Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

Analysis of Influencing Risk Factors for Covid-19 Infection Based on the Predictive Models Using Machine Learning Algorithms

[1]Dr.G.Sofia Jonathan, [2]N.Anjali

[1]Associate Professor, Lady Doak College, Madurai
[2]Lady Doak College, Madurai
*Corresponding Author : gsofia@ldc.edu.in,

Abstract.
Over recent years, machine learning has turned very reliable in the medical field. Prediction of COVID-19 by using Machine Learning algorithms would help to increase the speed of disease identification resulting in reduced mortality rate.This work focuses on therole of machine learning algorithms for the identification of risk factors that influences the prediction of COVID-19 infection. This will enable the people to be cautious about the risk factors and directs for proper medical treatment. COVID-19 data provided by WHO with 5434 patient's records and 19 features collected from kaggle.com is considered for this study.Collected data is preprocessed to make it suitable for further analysis. Dependent features are identified and channelized based on the influencing risk factor through exploratory data analysis. Based on the influencing factors, Predictive models are built by considering 70% of data for training and the remaining 30% for testing. These models are evaluated and experimental results are analyzed thatindicates that the Random Forest algorithm gives higher accuracy by minimizing RMSE. This work also endorses that the higher accuracy 93.25% with minimum RMSE value 25.96% is attained when the prediction is based on the symptom. Hence this work concludes that symptom is the higher influencing risk factor that needs to be addressed promptly during COVID-19 infection.

Keywords: COVID-19, Machine Learning, Correlation, Feature Engineering, Predictive Model

INTRODUCTION:

[5]The 2019 novel coronavirus, better known as COVID-19, was reported in Wuhan, China, for the very first time on 31st December, 2019. Coronaviruses are family of viruses that can cause illnesses such as the common cold, severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS).In March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak as a pandemic.The severity of COVID-19 symptoms can range from very mild to severe. Some people may have only a few symptoms, and some may have no symptoms at all. Some people may experience worsened symptoms, such as worsened shortness of breath and pneumonia, about a week after symptoms start.People who have existing chronic medical conditions also may have a higher risk of serious illness. [7]The basic symptoms for COVID-19 include fever, exhaustion, shortness of breath, and loss of smell and taste. Washing hands, covering the face, isolation, and social distancing may be a way to prevent this communicable disease, but it is not enough. When individuals are in close contact with the affected persons, the little droplets created by a contaminated individual while sniffling will infect the other person.

Having been declared a Public Health Emergency of International Concern, the pandemic is transmitted through direct contact with an infected person's bodily fluids either through sneezing and coughing[3,8]. Additionally, asymptomatic cases and lack of diagnosis kits result in delayed or even missed diagnosis, exposing patients, visitors, and healthcare workers to the COVID-19 infection. This poses a great threat to the healthcare and economic sectors. Therefore, it is clear that nonclinical techniques such as machine learning, data mining, expert system and other artifcial intelligence techniques must play critical roles in diagnosis and containment of the COVID-19 pandemic[3f[.Hence this work focuses on building the predictive models using machine learning algorithms to analyze the influencing risk factor for COVID-19.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

METHODS

[7]In the healthcare industry, there is a lot of evidence that machine learning algorithms can provide effective models to solve problems in order to identify patients.Machine learning (ML) is one of the most advanced concepts of ArtificialIntelligence (AI), and provides a strategic approach to developing automated, complex and objective algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis[4-3].
[3]ML techniques can be classified asfollows:
1. **Supervised learning** techniques are ML techniques or algorithms that bind existing data in the dataset with the help of labeled data to predict future events. The learning process begins with a training process and develops targeted activity to predict output values. These techniques are able to provide results based on input data with an adequate training process and compare results with actual results and expectations to identify errors and modify the model according to the results[3].
2. **Unsupervised learning** techniques are ML techniques that are used when the training dataset is non-labeled. These learning techniques deduce a function to extract hidden knowledge or a pattern from unlabeled dataset[3].
3. **Semi-supervised learning** techniques are ML techniques that lie between supervised learning techniques and unsupervised learning techniques, where labeled and non-labeled datasets are used in the training process. These learning techniques consider a smaller labeled dataset and a larger unlabeled dataset. The learning techniques are preferable when a labeled dataset needs competent and appropriate resources for training or learning in it[3].
4. **Reinforcement learning** techniques are ML techniques that interact with the learning environment by actions to locate errors.These techniques are used to identify the ideal behavior in a specific context to increase the performance of the model. Trial and error searches through feedbacks are common in reinforcement learning[3].

