ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.10, 8207–8215
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.10.50 Spec. Iss. on AOAOCEP-2020

A FRAMEWORK FOR PERFORMANCE ANALYSIS ON MACHINE LEARNING ALGORITHMS USING COVID-19 DATASET

BALAJEE MARAM[1], G. PADMAPRIYA, AND ARAVAPALLI RAMA SATISH

ABSTRACT. Data mining algorithms plays an important role for identifying and getting hidden patterns by using statistical data analytics. Now days, the humans in the world are being suffered from COVID-19 disease. To analyze the reasons and patterns for spreading COVID-19 in the year 2020, the COVID-19 disease dataset is essential for medical data mining. The identification of patterns for spreading COVID-19 may take more time for experienced doctors. In this work, four machine learning classification algorithms are used to develop a framework to analyze the COVID-19 dataset. The experimental setup is based on 4 classifiers with the help of the dataset from KAGGLE to find the performance of each classifier algorithm.

1. INTRODUCTION

In 21^{st} Century humans are busy with smart gadgets, Internet and Technology. So the people are neglecting small health problems, but it leads us to severe health problems or diseases. Definitely there is a need to focus on the small health problems also and need to avoid to go doctor for small health issues which would become severe diseases.

Now a days, the people are being suffered from many diseases. Every disease starts with some set of symptoms. So observing the symptoms are very important for humans as well as hospitals before going to provide the treatment. Now

[1]*corresponding author*

2020 *Mathematics Subject Classification.* 68V30, 68W50.

Key words and phrases. COVID-19, Classification, Data-mining, Machine Learning, Prediction.

a days, all the diseases can be identified with symptoms. In the technical world, a large number of datasets which contains symptoms, diseases of various patients in the world. So there is a need to analyze the entire datasets and give the conclusions like which type of treatment is perfectly suitable to the needy patient [1].

The objective of this paper is to provide the predictions of list of diseases which are being suffered by humans based on 4 classifier algorithms. For implementing the concept, there is a need to collect the real-time data in the society about current health status, analyzes the data and gives predictions using different techniques like Machine learning, Naïve Baye's, decision tree etc.

2. Literature survey

The paper "Comparison of Data Mining Classification Algorithms Determining the Default Risk, Scientific Programming" [2] predicts the default risks and avoid possible payment difficulties. In the paper "A study on data mining prediction techniques in healthcare sector" [3] has focused on Knowledge Discovery in Databases. It quotes that Decision tree prediction is not giving best when the attributes are not related. "Multi Disease Prediction Using Data Mining Techniques" (2017) compares the performance of data mining classification techniques for predicting various diseases. It explains to build effective algorithm for medical applications. In the paper [4] explains the construct is Feature choice ways. It trains support vector machine and Naïve-base algorithms, which gives the results for text classification.

This paper [5] presented the concept is, "Improving disease prediction by machine learning". It gives predictions based on the medical data, personal data and history of the patient which is available in the hospitals using KNN and SVM algorithms. On 10th March, 2020, a new type of pneumonia has been started in Wuhan, China. It has infected 114,350 humans and 4,023 were killed in 107 countries [6]. This new pneumonia has been named as coronavirus 2019(2019-ncov) and renamed as COVID-19 in Feb, 2020 [7]. Transmission rate of coronavirus is very high in public places. So promoting the use of face marks and reduce the travelling gives a fruitful outcome is 10% reduction in virus transmission rate and decrease in deaths is 23-49% [8-9].

3. DATA MINING TECHNIQUES

Data mining techniques are accustomed explore, analyze and extract helpful medical knowledge victimization complicated algorithms so as to find the unknown patterns. Researchers are applying altogether completely different processing techniques that unit of measurement as follows:

Classification. Classification is a data mining technique which analyzes the attributes of different data types in a dataset. Then the data is divided into different classes based on the identified characteristics. The data classification has 2 steps, which are Learning and Classification. In Learning, the predefined algorithms is applied and the classification rules will be constructed. Whereas in classification, predicts the class labels for the given data.

Clustering. Clustering is the process of taking all the objects and check that object belongs to which group and finally it forms a group of objects in such a way that all the objects in each group are more similar.

Regression. Regression is a model which identifies the likelihood of variable using the values of other related variables. For example, the cost of the house in a city after 10 years.

