EasyChair Preprint

№ 2625

Comparing various Machine Learning Techniques for Predicting the Salary Status

Suyash Srivastava, Deepanshu Sharma and Priyanka Sharma

February 10, 2020

Comparing various Machine Learning Techniques for Predicting the Salary Status

Suyash Srivastava* and Deepanshu Sharma*(Student)

Ms. Priyanka Sharma*(Assistant Professor)

** SRM Institute of Science and Technology ,NCR Campus ,Ghaziabad.*

Abstract

Supervised Learning and Unsupervised Learning method is used for classification of the data for predicting that which machine learning technique will classify the data sets of salary status of the people that who are less than or equal to 50000 salary or greater than 50000 salary more efficiently. We take the attributes as age, job_type, ed_type(education type), capital gain, capital loss, race, work hours per week, native country, salary status , relationship ,occupation, marital status, gender. We use four classifier methods Naïve Bayes, Random Tree, Random Forest, REPTree for classifying the data sets. After classifications we apply K-means algorithm for clustering the data.

Keywords

Supervised Learning ,Naïve Bayes ,Random Tree , Random Forest , REPTree ,K-means.

1.INTRODUCTION

In this research paper, we use Weka Software for visualizing the data sets that we take for predicting the solutions for the data sets by taking the age attribute for taking predictions for the salary status.

WEKA

Weka is a software i.e. a group of machine learning procedures and proficiented to examine the data set by the Data Mining and Machine Learning techniques. The algorithms can be linked directly to the dataset. Weka software contains of data preprocessing, classification, regression, clustering, association rules, and visualization methods. It is also an appropriate approach for increasing new Machine Learning ideas. WEKA software is produced by the University of Waikato2. The Weka tool includes the four applications shown in Figure 1.

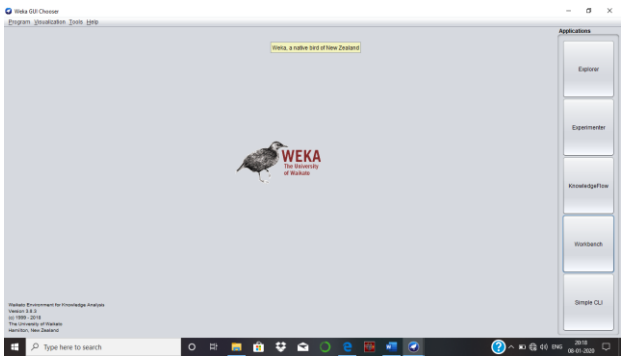

Figure 1: Weka Interface

Weka Tool

Weka Tool is used to find the outcome for the dataset. In 2nd diagram , encompasses the succeeding categories:

- Pre-process: Preprocess is used to ignore repeated data , missing data, error containing data and inconsistent data.

•Classify: It is used for the categorization task. A large quantity of classifiers are used as a part of Weka.

•Cluster: Grouping of the data sets into clusters.

•Associate: Create the connotation instructions for the data sets.

•Select attributes: Choose the attributes in the data.

•Visualize: 2-Dimensional design of the data3. Inside the paper 31,978 numbers of samples are used with machine learning algorithms such as Naïve Bayes,Random Forest,Random Tree and REPTree classifications and K-means used for

clustering. It continues to analyze and find the important attribute for income datasets . The research paper is represented into four parts. Part 2 presents the Methods and Results &Analysis over income dataset implying Weka software as showed at Part 3. Part 4 concludes the paper.

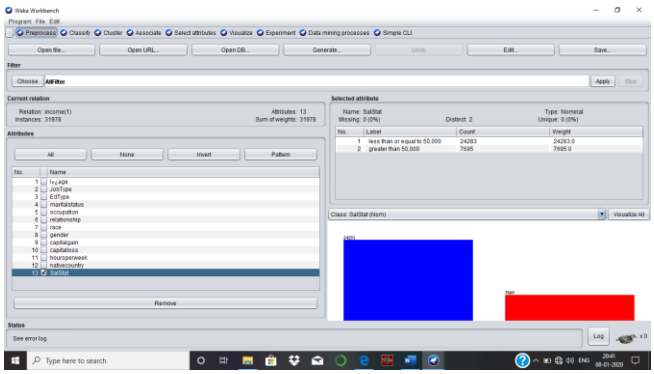

Figure 2: Pre-processing Screen

Visualising

Visualising all the attributes of the data set by plotting bar chart graph of all the attributes.

Figure 3: Bar Chart

2. METHODS

2.1. Classification method in Weka

It consists of three steps for data set in Weka.

1.Prepare the data sets.

