MACHINE LEARNING TOOLS FOR DATASET CLASSIFICATION

[1]**Kasarapu Ramani**
[1]Professor and Head
[1]Department of IT
[1]Sree VidyanikethanEngg College (Autonomous), Tirupati, India

Abstract*—Mushrooms, as a kind of food, are very special due to their edibility. Some countries treat mushrooms as a kind of high nutrition food. Still, only tiny portions of them are edible. It is really toxic to consume a poisonous mushroom. Thus, we used Decision Tree and Naïve Bayes classification algorithms to develop a best model to classify mushrooms that are edible based on the detected data of the mushrooms. Also the performance of these classification algorithms is tested on two different tools such as Weka and Spark. The performance is expressed in terms of parameters correctly classified instances, incorrectly classified instances, errorrate and precision. The Decision Tree algorithm has outperformed with respect to Weka tool while the Naïve Bayes algorithm outperformed in the Spark.

***IndexTerms*—Machine Learning, Classification, Decision Tree, Naïve Bayes, Mushroom Classification.**

I. INTRODUCTION

The main objective of paper is to study the impact of Decision Tree and Naïve Bayes classification algorithms on the Mushroom Classification dataset in Weka and Spark tools. The parameters for judging the algorithms are correctly classified instances, incorrectly classified instances, errorrate and precision. These are helpful when training data is used instead of testing data and comparing them to know the correctly classified instances, incorrectly classified instances, errorrate and precision of the particular algorithm. This paper is categorized as follows. Section IIinclines the related work. Section IIIgives the procedure and discusses the characteristics of the classification algorithms and the dataset. Section IVgives analysis of the generated by the algorithms. Section Vconcludes the paper.

II. RELATED WORK

The results of [1] proved that the Random Forest Algorithm gives better results on large datasets keeping the same number of attributes while Decision Tree is a finest and easy method for smaller datasets with less number of instances.[2] performed classification of mushrooms using Naïve Bayes, Bayesnet and ZeroR classification algorithms and concluded that the Bayesnet algorithm outperformed the considered three classifiers. In [3],performance of different classifier algorithms namely Naive Bayes, Multilayer perceptron Instance Based K-Nearest Neighbor (IBK), J48 Decision Tree, Simple Cart, ZeroR, CVParameter and Filtered Classifier is analysed using diabetes datasets, nutrition datasets, ecoli protein datasets and mushrooms datasets.

III. METHODOLOGY

The following are the steps included in the classification process carried out in this work:

- Mushroom classification dataset is chosen for the classification process.
- Two different classifiers namely-Decision Treeand Naive Bayes are chosen.
- Two different tools Weka and Spark are used to perform the classification by each of the classifier.
- The correctly classified instances, incorrectly classified instances, errorrate and precision of each classifier are calculated.
- Finally the results are analysed and the best suited algorithm for the chosen dataset is found. The performance of both the tools is also analysed.

III.IDataset

The dataset considered in this work is Mushroom classification dataset from the kaggle data repository. This dataset includes 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). Each species is identified as definitely edible, definitely poisonous. The dataset is composed of 22 attributes with one attribute for the class label. The dataset has atotal of 8124 instances. The class distribution of the dataset is as follows:

"poisonous state" (class 'p') : 3916 (48.2%)
"edible state" (class 'e'): 4203 (51.8%)

III.IIClassifiers

III.II.IDecision Tree

A decision tree classifier is a classifier that classifies the given input model into one of its possible classes. Decision tree classifier is a tree structured classifier that classifies byextracting knowledgethrough making decision rules from the huge amount data. A decision tree classifier is a simple form of classification which is briefly stored and canpowerfully classify new data. The advantages of decision tree classifier are its ability to handle different types of input data such as textual,numerical and nominal. Its ability to handle missing values and errors in the datasets.Itsavailability across various platforms in different packages.

III.II.II Naive Bayes

A Naive Bayes classifier assumes that the incidence of a particular feature in a class is not related to the incidence of any other feature. Naive Bayes classifier is a simple classifier that is based on the Bayes Theorem of conditional probability along with strong independent assumptions. This classifier emphasizes on measure of probability that whether the document belong to a particular class or not. It is an independent feature model. It is based on the assumption that the occurrence or non-occurrence of a specific attribute is unrelated to the occurrence or non-occurrence of a specific attribute. The major benefit of Bayesian classifier is that it needsonly a small training data set for classification. It is efficient, easier for implementation and fast to classify. It is non-sensitive to extraneous features.

III.III Tools

III.III.I WEKA

The full form of WEKA is Waikato Environment for Knowledge Learning. Data pre-processing, classification, clustering, association, regression and feature selection are the standard data mining tasks supported by Weka tool. It is an open source application available. In Weka datasets should be structured to the ARFF format. Weka Explorer provides the classification tasks through the classify tab. Weka uses a variety of classifiers such as Bayes, function, tree etc.

