International Journal of Safety and Security Engineering
Vol. 14, No. 6, December, 2024, pp. 1743-1750
Journal homepage: http://iieta.org/journals/ijsse

Predicting the Risk of COVID-19 Patients Using Machine Learning Techniques for Simplification of the Task of Life Insurance Companies

Prasanta Baruah, Pankaj Pratap Singh*, Jwngdao Daimary

Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, Assam 783370, India

Corresponding Author Email: pankajp.singh@cit.ac.in

https://doi.org/10.18280/ijsse.140609

Received: 6 December 2023
Revised: 26 August 2024
Accepted: 9 September 2024
Available online: 31 December 2024

Keywords:
risk prediction, life insurance, COVID-19, machine learning (ML), feature reduction, ROC curve

ABSTRACT

Evaluation of risk is the most important task of the industries of life insurance businesses to place the applicants in different categories based on their risk assessment. Underwriters estimate the pricing of the policies for the applicants based on the risk of the categories. For better assessment of risks on huge dataset of applicants, a predictive model is a suitable tool. The risk of an applicant suffering from COVID-19 is higher than a normal applicant. Our research is aimed at finding the risk of COVID-19 applicants by predicting the mortality rate of the applicants suffered from it. It helps to predict the accurately risk patients with the better approach and can also be used for insurance companies in materializing the health insurance. A dataset of 10000 patients was considered for a predictive model to assess the risks of the COVID-19 patients. Some prominent features are selected from the original dataset. The mortality prediction of the patients infected by COVID-19 is evaluated using predictive algorithms of machine learning on the dataset. Findings revealed that the model of logistic regression produced highest accuracy in comparison to the other algorithms. The novelty of this research lies in its integration of data sources and advanced machine learning techniques to enhance prediction accuracy, which is crucial for improving health outcomes and optimizing healthcare interventions during the pandemic.

1. INTRODUCTION

Availability of large dataset has made an exceptional impact in the revolution of life insurance business. Various machines learning algorithms can be implemented on big data and thereby making the process of data analysis more efficient day by day. Moreover, big data has rejuvenated the life insurance industries enabling them to enhance the prediction in risk assessment, underwriting process, fraud analysis, and insurance coverage [1, 2]. Premium calculation for a customer in a life insurance business is the job of underwriters. The underwriters assess the factors that contribute to the risk of the applicants based on various available data of the applicants such as medical history, family history, food habits, life style, locality etc. Applicants with similar level of risks are placed in the same category. After categorization of the customers into different levels based on similarity of risk, pricing of the premiums for each level is determined. Normally with conventional methods, underwriters need very good amount of time, may be several weeks to several months, which is not only a big disappointment for the applicants but also major barrier for the growth of the industry [3].

Due to the lack of efficiency of manual evaluation, conventional underwriting processes may not provide accurate risk prediction of or the applicants. Thus, conventional methods of underwriting process may affect the accuracy of premium calculation and thereby may badly affect the profit of the business as well. Also, the cost involves in a conventional predictive model is very high as the costs incurred in different medical tests need to be borne by the company itself. Due to these drawbacks associated with the conventions underwriting methods, alternative solution is quite important not only to make the business profitable, but also providing the business a competitive advantage as well as customer satisfaction [2]. Big data has provided a platform to develop predictive analytic models which not only decrease the cost associated with the underwriting process, also improves the efficiency in pricing the policies more appropriately. This is due to the fact that predictive models can identify the risk levels of applicants and according to the risk levels, pricing can be made appropriately [4]. Big data has revolutionized risk assessment by enabling more accurate and granular analyses of individual risks. In underwriting, it speeds up decision-making and improves accuracy with comprehensive data insights. For fraud analysis, it enhances detection through advanced pattern recognition and anomaly detection. In insurance coverage, big data allows for more personalized and dynamic pricing, leading to tailored policies and better customer service. Big data has notably impacted the insurance industry as Risk Assessment like 'Progressive's Snapshot uses telematics data to monitor driving behavior,

allowing for personalized insurance premiums based on actual driving habits rather than broad demographics.

