GS International Conference on Computer Science and Engineering 2020 (GSICCSE 2020), AUGUST 2, 2020

Comparison and Analysis of Classification Algorithms on COVID-19 Dataset using WEKA

Syed Mujtaba Hassan
Department of Computer Science
Riphah International University
Lahore, Pakistan
smhassan63@gmail.com

Hafiz Muneeb Ahmad
Department of Computer Science
Riphah International University
Lahore, Pakistan
ahmad.muneeb8470@gmail.com

Azr
a
Department of Computer Science
Riphah International University
Lahore, Pakistan
meshal.zanib@yahoo.com

Syeda Tahira Batool
Department of Computer Science
Riphah International University
Lahore, Pakistan
tahirabatool55@gmail.com

Imdad Hussain
Department of Computer Science
Riphah International University
Lahore, Pakistan
imdaadhussain@gmail.com

***Abstract* — this research study is based on very well-known classification techniques majorly usage by Machine Learning systems, especially in Artificial Intelligent (AI) systems. In this paper main concern is to implementing application using classification techniques on the Pakistan's data set of COVID-19. Purposed Classification algorithms applied to analyze their results and study which techniques is most suitable for the given data set. Weka tool is used to perform the classification algorithm to collect the results.**

Keywords; COVID 19 Dataset; analysis; classification; Naïve Bayes; Random tree; SMO; Tree j48; Lazy K* algorithms; Weka tool

I. INTRODUCTION

Our piratical work are to test implementation, classification algorithm analysis, on the Pakistan COVID-19 data set [1]. Eventually we successfully implement the work on Weka Tool [2] to import the dataset with "Pakistan COVID-19" which is the downloaded from Kaggle.com. In very first part of the technique looks at the characteristic of the data set, including the content of the data and the correlation with the data set. The second part of the article analyzes the algorithm [3] of "Naive Bayes", "SMO Function", "Random Tree", "Trees j48" and "Lazy K*" to control data set. In third step we analyzes the final results of the research listed in the algorithms which used in the Weka tool for analysis.

II. LITERATURE REVIEW

Some researchers used dataset of Breast Cancer and they applied different classification algorithm via using Weka Tool, after applying these classification algorithms they conclude the best classification algorithm accuracy [2]. Some researcher used three different data sets which are Diabetes data set, Spam base data set and credit approval data set, they used Weka Tool to analyses the performance of algorithm in these data sets [3]. Some researchers perform classification algorithm on different data set which are Balance Scale, Diabetes, Lomography and Vehicle using Weka Tool and calculate classification accuracy. Some researchers perform comparison study on Data Mining tool via using some classification methods. Some researcher used Big Data it's around about 5,000+ products types and they performed classification algorithm and check their accuracy classification accuracy on Big Data. Some researchers perform educational Data Mining using classification algorithms and check student performance. Some researchers perform comparative study of different classification techniques in the algorithms of data mining.

III. RESEARCH QUESTIONS

In following research paper we discussed five different classification algorithms and these algorithms are applied on COVID-19 data set, after applying we compare their accuracy and conclude that which algorithm give better result on the given data set.

1. Which algorithm give better accuracy on the Pakistani COVID-19 data set? Do Comparative analysis of different classification algorithms and conclude the accuracy result.

IV. METHODOLOGY

Weka which is free to use and one of the best tool for Data Mining to analyze the result is used, and the Weka Tool version is 3.7.8 The Pakistani data set COVID-19 is taken for checking the performance of each classification algorithms which is used. All the worked done on an Operating System Microsoft Windows 10 with the following specifications Intel @ CoreTM i5 vPro Central Processing Unit is 2.3 GHz processor and the system RAM is 12.00 GB. Pakistani COVID-19 dataset sets accordingly differ in the size like in term of number of attributes.

A. Pakistan COVID-19 Data set

Pakistan COVID-19 Dataset which download from Kaggle.com. Different large amount of instances were selected from this data set, the COVID-19 Data set. This data set include the data of the patients which are affected by the Corona Virus in Pakistan.

Pakistan COVID-19 data set hold in total 42 number of attributes and 112 number of instances. So it's help us to decide either the following person is affected by Corona Virus (COVID-19) or is not. The main attributes are "Suspected Case" (How many of them being suspected), "Confirmed Case" (how many cases are confirmed), "Deaths" (How many die due to this virus), "Recovered" (How many recovered from this virus infection), "Positive Case" (How many cases have positive results), "Negative Case" (How many cases have negative results), "Quarantines Facilities" and "Total tests". The COVID-19 dataset were imported in Weka Tool and applying preprocessing on it.

COVID-19 data set is download from the Kaggle.com as already mentioned in previously discussion. The data set is contained both Nominal attributes with high scale values and the Continual attributes with small scale values and applying preprocessing technique to convert all attributes into Nominal data.

V. ALGORITHMS USED

We choose total five algorithms from many of classification techniques. And these algorithms are SMO Function, Random Tree, Trees j48, Naïve Bayes and Lazy K*, After applying these algorithms we try to find out which algorithm give best results on Covid-19 data set. For this classification techniques purpose we use Weka tool for analysis.

A. Naive Bayes

It is established on the "Bayes" theory and used to determine novel or test data in machine learning [4]. Weka tool is used to carry out this algorithm which offers the chance to configure the algorithm above stated by evaluating the kernel (estimator), for numerical attributes and to carry the use of unsupervised discretization to transform numerical attributes into nominal attributes.

B. Random Tree

This algorithm is construct in Weka Tool under Classification -> tree -> Random tree algorithm. Weka Tool offers opportunity to this algorithm that can be designed to enhance performance. It uses a bagging concept to generate a random set of data for building a decision tree. Classification and regression issues can handle with this algorithm.

C. Lazy K* Algorithm

The Lazy K * sorting algorithm [5] is member of the "Lazy Learners" group. Use the near neighbor technique with a transform vector function. The space among two instances is resolute as a complicate of transforming one instance into another via a predefined order of operations. The execution of K * algorithm is very basic in Weka Tool; it has an "Entropic Auto Blend" indicator that automatically establish the global Blend parameter. Like the other classifiers with the neighborhood of the neighborhood, the K * algorithm can be compared to issues as of the Curse Dimensionality phenomenon. In this study, we will attempt to solve this issue by selecting several qualities.

D. SMO Function Algorithm

During the training of support-vector machines (SVM), quadratic programming (QP) problem appears that is resolve by using Sequential minimal optimization (SMO) algorithm. It is scaled among linear and quadratic in the scope of training data size. It is broadly used for training SVM and is carried out by the well-known Weka tool.

E. Trees j48 Algorithm

This algorithm uses to create a decision tree established by Ross Quinlan. C4.5 is an expansion of Quinlan's prior ID3 algorithm. The decision trees created by C4.5 can be used for classification, and because of this, C4.5 is usually said a statistical classifier. It has been pretty famous after grading no. 1 in the Top 10 Algorithms in Data Mining leading study issued by Springer LNCS in 2008.

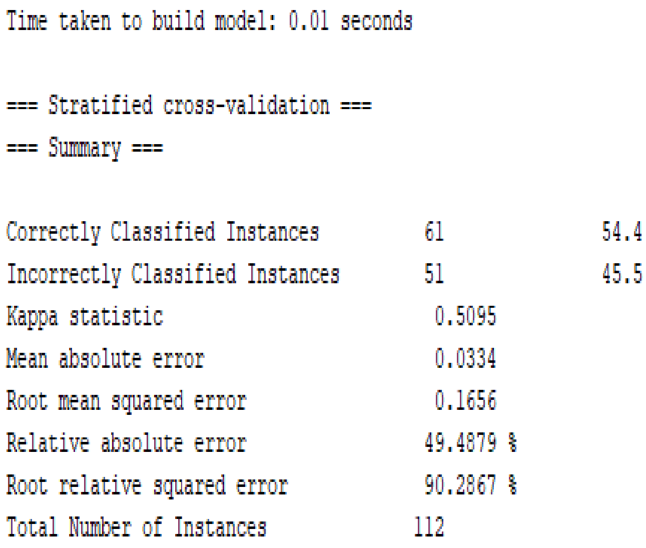
VI. DATASET ANALYSIS

A. Analysis using Naive Bayes Algorithm:

TABLE I. Dataset size and their specifications

Table 1	Datasets sizes and specifications
	COVID-19
Date	04-04-2020
Attributes are	42
Instances are	112
Missing values	N0
Field	Social

COVID-19 Data set: Default parameters in Weka Tool are used to perform initial implementation. This classification is based on carried out at choosing (66% of the training set), future with distant where data were used for training is 80% and 20% remaining is for testing purpose. Final analysis in COVID-19 data set via using Naïve Bayes algorithm is carried out by operationalized of n Supervised Discretization [7]. The framework that is used for turning all the numeric attributes into the nominal attributes. Statistic findings of three used cases are presented below.

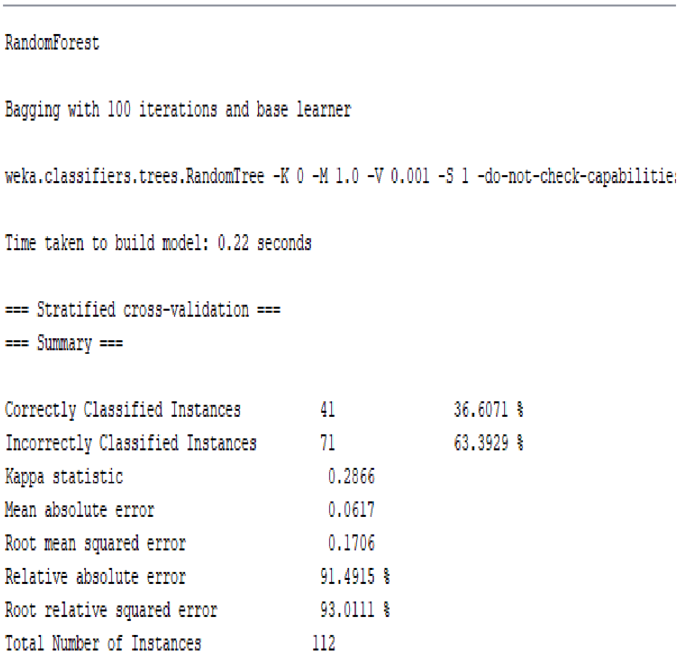


```
Time taken to build model: 0.01 seconds


=== Stratified cross-validation ===
=== Summary ===


Correctly Classified Instances          61               54.4
Incorrectly Classified Instances        51               45.5
Kappa statistic                          0.5095
Mean absolute error                      0.0334
Root mean squared error                  0.1656
Relative absolute error                 49.4879 %
Root relative squared error             90.2867 %
Total Number of Instances              112
```

FIG I. Analysis result of Naïve Bayes algorithm

B. Analysis using Random Tree (RT) Algorithm

l) COVID-19 dataset: In the former section Naïve Bayes algorithm used, now in this section RT is performed in the case of COVID-19 with parameters and not default parameters or values. As a result of the huge number of instances in COVID-19 data set, the time classified for the parameter used will take longer. It will take 0.22s compared to the algorithm used previously. NB will provide more accurate results compare to Random Tree [6].

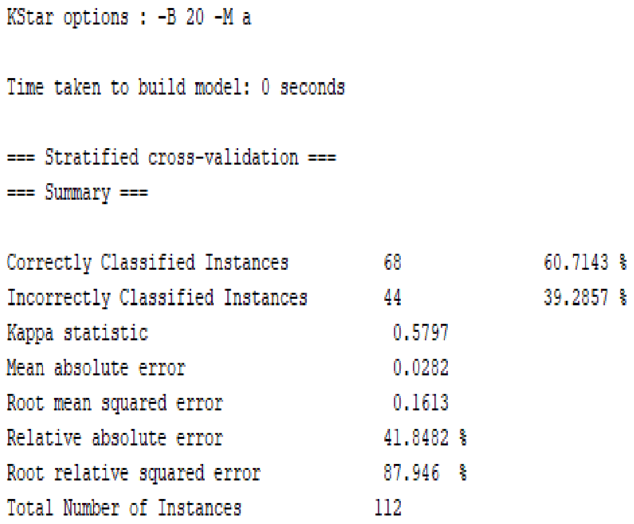


```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilitie

Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          41               36.6071 %
Incorrectly Classified Instances        71               63.3929 %
Kappa statistic                          0.2866
Mean absolute error                      0.0617
Root mean squared error                  0.1706
Relative absolute error                 91.4915 %
Root relative squared error             93.0111 %
Total Number of Instances              112
```

FIG 2. Analysis result of Random Tree algorithm

C. Analysis using K* Algorithm

l) COVID-19 Dataset: As stated over that amount of instances is so enormous in this dataset, it will carry more implementation time however the modification of models will turn rapidly. It will ensure more efficient classification results in fewer time [8].

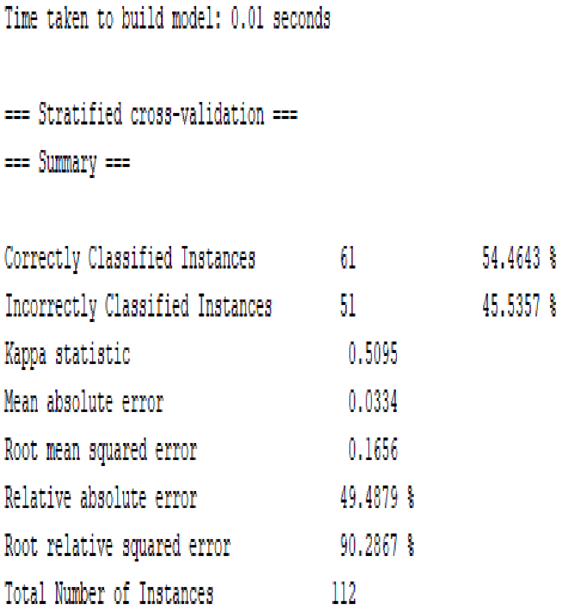


```
KStar options : -B 20 -M a

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          68               60.7143 %
Incorrectly Classified Instances        44               39.2857 %
Kappa statistic                          0.5797
Mean absolute error                      0.0282
Root mean squared error                  0.1613
Relative absolute error                 41.8482 %
Root relative squared error             87.946  %
Total Number of Instances              112
```

FIG 3. Analysis result of K* algorithm

D. Analysis using j48 Algorithm

COVID-19 Dataset: As is mentioned above that the Lazy K* provide more accuracy, it will take more execution time and comparatively show less result as compared to Lazy K* algorithm [9].

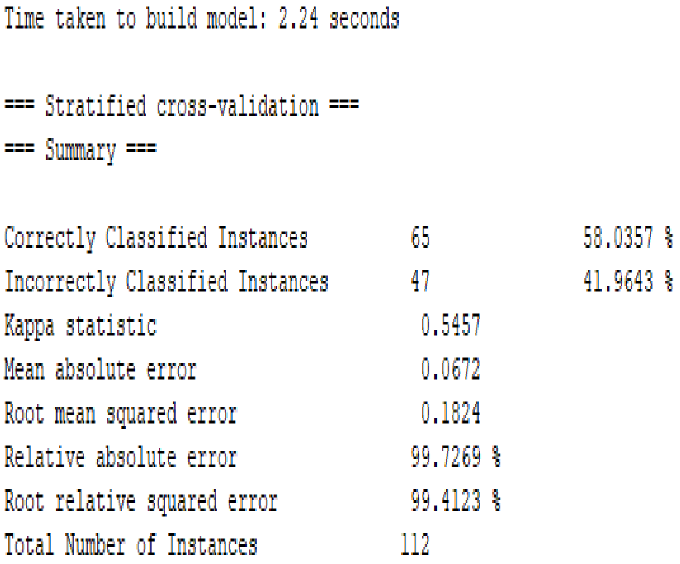


```
Time taken to build model: 0.01 seconds


=== Stratified cross-validation ===
=== Summary ===


Correctly Classified Instances          61               54.4643 %
Incorrectly Classified Instances        51               45.5357 %
Kappa statistic                          0.5095
Mean absolute error                      0.0334
Root mean squared error                  0.1656
Relative absolute error                 49.4879 %
Root relative squared error             90.2867 %
Total Number of Instances              112
```

FIG 4. Analysis result of j48 algorithm

E. Analysis using SMO Functions Algorithm

COVID-19 Dataset: SMO Function algorithms shows the following result drawn in table below