III.III.II Spark

Apache Spark is a general purpose cluster computing engine which is very fast and reliable. This system provides Application programing interfaces in various programing languages such as Java, Python, Scala. Spark tool is specialized at making data analysis faster. The in-memory processing capability of spark makes it much faster than any traditional data processing engine. Spark also provides enormous impressive high level tools such as machine learning tool M Lib, structured data processing, Spark SQL, graph processing took Graph X, stream processing engine called Spark Streaming, and Shark for fast interactive question device. The classification algorithms supported by Spark are part of the Spark machine learning tool mlib.

IV. RESULTS

The experimental setup used includes Windows 10 Operating System,intel core i5 processor, 8GB RAM,Weka tool version 3.8.1 and Spark tool version 1.6.1. The Results of following analysis on the Mushroom classification dataset are clearly given by the tables 1, 2 and 3. Tables 1 and 2 have given the positive and negative instances correctly classified with total number of training and testing instances in the dataset using Decision Tree and Naïve Bayes classifiers in Weka and Spark tools respectively. Table 3 listed the error rate and precision measures to analyse the classifiers in both Weka and Spark.

Comparing the Decision Tree and Naïve Bayes Classification Algorithms in both Weka and Spark tools, it can be concluded that the performance of the Decision Tree Classifier is better on the considered Mushroom classificationdataset in the Weka tool whereas the Naïve Bayes classifier of 100 percentage accurate in the Spark tool. The performance variation between the Decision Tree and Naïve Bayes classifiers is hardly less than 5 percentage with respect to both the tools. The pictorial representation of this analysis is provided through Fig. 1, 2, 3 and 4.

Table 1 Comparing Decision Tree and Naïve Bayes Classification Algorithms in Weka

WEKA Classification Algorithm	**No of Training instances**	**No of testing instances**	**No of positive instances correctly identified**	**No of negative instances correctly identified**	**No of correctly identified instances**
J48(Decision Tree)	5687	2437	1184	1253	2437
NaiveBayes	5687	2437	1083	1242	2325

Table 2 Comparing Decision Tree and Naïve Bayes Classification Algorithms in Spark

Spark Classification Algorithm	**No of Training instances**	**No of testing instances**	**No of positive instances correctly identified**	**No of negative instances correctly identified**	**No of correctly identified instances**
Decision Tree	5666	2458	1239	1163	2402
NaiveBayes	5714	2410	1132	1278	2410

Table 3 Comparing the performance of Classification Algorithms in Weka and Spark Tools

	Error rate	**Precision**
Weka Decision Tree	0	1
WekaNaïve Bayes	0.045	0.954
Spark Decision Tree	0.023	0.977
Spark Naïve Bayes	0	1