In this work, supervised ML techniques such as Logistic Regression(LR), Support Vector Machine(SVM), Decision Tree(DT), K-Nearest Neighbors(KNN), Naive Bayes(NB) and Random Forest(RF) are used to develop predictive models for analyzing the influencing risk factors for the COVID-19 infection.

A. Logistic Regression Algorithm:

Logistic Regression algorithm is a supervised learning algorithm that is used for the classification task in which the association of categorical dependent features against independent features is determined. This learning algorithm is used when the dependent features are dichotomous, which means there would be only two possible classes. They have binary values such as 0 and 1, true or false, positive ornegative, and yes or no[1]. [3]The mathematical equation used to calculate the association between dependent features and independent attributes of the dataset using logistic regression algorithm is given below:

$$i = \log\left(\frac{p}{1-p}\right)$$

where p is the probability of success.

B. Support Vector Machine :

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression.SVMs are designed for binary classification and it do not support classification tasks with more than two classes.This algorithm tries to find a hyperplane with the maximum margin by considering the maximum margin with the closest points known as support vectors.The hyperplanes are used to make predictions.
The equation of the line isy=ax+b, On renaming x with x1 and y with x2, the equation becomes ax1-x2+b = 0. By defining x as (x1,x2) and w=(a,-1),the equation of hyperplane is defined by w.x+b =0. From this,the hypothesis function h is defined as

$$h(x_i) = \begin{cases} +1, w._x + b \geq 0 \\ -1, w_x + b < 0 \end{cases}$$

The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

classified as class -1. The goal of the SVM learning algorithm is to find a hyperplane that could separate the data accurately. From many such hyperplanes, the best one, that is often referred as the optimal hyperplane has to be found.

C. Decision Tree Algorithm :

The easiest and most interpretable supervised learning algorithm is decision trees. Decision tree algorithms can be used for both classification and regression. A decision tree is an efficient algorithm for describing a way to traverse a dataset. The structure of a decision tree can be defined by a root node, which is the most important splitting feature. The internal nodes are tests on an attribute. The data points satisfying the conditions are on one side, and the rest on the other. The leaf nodes belong to the available classes that the dataset represents.

The goal of using a Decision Tree is to create a training model that can be used to predict the class label by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label traversing should be done from the root of the tree. The values of the root attribute are compared with the record's attribute. On the basis of comparison, the branch corresponding to that value is identified and classified.

D. K-Nearest Neighbors Algorithm :

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm trains using all the available data and classifies a new data based on the similarity. KNN algorithm classifiesby understanding the similarity between the new data and available data. The distance metric d, given by the following equation is to find the Euclidean distance through which similarity will be calculated.

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + .. + (x_n - x_n')^2}$$

Finally, the input x is assigned to the class with the largest probability.

E. Naive Bayes Algorithm :

Naive Bayes is a machine learning algorithm, more specifically a classification technique. Naive Bayes is used when the output variable is discrete. Naive Bayes algorithm is based on the Bayes Theorem for calculating probabilities and conditional probabilities. The equation for solving the probability of y, given input features X is

$$P(y \mid X) = \frac{P(X \mid y)P(y)}{P(X)}$$

The goal of Naive Bayes is to choose the class y with the maximum probability.

F. Random Forest Algorithm :

Random Forest is an ensemble learning, which combines multiple machine learning algorithms to obtain better predictive performance. A technique known as bagging is used to create an ensemble of trees where multiple training sets are generated with replacement. In the bagging technique, a data set is divided into **N** samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging in parallel.

