Nepal Journal of Mathematical Sciences (NJMS)
ISSN: **2738-9928 (*online*), 2738-9812 (*print*)**
Vol. 3, No. 1, 2022 (February): 111-116
DOI: https://doi.org/10.3126/njmathsci.v3i1.44130

Research Article
Received Date: December 25, 2021
Accepted Date: February 24, 2022
Published Date: February 28, 2022

Mushroom Classification using Random Forest and REP Tree Classifiers

Nawaraj Paudel[1] and Jagdish Bhatta[2]

[1,2]*Central Department of Computer Science and IT, Tribhuvan University, Kathmandu, Nepal*
Email: [1]*nawarajpaudel@cdcsit.edu.np*, [2]*jagdish@cdcsit.edu.np*
Corresponding Author: Jagdish Bhatta

Abstract: *Mushroom is a reproductive structure produced by some fungi that has a high level of protein and a rich source of vitamin B. It aids in the prevention of cancer, weight loss, and immune system enhancement. There are numerous thousands of mushroom species within the world and a few are edible and a few are noxious due to noteworthy poisons on them. Hence, it is a vital errand to distinguish between edible and harmful mushrooms. This paper focuses on comparing the performance of two tree-based classification algorithms, Random Forest and Reduced Error Pruning (REP) Tree, for the classification of edible and poisonous mushrooms. In this paper, mushroom dataset from UCI machine learning repository has been classified using Random Forest and REP Tree classifiers. The evaluation of these two algorithms using accuracy, precision, recall and F-measure shows that the Random Forest outperforms REP Tree algorithm with value of 100% for accuracy, precision, recall and F- measure. The performance of Random Forest is 100% and is better with respect to REP Tree classifier.*

Keywords: *Mushroom Dataset, Random Forest, REP Tree, 10-fold cross-validation Confusion Matrix.*

1. Introduction

Classification is a supervised machine learning technique that categorizes a data instance in a dataset into a predefined class. Classification algorithms require labeled data set that learn how to assign a class label to the data given to the algorithm after learning. There are many different classification tasks and many different classification algorithms that may be used for classification problems. Each class in the dataset is assigned a label and the main use of classification is to predict the class labels. "Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data)" [5].

As of late, numerous distinctive calculations are utilized in classification and numerous analysts have done their investigation to classify noxious and edible mushrooms utilizing distinctive classification calculations on mushroom dataset. The known species of mushroom are roughly 14,000 in the world and there are 2000 edible species. Among these edible mushrooms, approximately 200 are wild species [9].

This study focuses on comparing two tree-based classification algorithms Random Forest [2],[6],[10] and REP Tree [8],[13],[18] for the classification of edible and poisonous mushrooms based on mushroom datasets [14]. These algorithms have been evaluated based four performance evaluation parameters accuracy, precision, recall and F-measure. Finally, the comparison of these algorithms has been made in order to decide the better algorithm for mushroom classification.

2. Literature Review

There are different researches for mushroom classification that use different classification algorithms to classify poisonous and edible mushrooms.

Ottom and others in [12] implemented and analyzed k-Nearest Neighbor (kNN), ANN, SVM, and Decision Tree, algorithms on mushroom images dataset. They had extracted mushroom image features like Eigen, histogram and parametric. In the research, the kNN results with precision of 94% based on Eigen features and real dimensions and the accuracy of 87% resulted with virtual dimensions.

Ismail and others in [7] performed classification of mushrooms using J48 classifier with an accuracy of 100%. They have used behavioral features like population, habitat, etc. The authors employed Principal Component Analysis (PCA) algorithm for ranking features of the dataset used. Among the 21 attributes used, 'odour' feature is ranked highest with an average of 0.57.

The research done by Verma and Dutta includes use of Artificial Neural Network, Adaptive Nuero Fuzzy Inference System and Naïve Bayes techniques to categorize different mushrooms as edible or non-edible. The performance results are based on accuracy, MAE, and kappa statistic. With the 80% of the training size, the fuzzy based approach is found best among all with an accuracy of 99.87%, MAE of 0.0008 and kappa statistic of 0.9338 [16].