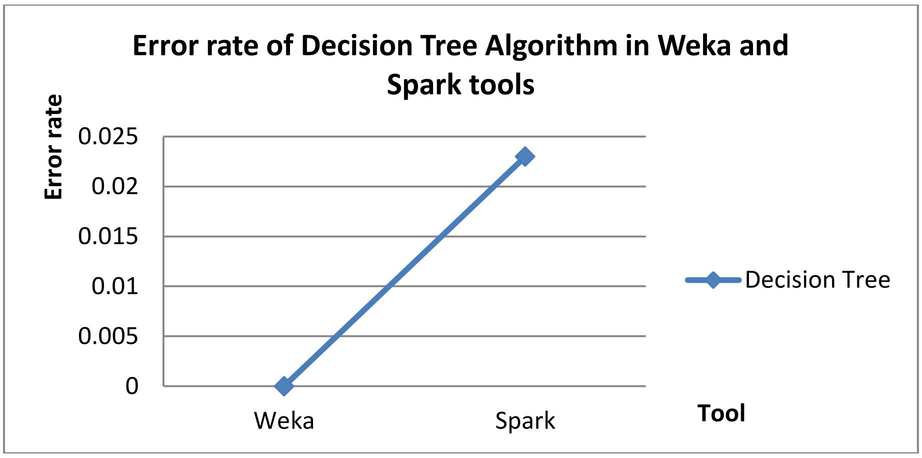

Fig. 1 Comparing Decision Tree with its Error rate in Weka and Spark

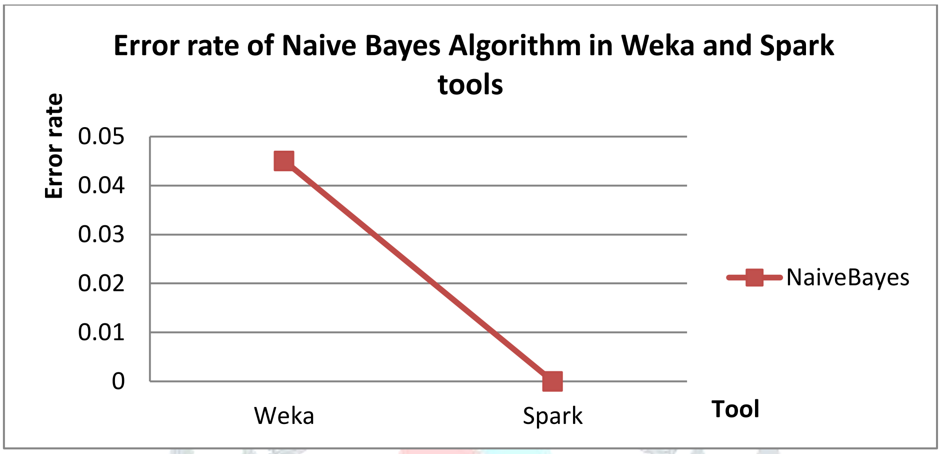

Fig. 2Comparing Naïve Bayes with its Error rate in Weka and Spark

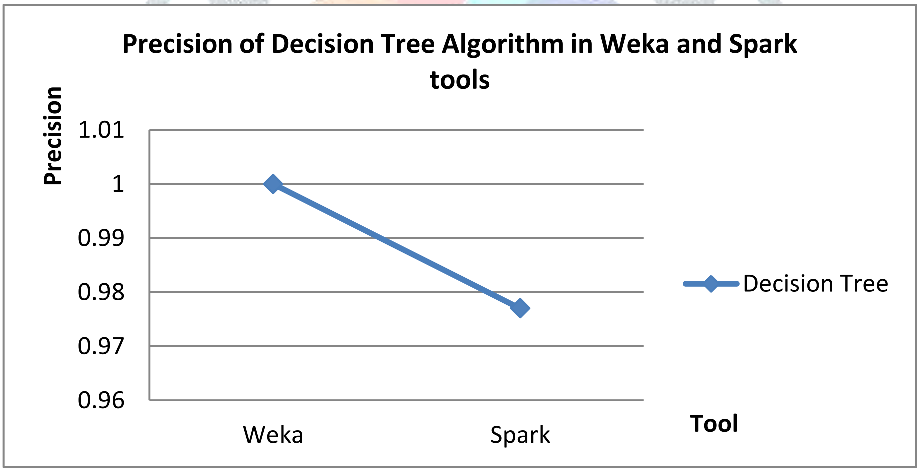

Fig. 3Comparing Decision Tree with its Precision inWeka and Spark

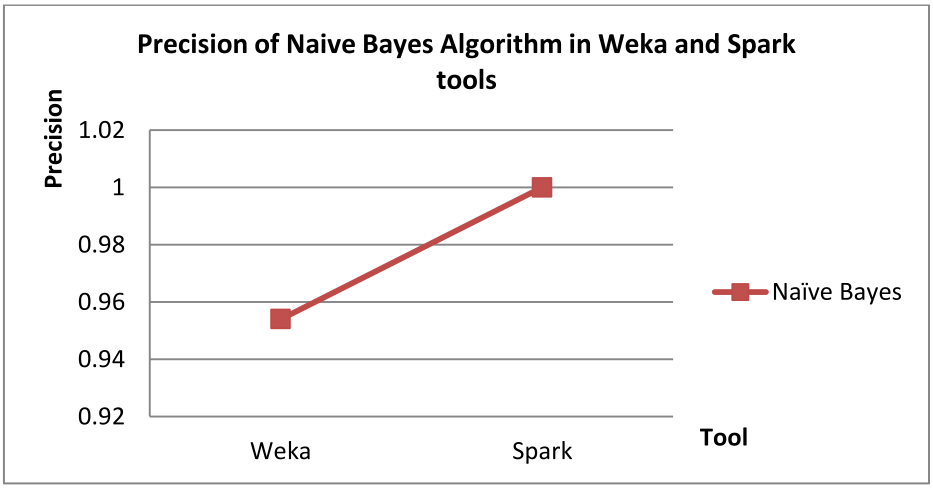

Fig. 4Comparing Naïve Bayeswith its Precision inWeka and Spark

V. CONCLUSION AND FUTURE WORK

In this paper we have compared the performance of Decision Tree and Naïve Bayes classifiers in both Weka and Spark tools. Mushroom classification dataset is used for experimentation from the Kaggledata repository. It is concluded that the performance of Decision Tree classification technique as well as the Naïve Bayes classification technique on the considered data set varied with the tool. The performance of Decision Tree is accurate in Weka while the performance of Naïve Bayes is accurate in Spark.Our future work will focus on improvement of the classification Technique thus improving the effectiveness of classification in reduced time.

REFERENCES

[1] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[2] Beniwal, Sunita& Das, B. “Mushroom classification using data mining techniques”, International Journal of Pharma and Bio Sciences, 6. B1170-B1176, 2015.

[3] A. Swarupa Rani and S. Jyothi, “Performance analysis of classification algorithms under different datasets”, 3rd International Conference on Computing for Sustainable Global Development, 2016.

[4] V. Vaithiyanathan, K. Rajeswari, KapilTajane and Rahul Pitale, “Comparison of Different Classification Techniques Using Different Datasets”, International Journal of Advances in Engineering &Technology,ISSN: 2231-1963, May 2013.

[5] Ananthi S and G. Thailambal, “Comparison of Classification Algorithms in Text Mining”, International Journal of Pure and Applied Mathematics, Volume 116 No. 22, 425-433, 2017.