RESULTS:

The machine learning process begins with collecting data separately from a variety of resources[2-3]. The next step is data pre-processing in which data-related issues are fixed andinteresting data is being selected by removing invalid data. Sometime, the value of the dataset might be inadequate to make decision. Therefore, machine learning algorithms are assimilatedwith other concepts such as statistics, theory control and probability to analyze

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

data and extract useful and novel knowledge or hidden patterns or information from past experiences[2]. The next step is building thepredictive model using training data followed by model evaluation for improving the model using test data. Figure 1 shows the steps / processinvolved in this work for the development of predictive models.

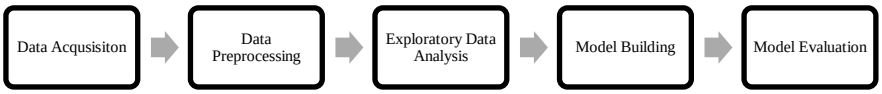

Fig No:1 Steps involved in this work

A. Data Acquisition :

A dataset of positive and negative COVID19 cases provided by WHO (World Health Organization) is used in this study. This dataset was obtained from the official website of kaggle.com, which has 5434 instances or patient records with 19 features who have undergone the viral respiratory diagnosis.The 18 features (breathing problem, fever, sore throat, running nose , asthma, lung disease, heart disease, headache, diabetes, hyper tension, fatigue, gastrointestinal, abroad travel, contact with covid affected persons, attended large gathering, visited to public exposed places, family working in public exposed places, sanitization and wearing mask) except the target classare grouped as symptoms, Chronic Medical Condition, External Contact and Preventive measure.

B. Data Preprocessing:

Data is preprocessed by Data Cleaning and Data Wrangling. Data cleaning identifies and handles missing data and noisy data there by it supports in better prediction. Data wrangling is applied to make the data suitable for further processing. As the dataset used in this work doesn't have any missing data, this step is skipped. Data wrangling is achieved by using Label Encoding method to make the data suitable for further analysis. Figure 2 shows the table of values after performing Label Encoding.

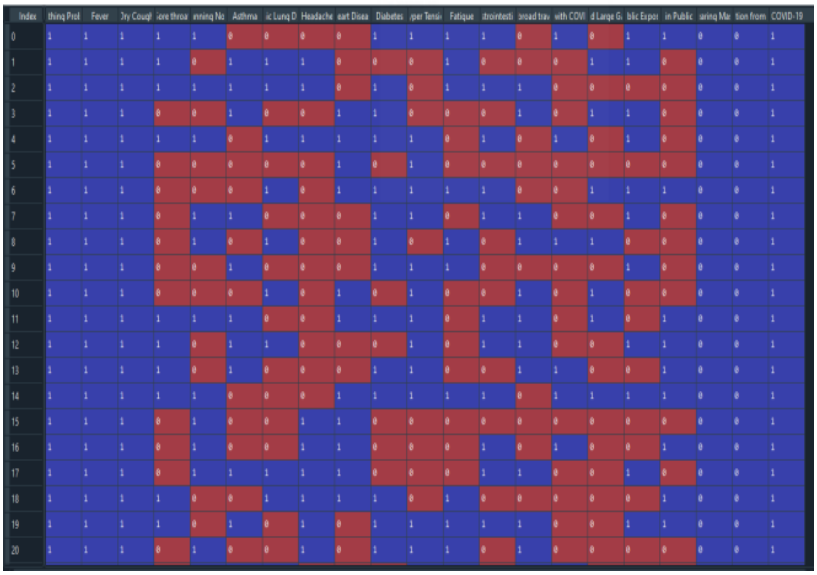

Fig no:2 Dataset after Data Wrangling

C. Exploratory Data Analysis

Dependent features from the dataset are identified and channelized based on the influencing risk factor through exploratory data analysis. Exploratory data analysis is performed on the dataset to support Feature Engineering and Data Reduction. [6]Correlation plays a great role in finding the dependency among the features of the dataset. In this work, the correlation between the dependent and independent features was found to determine the existence of relationship among the features in thethe dataset. Hence during feature engineering, the correlation between the features in the respective groups of influencing risk factors and the target class labels are being calculated to

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

finalize the features to be considered for prediction. Table1 shows the correlation between symptoms and the target class, Table 2shows the correlation between chronic medical condition and target class, Table 3shows the correlation between external contact and target class and Table 4shows the correlation between preventive measure and target class.