Wibowo and others compared decision Tree (C4.5), Naïve Bayes and Support Vector Machine (SVM) classifiers to classify mushroom data of Agaricus and Lepiota family taken from The Audubon Society Field Guide to North American Mushrooms in UCI machine learning repository. The authors have experimented using 10-fold cross validation with the results or 100% accuracy for decision tree and support vector and 95.82% of accuracy for Naïve Bayes. Despite of C4.5 and SVM performing same in accuracy level, C4.5 is found best from the dimension of process time aspect [17].

Chitayae and Sunyoto used K-Nearest Neighbor (KNN) and Decision Tree methods to classify mushroom using UCI mushroom dataset and compared performance of these two algorithms. The analysis of results in the research indicate the Decision Tree based CART algorithm classifies types of mushroom with an accuracy of 91.93% while the KNN with 89.61% accuracy [3].

Alkronz and others used Multi-Layer ANN with mushroom dataset to foresee whether it is edible or harmful. The ANN configured during the study includes with one input layer, three hidden layers and one output layer. The result showed that the classifier classifies whether mushroom is edible or poisonous with an average predictability rate of 99.25% [1].

Chumuang and others in [4] compared Naive Bayes Updateable, Naive Bayes, Naive Bayes Multinomial Text, SGD Text, LWL, K-Nearest Neighbor (k-NN) and stacking. The results showed that K-NN gave the highest classification accuracy rate of 100%. Both of the Naive Bayes Updateable and Naive Bayes resulted same accuracy of 96.38%. Similarly, LWL and SGD Text has performance of 92.88% and 50.06% respectively. The Naïve Bayes Multinomial Text as well as the stacking performed poor with accuracy of 49.94% only.

The authors of [15] used machine learning algorithm namely support vector machines (SVMs), nearest centroid classifier (NCC), k-nearest neighbor (KNN), deep neural network (DNN) and decision trees for classification of oyster mushroom spawns. The feature used was the trivariate histograms. The study revels contaminated spawns in polypropylene bags can be effectively identified. The classifiers were compared for performance using 4-fold cross validation and the result showed that DNN classifier had the highest accuracy at 98.8%. The Cohen's kappa of 0.93 for both NCC and DNN demonstrated both of the models being robust.

Masoudian and Kenneth used Support Vector Machine algorithm to improve the recognition accuracy and efficiency of the robot to detect mushroom damage either caused by microbial or mechanical origin. They used SIFT algorithm for feature extraction of mushroom images. The sample datasets are classified as class 1 of unhealthy mushrooms and class 2 of healthy mushrooms. The SVM with kernel functions namely, RBF, Polynomial, Sigmoid and Linear are evaluated each having accuracies of 92.6%, 91.33%, 88.66% and 86% respectively [11].

Although lots of researches that have been carried out to classify mushroom dataset using different classification algorithms, this research focuses on using Random Forest and REP Tree to classify mushroom dataset to predict whether the mushroom is edible or poisonous. In particular, this research focuses on comparing the performance of these two classification algorithms using different performance measures.

3. Methodology

The mushroom dataset from Kaggle (https://www.kaggle.com/uciml/mushroom-classification) was used to train and test Random Forest and REP Tree classifiers with 10-fold cross validation using Python 3.10 and IDLE editor. These two algorithms were compared using accuracy, precision, recall, and F-measure. The 10-fold cross-validation is a resampling technique that uses 10 splits of a given data sample. Training and testing is performed 10 times with i^{th} split is reserved as the test set in iteration i and remaining splits are used to train the model.