Outlier Detection. Recognizing the pattern is not a big task. But in rare cases, it is little bit difficult to understand a dataset. Because the dataset contains some outliers. Detecting outliers is a major task in data analysis. For example, a product is exclusively for male, but in some particular month, the same product is preferred by female also.

Tracking Patterns. Identifying the patterns plays a vital role in predictions. Before going to analyze the data, there is a need to identify the patterns in the dataset. For example, the sales of GOLD is increasing before festivals, marriages and other major functions.

Prediction. It is a model which gives the future value based on the historical data. For example, the sales of a product in the month of May in next year based on the sales of the same product in earlier years.

Associate rules. The associate rule says that support and confidence are used to find out the associated items. Here the confidence gives all the transactions where is the item1 and item2 are selected both at a time.

4. Machine Learning Classification

Machine learning is the latest technology where the computer system learns the different data sets and enables the machine to learn and give its own decisions. The machine learning algorithms learns the patterns given by the same user earlier. It can predict the marks of a student, rainfall on specific day, traffic on specific day, road accidents in specific month, sales in a month etc.
The k-nearest neighbors formula stores all the cases and finds new cases supported keep cases. Thus it acts as each classifier and regression. It's wont to predict the values of a brand new knowledge points and assigns a worth supported however closely it matches the points within the coaching set. Random Forest Classifier is most power formula for prophetic analytics. It's supported multiple call trees. Every branch of the tree is one form of call and gets the prediction. Finally it selects the most effective answer. It's appropriate for big knowledge bases and offers correct results by estimating missing data.

Decision-tree is a classifier, where the root of the tree mustn't have incoming edges and one incoming edge is allowed for others. Every internal node splits the input attribute values into 2 or a lot of sub-spaces supported the need. Every leaf is allotted to most applicable target worth. This algorithm works on both categorical and numerical data.

Support Vector Machine could be a machine learning methodology for supervised categoryification that is employed to search out the optimized hyperplane by analyzing the coaching dataset samples set at the sting of the potential class. The performance indicators of this study are as follows:

Mean Squared Error (MSE). It is an absolute difference between the actual and predicted values. The formula is given in Eq. 4.1. In which, A is the actual data value, B is the forecasted data value, and n is the sample size.

$$MSE = \text{Average of (True values - Predicted values)} = \frac{1}{n} \times \sum (A - B)^2. \tag{4.1}$$

Root mean squared error (RMSE). It's the square root of the average of absolute difference between the actual and predicted values. Hence, the expression

for RMSE is given in Eq. 4.2. In which, C_i is i^{th} predicted value in the dataset and D_i is i^{th} observed value in the dataset.

$$RMSE = \sqrt{MSE} = \sqrt{\sum \frac{(C_i - D_i)^2}{n}}. \tag{4.2}$$

Classifier performance index. The ability of the classifier can be determined by the confusion matrix which analyzes the performance of classifier. The list of parameters are True Positives (TP), True Negatives (TN), False Positives(FP) and False Negatives(FN).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.3}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4.4}$$

5. PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS

5.1. **DATA SET DESCRIPTION.** The objective of this research work is to check the performance and identifying the best classification algorithm. The 1st step is to train the classifier with some set of records. Here the entire dataset is split training-data and testing-data with 70%, 30% respectively. The above procedure is explained by the following flowchart Fig. 1. The database for this research