2. Put a classification algorithm

3. Analyze the result .

First, organize the data sets in .csv extension and upload the data sets then, select classifying algorithms. At last, Analyze the results.

2.1.1.Prepare the data sets-

Data set is used as a portion in Weka. It consists of various parts of attributes as age, JobType, EdType, maritalstatus, occupation, relationship, race, gender, capitalgain, capitalloss, hoursperweek, nativecountry and SalStat.

2.2. Classification Methods-

2.2.1. Clustering

Clustering is an Unsupervised Machine Learning technique that involves the grouping of data points into clusters. If we given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.

2.2.2. Naïve Bayes

Naïve Bayes is a kind of Supervised machine learning technique of mixture model that can be used for classification or for clustering (or a mix of both), depending on which labels for items are observed. ... Naive Bayes classification and clustering can be applied to any data with multinomial structure.

2.2.3. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

2.2.4. Random Tree

In mathematics and computer science, a random tree is a tree or arborescence that is formed by a stochastic process. Types of random trees include: Uniform spanning tree, a spanning tree of a given graph in which each different tree is equally likely to be selected.

2.2.5. REPTree

REPTree: algorithm is a fast decision tree learner it is also based on C4.5 algorithm and can produce

classification (discrete outcome) or regression trees (continuous outcome). This early random decision trees method combines bagging and random feature selection methods to generate multiple classifiers.

2.2.6. K-means Clustering

It aims to partition n-observations into k-clusters into which each observation belong to the cluster to the nearest mean, serving as a prototype of the cluster.

3.ANALYZE THE RESULT

3.1. Classifiers

Naïve Bayes, Random Tree, Random Forest, REPTree.

3.2. Analysis of the Results

The result of the classification is analyzed and stated built on performance procedures. 10-Fold Cross-Validation technique is used to evaluate the implementation of assembling techniques. By this method, data is classified in ten equally weighted divisions , by the divisions 9 of those are utilised as exercise set and the last one is utilised as a testing set. This is used to elaborate the presentation of classifying techniques. Output is measured by Kappa Statistics, Mean Absolute Error and Root Mean Squared Error and ROC curve metrics7.

3.3. Kappa Statistics

The Kappa Statistics calculates the arrangement of prediction with the original class - 1 defines whole arrangement. If K is equal to 1 is perfect arrangement or If K is equal to 0 is chance of arrangement.

3.4. Mean Absolute Error

This calculates the real attributes explicitly whole magnitude of the individual errors. This is little smaller than the RMSE.

3.5. Root Mean Square Error (RMSE)

It analyzes the correctness. This finds the alterations in standards predicted by the model. The RMSE of the dataset will always be greater or equal to the MAE. If the RMSE = MAE, then every error are is of the equal magnitude.