Table no:1 Correlation between symptoms and target class

Index	athing Probl	Fever	Dry Cough	Sore throat	Running Nose	COVID-19
Breathing Problem	1	0.0899032	0.159562	0.303768	0.05519	0.443764
Fever	0.0899032	1	0.12758	0.322235	0.0817575	0.352891
Dry Cough	0.159562	0.12758	1	0.213907	-0.0307633	0.464292
Sore throat	0.303768	0.322235	0.213907	1	0.0394502	0.502848
Running Nose	0.05519	0.0817575	-0.0307633	0.0394502	1	-0.00565719
COVID-19	0.443764	0.352891	0.464292	0.502848	-0.00565719	1

Table No: 2 Correlation between chronic medical condition and target class

Index	Asthma	onic Lung Dise	Headache	Heart Disease	Diabetes	Hyper Tension	Fatigue	Gastrointestinal	COVID-19
Asthma	1	-0.0337712	0.0370642	0.076783	-0.0120599	0.0177068	0.00656357	0.101909	0.0899295
Chronic Lung Disease	-0.0337712	1	-0.0504804	-0.0398597	0.0467889	-0.0103314	-0.0476553	-0.0503334	-0.056837
Headache	0.0370642	-0.0504804	1	0.0484711	0.0323899	-0.207489	0.0520353	0.0977783	-0.0277931
Heart Disease	0.076783	-0.0398597	0.0484711	1	-0.0329555	0.0491393	-0.0589252	0.00412097	0.0270721
Diabetes	-0.0120599	0.0467889	0.0323899	-0.0329555	1	0.042543	-0.0439028	0.0406512	0.0406275
Hyper Tension	0.0177068	-0.0103314	-0.207489	0.0491393	0.042543	1	-0.027605	-0.0679723	0.102575
Fatigue	0.00656357	-0.0476553	0.0520353	-0.0589252	-0.0439028	-0.027605	1	0.00935596	-0.0441879
Gastrointestinal	0.101909	-0.0503334	0.0977783	0.00412097	0.0406512	-0.0679723	0.00935596	1	-0.00336686
COVID-19	0.0899295	-0.056837	-0.0277931	0.0270721	0.0406275	0.102575	-0.0441879	-0.00336686	1

Table No: 3 Correlation between external contact and target class

Index	Abroad trave	t with COVID	ed Large Gat	ublic Expose	ıg in Public Ex	COVID-19
Abroad travel	1	0.0802095	0.113399	0.069609	0.143094	0.443875
Contact with COVID Patient	0.0802095	1	0.234649	0.0797997	0.00690923	0.357122
Attended Large Gathering	0.113399	0.234649	1	0.0837954	0.0637757	0.390145
Visited Public Exposed Places	0.069609	0.0797997	0.0837954	1	0.0284861	0.119755
Family working in Public Exposed Places	0.143094	0.00690923	0.0637757	0.0284861	1	0.160208
COVID-19	0.443875	0.357122	0.390145	0.119755	0.160208	1

Table No: 4 Correlation between preventive measure and target class

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

Index	Wearing Masks	anitization from Mark	COVID-19
Wearing Masks	nan	nan	nan
Sanitization from Market	nan	nan	nan
COVID-19	nan	nan	1

Table 4, clearly depicts that there is no correlation between the preventive measure and the target class as per the source dataset. Hence the features that are included under the group preventive measure are considered for Data Reduction.Data reduction is performed to eliminate the non-influencing factor as the outcome of feature engineering. Based on the results of correlation, features base on preventive measures are dropped from further processing. Figure 3 shows the dataset after Data Reduction.