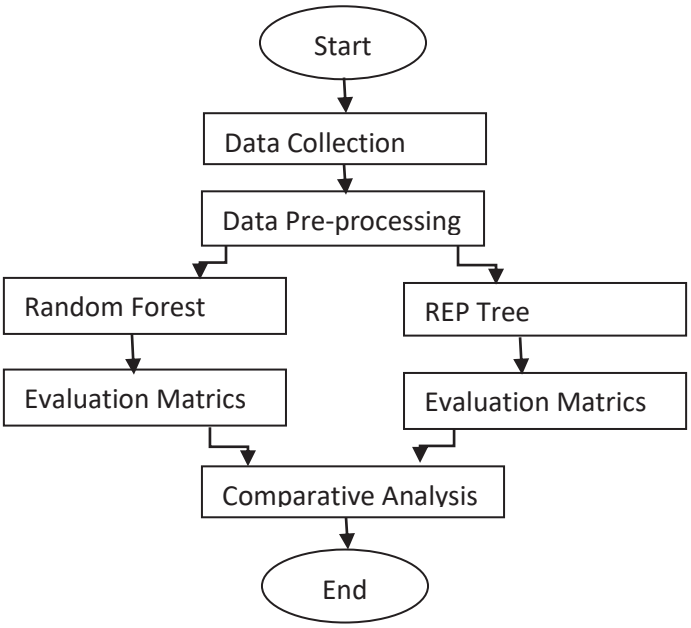

Figure 1 Methodology

3.1. Dataset Used

The data set used in this research is collected from Kaggle (https://www.kaggle.com/uciml/mushroom-classification) [14]. The dataset was originally contributed to the UCI Machine Learning on 27 April 1987. This dataset contains descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). The species in the dataset are either definitely edible, definitely poisonous, or unknown edibility. The unknown edibility is not also recommended and is combined with poisonous one. The dataset consists of 22 different attributes such as cap-shape, cap-surface, cap-color, etc. with different nominal values and a class with labels p for poisonous and e for edible values. There are 8124 total instances with some missing values represented with question mark (?). The dataset has 3916 instances of poisonous mushrooms and 4208 instances of edible mushrooms.

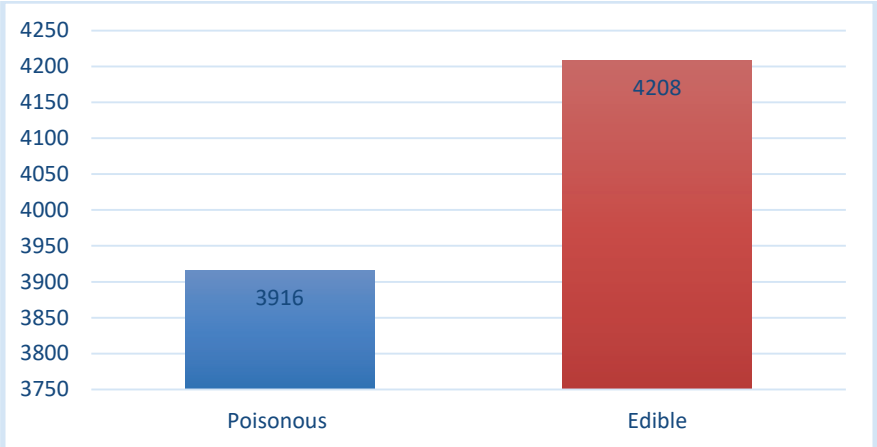

Figure 2 Data distribution in each class

3.2. Algorithms Used

This research has been conducted using two different tree-based classification algorithms as mentioned before.

3.2.1. Random Forest

Random Forest is an ensemble learning method that combines the output of multiple decision trees to reach a single result. This method is mainly used for classification. Ensemble classification is based on multiple classifiers and is more accurate than the individual classifiers. Random forest uses majority voting scheme to determine the class label for unlabeled instances. In majority voting, the class that receives the greatest number of votes is considered as the final decision of the ensemble after each classifier predict the class label of the instance being considered. As a base classifier, a random forest builds a collection of decision trees with controlled variation. Each decision tree in the ensemble is built via bagging, which involves replacing a sample with data from the training set. Each base classifier votes once for its projected class label, and the most popular class label is chosen to categorize the instance [2],[6],[10].

3.2.2. REP Tree

Reduces Error Pruning (REP) Tree is a fast decision tree learning classification technique based on the notion of computing information gain with entropy while minimizing variance-related error. This algorithm was first recommended in [13]. This algorithm applies regression tree logic and generates multiple trees in altered iterations and then picks best one from all spawned trees. This methodology uses information gain/variance to build a regression/decision tree, which is then pruned using the reduced-error pruning with back-fitting method. This approach arranges the values of numeric attributes once at the start of the model preparation process. This approach also deals with missing values by dividing the instances into pieces [8],[13],[18].