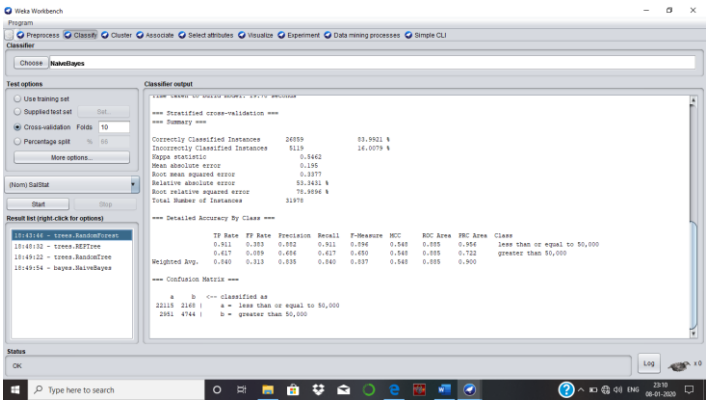

Figure 4:Output of the Random Forest Classifier

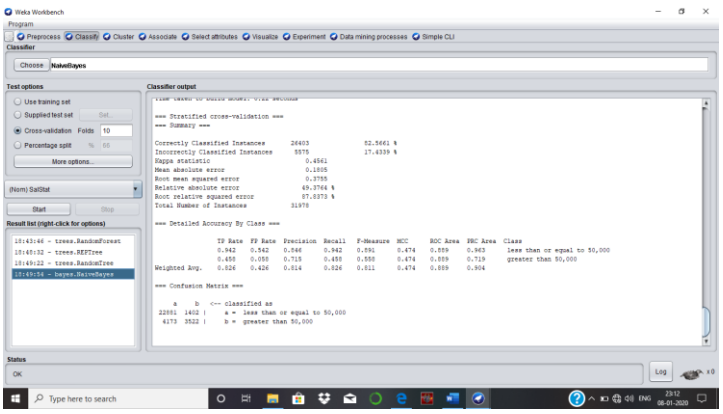

Figure 5: Output of the Naïve Bayes Classifier

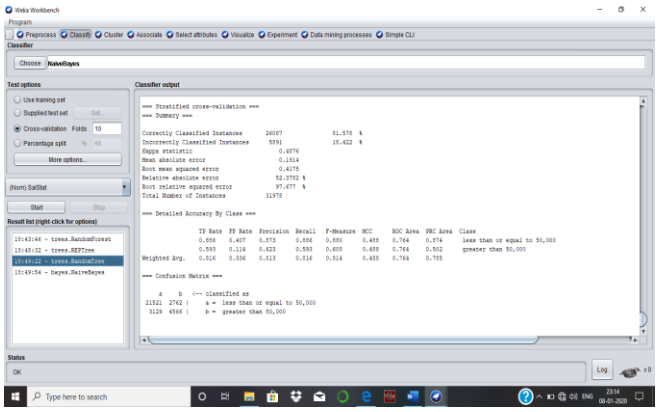

Figure 6: Output of the Random Tree Classifier

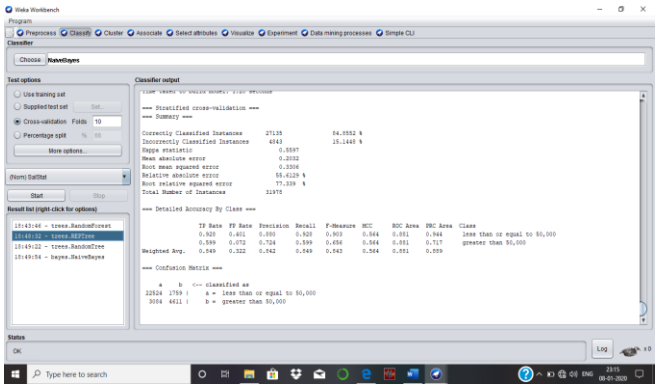

Figure 7: Output of the REPTree Classifier

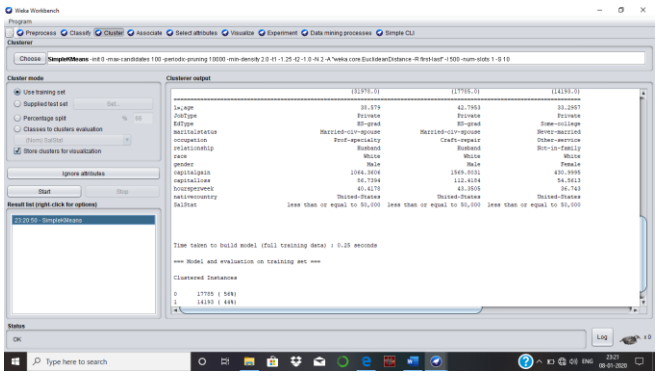

Figure 8 Output of the K-means Clustering

4. CONCLUSION

In this paper we compare unsupervised and supervised algorithms like Naïve Bayes, Random Forest, Random Tree REPTree and K-means for predicting the Salary Status of people using Machine Learning Technique . An income dataset contains 31978 original datasets are utilised. The dataset is analyzed by Weka software and the output of the classification is calculated in the form of their classification correctness by various measurements as Kappa Statistics, Mean Absolute Error and Root Mean Squared Error. It is observed that the correctness of REPTree Technique is most accurate in income dataset prediction in comparison of Naïve Bayes, Random Forest, and Random Tree classifier.

References

1. Dogra AK. A comparative study of selected classification algorithms of data mining. International Journal of Computer Science and Mobile Computing, IJCSMC. 2015 Jun; 4(6):220–9.

2. D. Gao, YX Zhang. Random forest algorithm for classification of multiwavelength data, 2009 National Astronomical Observatories of Chinese Academy of Sciences and IOP Publishing Ltd.

3. Sushilkumar Kalmegh. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, Associate Professor, Department of Computer Science, Sant Gadge Baba Amravati University Amravati, Maharashtra- 444602, India.

4. S.L. Ting, W.H. Ip, Albert H.C. Tsang. Is Naïve Bayes a Good Classifier for Document Classification?, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hum, Kowloon, Hong Kong jacky.ting@polyu.edu.hk .

5. Hlaudi Daniel Masethe, Mosima Anna Masethe. Prediction of Heart Disease using Classification Algorithms, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

6. Weka Data Mining Tool. 23 Dec 2015. Available from: http://www.cs.waikato.ac.nz/ml/wek.