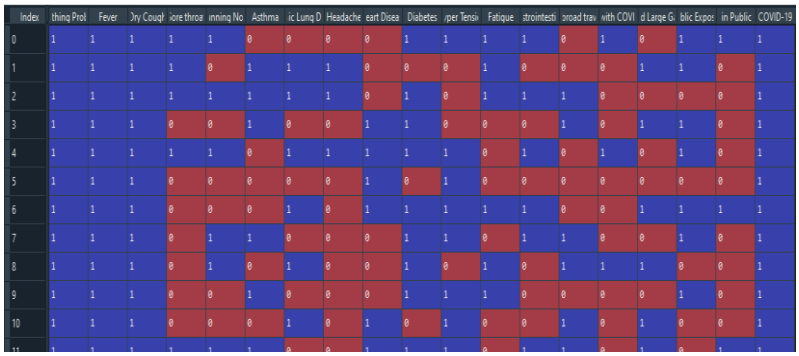

Fig no:3 Dataset after Data Reduction based on Correlation

D. Model Building :

Based on the influencing factors, Predictive models are built using the Machine Learning algorithms such asLogistic Regression, Decision Tree Algorithm, Support Vector Machine, K-Nearest Neighbors Algorithm, Naive Bayes Algorithm And Random Forest Algorithm. 70% of data is considered as training set and remaining 30% is considered as testing set.

E. Model Evaluation :

The models are evaluated using various metrics such as Accuracy, Precision, Recall, RMSE and F1Score.

Table No: 5 Evaluation based on symptoms

Classifier	Accuracy	Precision	Recall	RMSE	F1 Score
LR	91.78	92.13	98.17	28.66	95.05
SVM	92.27	93.03	97.71	27.79	95.31
DT	93.19	92.80	99.23	26.08	95.91
KNN	92.45	93.11	97.86	27.46	95.43
NB	87.92	94.28	90.47	34.75	92.34
RF	93.25	93.01	99.00	25.96	95.91

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

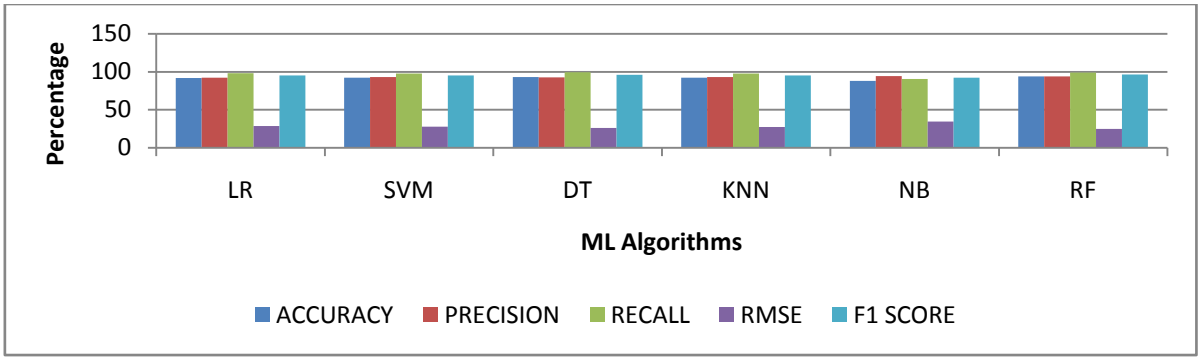

Fig No: 4 Evaluation based on symptoms

Table no: 6 Evaluation based on Chronic medical condition

Classifier	Accuracy	Precision	Recall	RMSE	F1 Score
LR	80.50	80.50	100	44.15	89.19
SVM	80.50	80.50	100	44.15	89.19
DT	86.48	90.14	93.37	36.81	91.73
KNN	85.59	90.10	92.23	37.95	91.11
NB	80.50	80.50	100	44.15	89.19
RF	87.67	89.95	95.46	35.10	92.62