3.3. Model Evaluation

Various evaluation matrices can be used to predict accuracy of a classifier. The most commonly used matrices are accuracy, precession, recall, and F-measure [10]. The comparative analysis of two classification algorithms in this research for mushroom classification has been made by measuring the performance of each algorithm using these four matrices.

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

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

Where, True Positive (TP) is the number of observations with label yes that are correctly labeled by the algorithm. True Negative (TN) is the numbers of observations with label no, that are correctly labeled by algorithm. False Positive (FP) is the number of observations with label no that are incorrectly labeled as yes. False Negative (FN) is the number of observations with label yes that are mislabeled as no.

Precision refers to the measure of exactness. That is, what percentage of tuples labeled as positive are actually such.

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

Recall is a measure of completeness. That is, what percentage of positive tuples are labeled as such.

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

The F-measure (also known as the F_1 score or F-score) combines both measures precision and recall as the harmonic mean.

$$\text{F-measure} = \frac{2\times precision\times recall}{pression+recall}$$

4. Experiments and Results

The two classification algorithms were executed on the mushroom dataset using 10-folds cross-validation for the classification of mushrooms based on their class labels. The table below shows confusion matrix of the classification report that has been obtained after testing Random Forest algorithm.

Table 1 Confusion Matrix of Random Forest Algorithm

	Predicted class			
		poisonous	edible	Total
Actual Class	Poisonous	3916	0	3916
	Edible	0	4208	4208
	Total	3916	4208	8124

The table below shows confusion matrix of the classification report that has been obtained after testing REPT Tree algorithm.

Table 2 Confusion Matrix of REP Tree Algorithm

	Predicted class			
		poisonous	edible	Total
Actual Class	Poisonous	3916	0	3916
	Edible	2	4206	4208
	Total	3918	4206	8124

Based on the classification reports shown in Table 1 and Table 2, the calculated summary performance result for the comparison of two algorithms applied on mushroom dataset is shown in the table below. The accuracy, precision, recall and F-measure value are shown is the average of precision, recall and F-measure for both categories.

Table 3 Performance result of two algorithms

Algorithm	Accuracy	Precision	Recall	F-measure
Random Forest	100%	100%	100%	100%
REP Tree	99.98%	99.95%	100%	99.97%

It is clearly seen that the accuracy, precision, recall, and F-measure values of Random Forest is 100% and that of REP Tree is 99.98%, 99.95%, 100%, and 99.97% respectively.

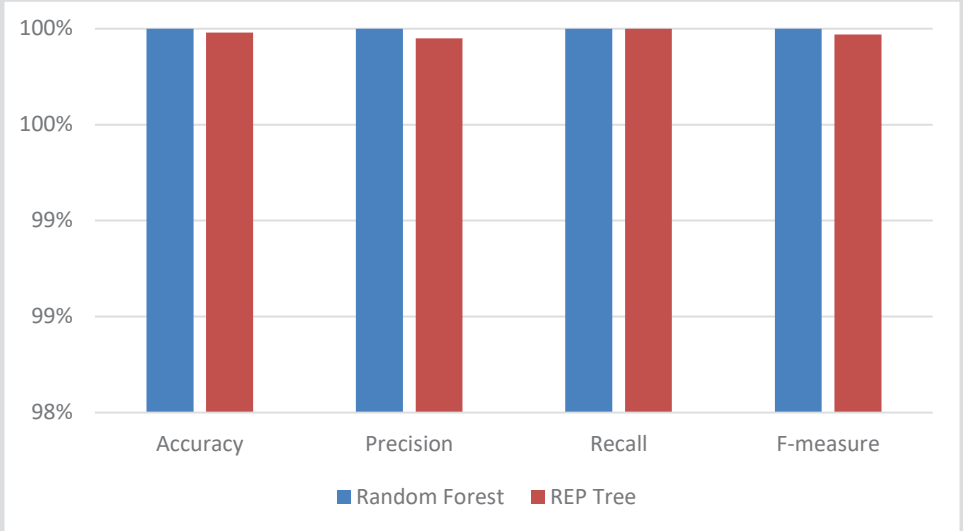

Figure 3 Comparison Chart of Random Forest and REP Tree

5. Conclusion and Future Work

This paper has presented comparative study of Random Forest and REP Tree classification algorithms for classifying mushroom dataset. The final result showed that Random Forest algorithm was superior and more accurate for mushroom dataset classification. This algorithm had higher accuracy, precision, recall and F-Measure score of 100%, 100%, 100%, and 100% respectively whereas, REP Tree had accuracy, precision, recall and F-Measure score of 99.98%, 99.95%, 100%, and 99.97% respectively. Hence,

Random Forest algorithm has high potential to classify mushroom dataset correctly. This work can also be extended to compare other classification algorithms to classify mushroom as well as other datasets.