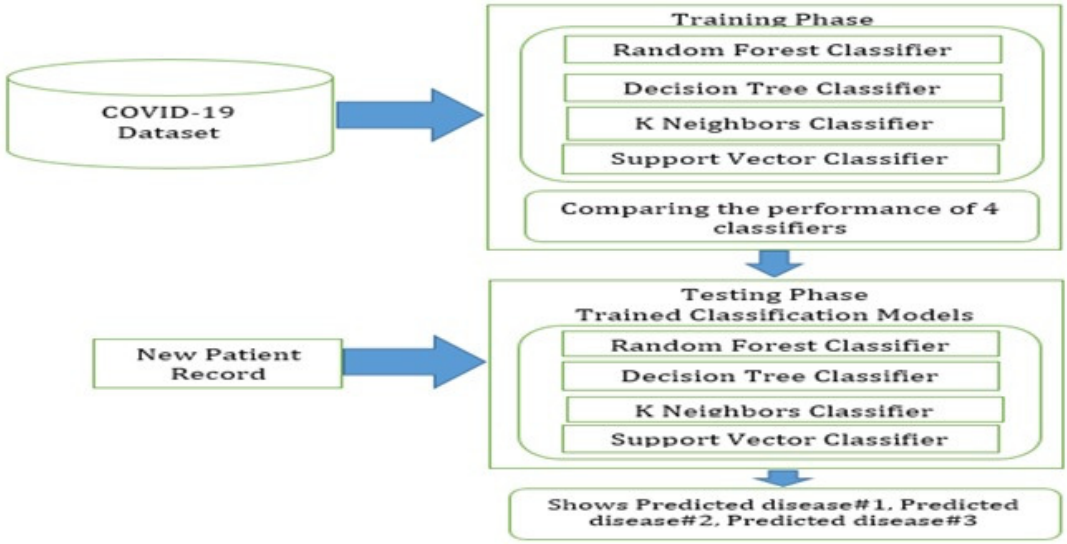

FIGURE 1. Performance comparison of classification algorithms

work has been taken from KAGGLE repository. It contains various dataset files. A single file has been taken for this research work which contains 21 attributes. Out of 21 attributes, 17 useful attributes have been considered for this research work. This dataset contains 1084 records.

6. Results and discussion

The performance of Support Vector Classifier is shown in Table 1. According

TABLE 1. Statistics of performance metrics of Support Vector Classifier Algorithm

SVM Kernel-Name	**Accuracy on training set**	**Accuracy on testing set**	**Accuracy of test-set and predicted-set**	**MSE**	**RMSE**
Linear	0.86	0.89	0.886	0.113	0.336
Polynomial	0.86	0.88	0.877	0.122	0.350
Gaussian radial basis function	0.86	0.89	0.886	0.113	0.336
Sigmoid	0.81	0.85	0.846	0.153	0.391

to Table 1, the accuracy of test-set and predicted-set is good in the kernel like Linear and Gaussian radial basis function than polynomial and sigmoid. The performance of K-Neighbors Classifier Algorithm is shown in Table 2.

According to Table 2, the accuracy of test-dataset and predicted-dataset is maximum and minimum where the number of neighbors are 5, 8 and 9, 13, respectively. The performance of Decision-Tree Classifier Algorithm is shown in Table 3. According to Table 3, the accuracy on training-dataset is good when the number of maximum features are 1,2,3,4. The accuracy of test-set and predicted-set is maximum when the number of features are 1 and minimum when 2. The performance of Random Forest Classifier Algorithm is shown in Table 4.

According to Table 4, the accuracy on training-set is maximum when the number of estimators are 2,3,4,5. The accuracy of test-set and predicted-set is maximum when the number of estimators are 3 and 5.

TABLE 2. Statistics of performance metrics of K-Neighbors Classifier Algorithm

Number of neighbors	**Accuracy on training set**	**Accuracy on testing set**	**Accuracy of test-set and predicted-set**	**MSE**	**RMSE**
1	0.98	0.83	0.828	0.171	0.414
2	0.91	0.89	0.886	0.113	0.336
3	0.91	0.87	0.874	0.125	0.354
4	0.88	0.89	0.886	0.113	0.336
5	0.88	0.89	0.889	0.110	0.332
6	0.86	0.89	0.886	0.113	0.336
7	0.87	0.88	0.883	0.116	0.341
8	0.86	0.89	0.889	0.110	0.332
9	0.87	0.88	0.880	0.119	0.345
10	0.86	0.88	0.883	0.116	0.341
11	0.86	0.89	0.886	0.113	0.336
12	0.86	0.88	0.883	0.116	0.341
13	0.86	0.88	0.880	0.119	0.345
14	0.86	0.88	0.883	0.116	0.341
15	0.86	0.88	0.883	0.116	0.341
16	0.86	0.88	0.883	0.116	0.341
17	0.86	0.88	0.883	0.116	0.341

TABLE 3. Statistics of performance metrics of Decision-Tree Classifier Algorithm

Number of Maximum Features	**Accuracy on training set**	**Accuracy on testing set**	**Accuracy of test-set and predicted-set**	**MSE**	**RMSE**
1	0.99	0.91	0.911	0.880	0.298
2	0.99	0.89	0.892	0.107	0.327
3	0.99	0.90	0.898	0.101	0.318
4	0.99	0.90	0.901	0.098	0.313