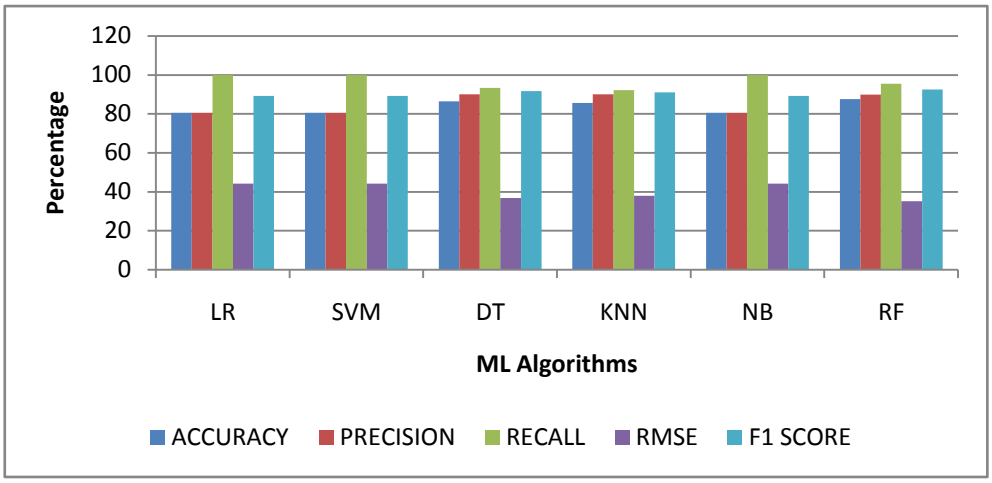

Fig No: 5 Evaluation based on Chronic medical condition

Table no: 7 Evaluation based on External contact

Classifier	Accuracy	Precision	Recall	RMSE	F1 Score
LR	89.88	95.33	91.92	31.80	93.60
SVM	89.88	95.33	91.92	31.80	93.60
DT	89.88	95.33	91.92	31.80	93.60

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

KNN	87.36	88.73	96.57	35.53	92.48
NB	67.88	99.74	60.01	56.84	74.69
RF	89.88	95.73	91.53	31.90	93.58

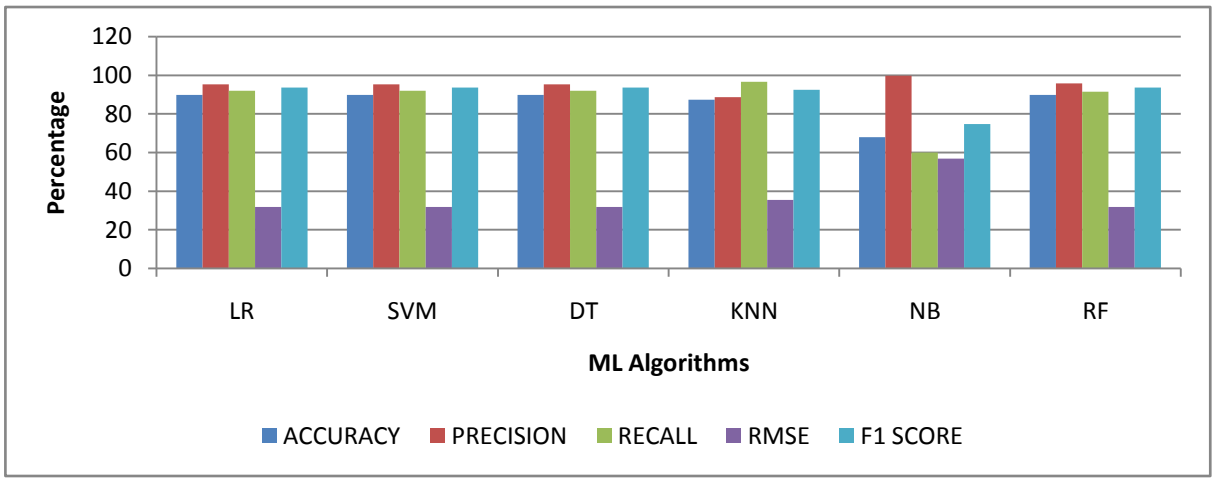

Fig No:6 Evaluation based on External contact

EXPERIMENTAL RESULTS

By analyzing the experimental results in Python, this study proved that the Random Forest algorithm gives higher accuracy with minimum RMSE. As shown in Fig.7, The higher accuracy 93.19% with minimum RMSE value 26.08% is attained when the prediction is based on the symptom. Hence this work concludes that symptom is the dominant influencing risk factor to predict the COVID-19 infection.

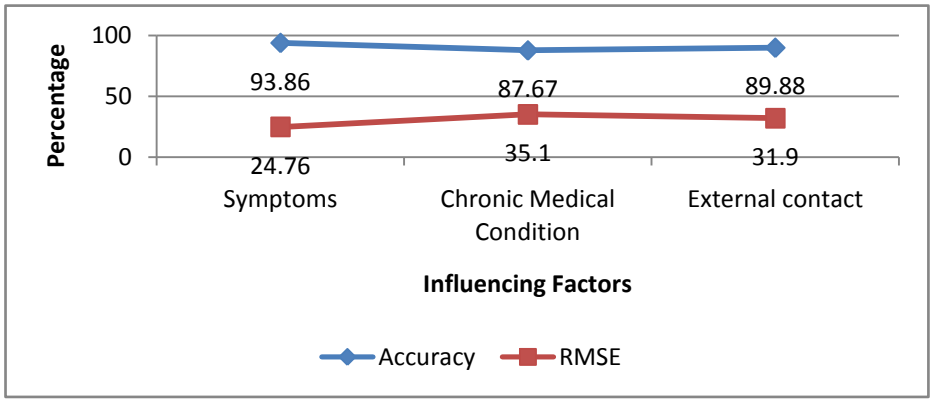

Fig No: 7 Performance Evaluation

CONCLUSION

The novel coronavirus (COVID-19) presents a significant and urgent threat to global health. There is a sharp swift increase in the number of cases to test for coronavirus, which is constantly taking people under its arrest. Various influencing risk factors of that everyone has to be cautious include symptoms, chronic medical conditions and contact with affected persons.ML Techniques are being used in a wide variety of areas such as medicine, engineering, education, manufacturing production, forecast, traffic management and robotics. Hence, in this work, Supervised ML Predictivemodels were developed for the risk factors of COVID-19 infection using the labeled dataset of positive and negative COVID-19 cases. The model developed with Random Forest happened to be the best model among all models developed in terms of accuracy and RMSE values and symptoms are identified as the dominant influencing risk factor in COVID -19 infections. As per the analysis of the experimental results, this work concludes that

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 73-81
Received 15 April 2021; Accepted 05 May 2021.

symptom is the influencing risk factor that needs to be addressed promptly during COVID-19 infection.

CONFLICTS OF INTEREST:

The author have declared no conflicts of interest

REFERENCES

[1] Edison O, Mei UW, Anthony H et al. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. bioRxiv.

[2] Huang GB, Zhu QY, Siew CK (2006). Extreme learning machine: theory and applications. Neurocomputing.

[3] Muhammad L. J, Ebrahem A. Algehyne, Sani Sharif Usman · Abdulkadir Ahmad,Chinmay Chakraborty, Mohammed I. A (2020). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset, SN Computer Science, © Springer Nature Singapore Pte Ltd.

[4] Paul S (2006). Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng.

[5] Rajan Gupta, Gaurav Pandey, Poonam Chaudhary, Saibal K. Pal (2020). Machine Learning Models for Government to Predict COVID-19 Outbreak, Digital Government: Research and Practice.6.Sujath R, Jyotir Moy Chatterjee, Aboul Ella Hassanien (2020). A machine learning forecasting model for COVID-19 pandemic in India, Stochastic Environmental Research and Risk Assessment, Springer.

[6] VikasChaurasia, Saurabh Pal (2020). Application of machine learning time series analysis for prediction COVID-19 pandemic, Research on Biomedical Engineering, Springer

[7] S.UshaKiruthika, S.Kanaga Suba Raja , Jaichandran R, Priyadharshini C, 'Detection and Classification of Paddy Crop Disease using Deep Learning Techniques', International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume 8, Issue - 3,2019

[8] S.Jayachitra, A.Prasanth, 'Multi-Feature Analysis for Automated Brain Stroke Classification Using Weighted Gaussian Naïve Baye's Classifier', Journal of Circuits, Systems, and Computers, 2021

[9] Murugan, S., Jeyalaksshmi, S., Mahalakshmi, B., Suseendran, G., Jabeen, T. N., & Manikandan, R. (2020). Comparison of ACO and PSO algorithm using energy consumption and load balancing in emerging MANET and VANET infrastructure. Journal of Critical Reviews, 7(9), 2020.