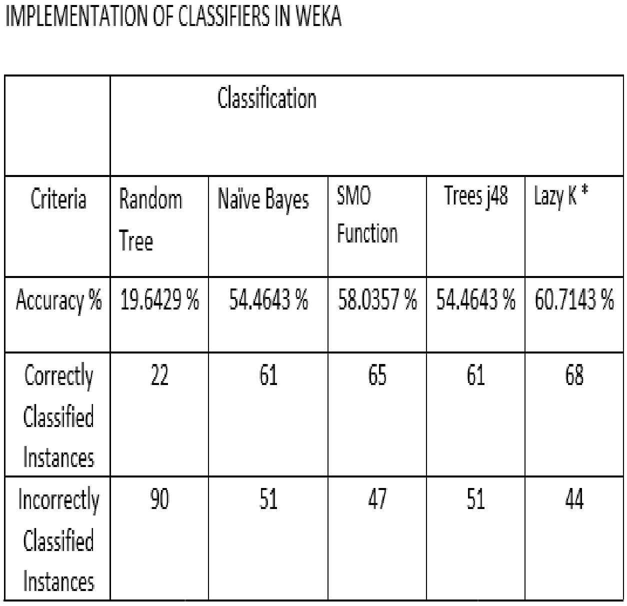
```
Time taken to build model: 2.24 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          65               58.0357 %
Incorrectly Classified Instances        47               41.9643 %
Kappa statistic                          0.5457
Mean absolute error                      0.0672
Root mean squared error                  0.1824
Relative absolute error                 99.7269 %
Root relative squared error             99.4123 %
Total Number of Instances              112
```

FIG 5. Analysis result of SMO function algorithm

VII. RESULTS AND DISCUSSION

After the analysis of all five different algorithms Lazy K* giver better accuracy among all of the algorithms and their accuracy is 60.7143%. Random Tree algorithm give lowest accuracy which is 19.6429% on the data set of Pakistan COVID-19. Comparison table 2 also attached to show the working of all applied five different algorithms on given data set.

TABLE II. Correctly and incorrect classified.



IMPLEMENTATION OF CLASSIFIERS IN WEKA

	Classification				
Criteria	Random Tree	Naïve Bayes	SMO Function	Trees j48	Lazy K *
Accuracy %	19.6429 %	54.4643 %	58.0357 %	54.4643 %	60.7143 %
Correctly Classified Instances	22	61	65	61	68
Incorrectly Classified Instances	90	51	47	51	44

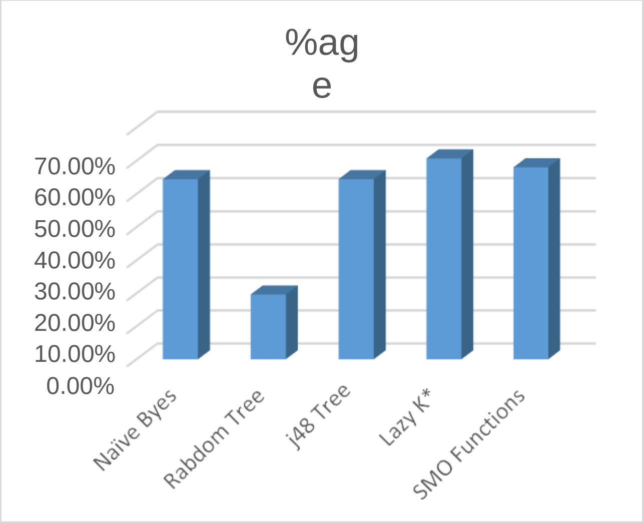


FIG 6. Comparison result of all algorithms

VIII. CONCLUSION

In this study five different classification algorithms performed on COVID19 dataset for analysis. The data set used in our research is different in size and in content too that's why the produced result are different. However when we applied the Lazy K* algorithm on the data set, it has built comparatively high results which is 60.7143%. Lazy K* algorithm was applied in "COVID-19", and it give high accuracy results in the shortest time. Overall, we conclude that Lazy K* technique is best approach among other algorithms on the given data set.

REFERENCES

[1] https://www.kaggle.com/mesumraza/corona-virus- pakistan-dataset-2020

[2] Weka: Java Data Mining Software, University of Waikato.

[3] Mohd Fauzi bin Othman, "Comparison of Different Classification Techniques Using WEKA for Breast Cance., 2007

[4] Rafet Duriqi, Vigan Raca, "Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA", 5th Mediterranean Conference on Embedded Computing2016.

[5] Rohit Arora, Suman,“ Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", International Journal of Computer Application September 2012.

[6] SAGAR S. NIkAM" A Comparative Study of Classification Techniques in Data Mining Algorithms", Computers & Education, 2015

[7] Chong Sun1, Narasimhan Rampalli1, Frank Yang,“ Chimera: Large-Scale Classification using Machine Learning, Rules, and Crowdsourcing".

[8] Chitra Jalota, Rashmi Agrawal,“ Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing Feb 2019.

[9] SAGAR S. NIkAM,“ A Comparative Study of Classification Techniques in Data Mining Algorithms", oriental journal of computer science & technology April 2015.