References

[1] Alkronz, E. S. , Moghayer, K. A. , Meimeh, M., Gazzaz, M., Abu-Nasser, B. S. and Abu-Naser, S. S. **(2019)**. Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network. *International Journal of Academic and Applied Research (IJAAR),* 3(2): 1-8.

[2] Breiman, L. **(2001)**. Random Forest. *Machine Learning,* 45(1): 5-32.

[3] Chitayae, N. and Sunyoto, A. **(2020)**.Performance Comparison of Mushroom Types Classification Using K-Nearest Neighbor Method and Decision Tree Method. *3rd International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia.

[4] Chumuang, N. , Sukkanchana, K. , Ketcham, M.,Yimyam, W. , Chalermdit, J. , Wittayakhom, N. and Pramkeaw, P. **(2020)**. Mushroom Classification by Physical Characteristics by Technique of k-Nearest Neighbor. *15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), IEEE Xplore*, Bangkok, Thailand.

[5] Han J., Kamber M. and Pei J. **(2012).** Data Mining: Concepts and Techniques, Morgan Kaufmann.

[6] Han, J., Kamber, M. and Pei, J. **(2012)**.Classification: Basic Concepts and Advanced Methods in Data Mining: Concepts and Techniques, 3rd ed., Waltham, Massachusetts: Morgan Kaufmann Publishers, , pp. 327-442.

[7] Ismail, S., Zainal, A. R. and Mustapha, A. **(2018)**.Behavioural features for mushroom classification in *IEEE . Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia.

[8] Jayanthi, S. K. and Sasikala, S. **(2013)**. REPTree Classifier for indentifying Link Spam in Web Search Engines. *ICTACT Journal on Soft Computing (IJSC),* 3(2): 498-505.

[9] Kalač, P. **(2009).** Chemical composition and nutritional value of European species of wild growing mushrooms: A review. *Food Chemistry,* 113(1): 9-16.

[10] Kirasich, K.,Smith, T. and Sadler, B. **(2018)** .Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review,* vol. 1, no. 3.

[11] Masoudian, A. and Mcisaac, K. **(2013)**.Application of Support Vector Machine to Detect Microbial Spoilage of Mushrooms. *International Conference on Computer and Robot Vision*.

[12] Ottom, M. A. , Alawad, N. A. and Nahar, K. M. O. **(2019)**. Classification of Mushroom Fungi Using Machine Learning Techniques. *International Journal of Advanced Trends in Computer Science and Engineering,* 8(5): 2378-2385.

[13] Quinlan, J. R. **(1987)**.Simplifying decision trees. *International Journal of Man-Machine Studies,* 27(3): 221-234.

[14] Schlimmer, J. **(2016)**. Mushroom Classification.Safe to eat or deadly poison? Kaggle, 1 December 2016. [Online]. Available: https://www.kaggle.com/uciml/mushroom-classification. [Accessed 7 7 2021].

[15] Tongchama, P. , Supaa, P., Pornwongt, P. and Prasitmeeboon, P. **(2020)**. Mushroom spawn quality classification with machine learning. *Computers and Electronics in Agriculture,* vol. 179.

[16] Verma, S. and Dutta, M. **(2018)**.Mushroom Classification Using ANN and ANFIS Algorithm. *IOSR Journal of Engineering (IOSRJEN),* 8(1): 94-100.

[17] Wibowo, A. , Rahayu, Y., Riyanto, A. and Hidayatulloh, T. **(2018)**.Classification algorithm for edible mushroom identification in 2018, *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia.

[18] Witten, I. H. , Frank, E., Hall, M. A. and Pal, C. J. **(2017)**. Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann.