	False Positive
	No	False Negative	True Negative

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

Equation 5: Accuracy Result Formula

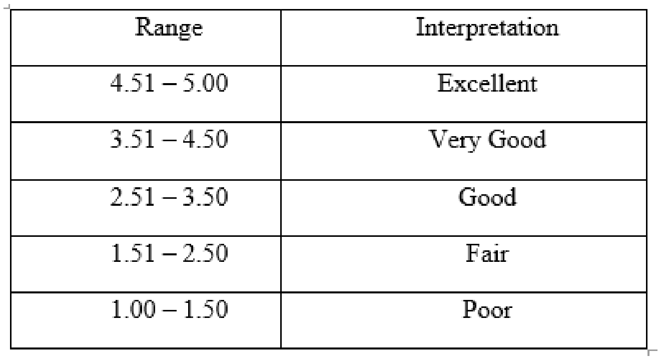
Equation 5 pertains to the accuracy formula, which serves as one of the basis used to measure the performance of an algorithm.

An evaluation instrument was formulated to evaluate the prototype based on quality characteristics of the ISO 9126 software quality model consisting of criteria, which include functionality, reliability, usability, maintenance, and efficiency.

The researches focused only to use the first three characteristics and included design as one of the criteria. The functionality criteria are based on the accurateness and suitability. The reliability focused on metrics on fault tolerance and maturity.

The Five-point Likert Scale was used in rating the system. Mean performance was used to determine the acceptability of each criterion

Table 4: Likert Model



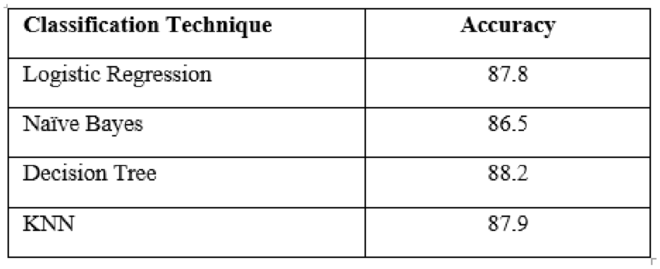
Range	Interpretation
4.51 – 5.00	Excellent
3.51 – 4.50	Very Good
2.51 – 3.50	Good
1.51 – 2.50	Fair
1.00 – 1.50	Poor

4. RESULTS AND DISCUSSION

4.1 Model Comparison Techniques

The confusion matrix table illustrates a tabular display that evaluates the forecasting precision of a predictive model. The summary of the performance of the algorithms is presented below.

Table 5: Summary of Classification Algorithms Accuracy Result

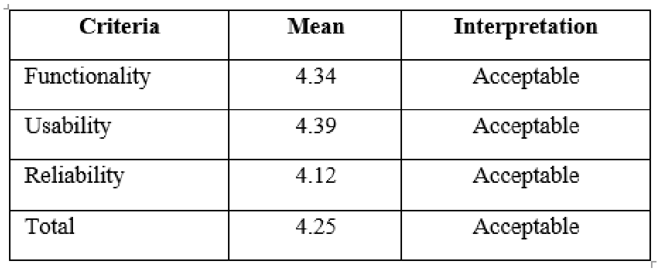


Classification Technique	Accuracy
Logistic Regression	87.8
Naïve Bayes	86.5
Decision Tree	88.2
KNN	87.9

Table 5 presents that decision tree algorithm has the highest accuracy result compare to other algorithms.

The second research question provides the acceptability perspective of the respondents in terms of the developed software application. Table below indicates the result of the survey and its corresponding interpretation.

Table 6: Summary of Evaluation Results Based on Experts Response



Criteria	Mean	Interpretation
Functionality	4.34	Acceptable
Usability	4.39	Acceptable
Reliability	4.12	Acceptable
Total	4.25	Acceptable

As gleaned in the table above, the developed prototype recorded a total mean performance of 4.25, which has an interpretation of acceptable. Noticeably, all other criteria, which include functionality, usability and reliability, have also an interpretation of acceptable. This result conforms that the system is working and can be used in prediction purposes.

5. CONCLUSION AND FUTURE WORKS

The procedures and steps in KDD methodology are very effective in any classification problems. The most crucial parts of KDD refer to preprocessing and modeling. Pre-processing involves different techniques to improve the quality of dataset while modeling checks algorithms' performance and develop predictive and cluster models used to profile and predict future instances.

The Decision Tree algorithm has been one of the most effective machine learning algorithms for predictive modeling. The decision tree produces a tree-based classification used by experts and converted them into powerful classification rules. The data model that can be extracted from the decision tree algorithm can be embedded in a system that will make it as a decision support application.

The developed prototype recorded a total mean performance of 4.25, which has an interpretation of acceptable. This concludes that the software can now be used for prediction.

The study also can be improved by increasing number of data instances and other attributes. This can change the result of the performance of the algorithm and it can lead to extraction of new patterns that can be used for decision support system. The researchers also can utilize the power of other classification algorithms or in combination using ensemble models to increase the performance of the algorithm

6. ACKNOWLEDGEMENT

The researchers would like to thank the UCI Learning Repository – Center for Machine Learning and Intelligent Systems for the dataset by DheeruDua and Casey Graff of University of California that is used in this study.

REFERENCES

1. Shen, Y. (2013). **Wild Mushrooms Classification – Edible or Poisonous.** Retrieved from http://homepages.cae.wisc.edu/~ece539/fall13/project/Shen_rpt.pdf
2. Alkronz, E., Meimeh, M., Moghayer, K., & Gazzaz, M. (2019). **Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation.** Neural Network. International Journal of Corpus Linguistics, 3(2):1-8.
3. Stoltzman, S. (2017). **Random Forest Classification of Mushrooms**. Retrieved from Stoltzmaniac: https://www.stoltzmaniac.com/random-forest-classification-of-mushrooms/
4. Turksoy, O. (2019, April). **Classification Methods on Mushroom Dataset**. Retrieved from Kaggle: https://www.kaggle.com/turksoyomer/classification-methods-on-mushroom-dataset
5. Clifton, Christopher (2010). **Encyclopædia Britannica: Definition of Data Mining**. Retrieved 2010-12-09.
6. Ottom, M., Alawad, N.A., & Nahar, K. (2019). **Classification of Mushroom Fungi Using Machine Learning Techniques**. International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8. No. 5. September-October 2019. https://doi.org/10.30534/ijatsce/2019/78852019
7. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). **From Data Mining to Knowledge Discovery in Databases (PDF)**. Retrieved 17 December 2008.
8. Luqing (2004). **A Preprocessing Method For Nominal Attributes In Classfication And Prediction Problems**. Computer Engineering. Vol. 30, no.3, pp. 92–94.
9. McCallum, Andrew. **Graphical Models, Lecture2: Bayesian Network Represention (PDF)**. Retrieved 22 October 2019.
10. El-Bakry, H. M. Mastorakis N. (2010). **Advanced Technology for E-Learning Development**. Recent Advances in Applied Mathematics and Computational and Information Sciences, Volume II. Pp. 501-522
11. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
12. Jiyadi, R., Firmantyo, H., Dzaka, M., Suaidy, M., & Putra, A. (2019). **Employee Performance Prediction Using Naïve-Bayes**. International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8. No. 6. November – December 2019. https://doi.org/10.30534/ijatcse/2019/59862019
13. Han J. (2006).**Data Mining: Concepts and Techniques Second Edition**. University of Illinois at Champaign Micheline Kamber:Classifier Accuracy Measures
14. Machica, I. Gerardo, B., and Medina, R.(2020). **One-Class Conditional Anomaly Detection Algorithm (OCCADA) for MultipleLinear and Logistic Regression**. International Journal of Advanced Trends in Computer Science and Engineering. Vol. 9. No.16. January-February 2020. https://doi.org/10.30534/ijatcse/2020/66912020