TABLE 4. Statistics of performance metrics of Random Forest Classifier Algorithm

Number of Estimators	**Accuracy on training set**	**Accuracy on testing set**	**Accuracy of test-set and predicted-set**	**MSE**	**RMSE**
10	0.98	0.91	0.907	0.092	0.303
100	0.99	0.92	0.917	0.082	0.287
200	0.99	0.91	0.914	0.085	0.293
500	0.99	0.91	0.911	0.088	0.298
1000	0.99	0.91	0.914	0.085	0.293

7. CONCLUSION

This paper evaluates the performance of classifier algorithms using the dataset which belongs to COVID-19. Accuracy of the pair's test-set, predicted-set and training-dataset and test-dataset are best with Random Forest Classifier and Decision-Tree Classifier. The class label is very important to check the test-data. Another parameter Root Mean Squared Error in K-Neighbors Classifier Algorithm and Support Vector Classifier algorithm is high. So the performance of Random Forest Classifier Algorithm is good comparatively other 3 classifier algorithms. It concludes that the performance of any algorithm is depends on the type of the algorithm and number of attributes in the algorithm. For COVID-19 dataset, Random Forest Classifier Algorithm generates more accuracy than other three classifier algorithms.

REFERENCES

[1] Y. YE, Y. XIONG, Q. ZHOU, J. WU, X. LI,X. XIAO: *Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study*, J Diabe Res., **2020** (2020), 1–10.

[2] B. ÇIĞŞAR, D. ÜNAL: *Comparison of Data Mining Classification Algorithms Determining the Default Risk*, Scienti Program., **2019** (2019), 8706505.

[3] B. SRINIVASAN, K. PAVYA: *A Study on Data Mining Prediction Techniques in Healthcare Sector*, Inter Res J Engi Tech., **3**(3) (2016), 552–556.

[4] B. VARMA, B. SENTHIL: *A Different Type of Feature Selection Methods for Text Categorization on Imbalanced Data*, J Netw Comm Emer Tech., **5**(9) (2016), 297–303.

[5] S. MUKESH SINGH, D.B. HANCHATE: *Improving Disease Prediction by Machine Learning*, Inter Res J Engi Tech., **5**(6) (2018), 1542–1548.

[6] DEBORAH BIRX: *White House Coronavirus Task Force*, Centers for Disease Control and Prevention - Coronavirus Disease 2019, (2020), 1–4.

[7] N. IMAI, I. DORIGATTI, A. CORI, S. RILEY, N.M. FERGUSON: *Report 2: Estimating the Potential Total Number of Novel Coronavirus Cases in Wuhan City, China*, Imper Col London COVID-19 Res Team., (2020), 1–5.

[8] M. SHEN, Z. PENG, Y. XIAO, L. ZHANG: *Modelling the Epidemic Trend of the 2019 Novel Coronavirus Outbreak in China*, Quanti Biology., **8** (2020), 11–19.

[9] C.B. SIVAPARTHIPAN, B. ANANDMUTHU, G. MANOGARAN, M. BALAJEE, R. SUNDARASEKAR, S. KRISHNAMOORTHY, C.H. HSU, C. KARTHIK: *Innovative and Efficient Method of Robotics for Helping the Parkinson's Disease Patient using IoT in Big Data Analytics*, Trans Emerging Tel Tech., **e3838** (2019), 1–11.

[10] B.S. KUMAR, S. KARTHIK, V.P. ARUNACHALAM: *Upkeeping Secrecy in Information Extraction using 'k' Division Graph based Postulates*, Cluster Comp., **22**(1) (2019), 1–7.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GMR INSTITUTE OF TECHNOLOGY
RAJAM, ANDHRA PRADESH, INDIA
Email address: balajee.m@gmrit.edu.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SAVEETHA SCHOOL OF ENGINEERING
SAVEETHE INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES, CHENNAI, TAMIL NADU, INDIA
Email address: padmapriyag.sse@saveetha.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
VIT-AP UNIVERSITY
GUNTUR, ANDHRA PRADESH, INDIA
Email address: rama.satish@vitap.ac.in