Extensive researches are going in the implementation of machine learning, artificial intelligence algorithms to enhance the predicting the risk level, fraud finding to minimize the loss or maximize the profit of the life insurance businesses [5-7]. Predictive analytic models play a crucial rule to predict the mortality rates of applicants of life insurance business. The use of these models will enhance the decisions in the underwriting process, decreasing the amount of effort as well as time taken in processing the applicants [8]. Underwriters classify the applicants based on the risk level. They do analysis of the risk profiles of individual applicants considering various factors such as family history, health condition, food habits, occupation, living conditions and many other factors [9]. Then risks of the applicants are evaluated and the applicants are grouped according to their risk level. Thus, the application of predictive models in life insurance businesses not only be economical in use also make the application processing faster as well as will provide confidence to the underwriters in making more accurate decisions [2, 10]. Certainly, predictive models will address many of the problems such as higher cost, greater application processing time, inconvenience decision making associated with the conventional methods [11].

Predictive analytics models enhance underwriting by automating risk assessments, thereby reducing manual errors and costs. They improve efficiency through faster and more accurate evaluations, leading to quicker decisions. Additionally, they offer a competitive edge by enabling precise risk pricing and tailored offerings, which attract and retain customers. Predictive analytic machine learning models decrease costs by automating risk assessments and reducing manual errors. They improve efficiency through faster and more accurate evaluations. Additionally, they provide a competitive advantage by enabling precise risk pricing and personalized offerings, which attract and retain customers more effectively. I have incorporated it as per the suggestions.

The interest of doing this work is to build models based on the available historical data of COVID-19 patients and endorse the most suitable model to evaluate the risk levels of applicants. This is the key objective of this paper is to further proposed a detailed analysis of the predictive models on COVID-19 patient data. Ultimately goal of the model is to help the underwriters to make more appropriate decisions in pricing the policies for applicants in life insurance business so that the organisation possesses competitive advantage over its competitors. This will not only satisfy the customers but also provide growth in profitability in the business.

2. LITERATURE REVIEW

The objective of a life insurance company is to persuade the applicants and make them their customers. Every life insurance company tries to convince the applicants with their policies. But, to be successful in enrolling customers, suitable methodology having competitive advantage is very much important. Mustika et al. [12] developed a machine learning model based on life insurance data for assessing the level of risk of the applicants. The accuracy of the prediction was very satisfactory. Jain et al. [13] suggested a technique for assessing the risks of applicants using artificial neural network and gradient boosting algorithm on a dataset of 128 features to carry out the research work. Batty et al. [14] applied several machine learning algorithms on a dataset to identify the range of the risks for the applicants. Dwivedi et al. [15] used various supervised learning algorithms to predict the risk level of the applicants. Among the algorithms, random forest result was found to be the most accurate.

Prabhu et al. [16] proposed a predictive model based on neural network for risk prediction. Franch-Pardo et al. [17] revised several articles which reveal that geospatial and spatial-statistical analysis tools were used for identifying the highly affected areas of coronavirus (COVID-19) pandemic around the world. Some Modeling was done for handling threats in the Context of COVID-19 [18]. Saran et al. [19] mentioned the use of geospatial technologies such as geographic information system (GIS) for the surveillance of infectious diseases in the domain of public health. A detailed review is done on risk prediction in the field of insurance [20]. Kamel Boulos and Geraghty [21] discussed about GIS mapping dashboards for tracking coronavirus epidemic updates around the world in near-real-time. Diabetic's patient data is also an important risk factor for the analysis and classification using the machine learning techniques [22]. GIS and machine learning are utilized as combined framework for predicting the risk in health insurance sector [23]. An assessment is indeed for Health Risk due to the air pollution which raised high in the Air Conditioning Manufacturing Plants [24].

3. DATASET DESCRIPTION

Data is collected from online data databases to carry out this research. The collected data are analyzed and pre-processed to eliminate the unnecessary data including redundant ones. Then, the relationships between the variables of the dataset are established before applying machine learning algorithms for achieving our tasks. The dataset consisting of 10,000 records of COVID-19 patients of both male and female is collected from data library [25]. It contains 112 attributes for each patient, including the demographic and physiological data. This dataset is pre-processed and then dimensionality reduction is carried to keep only 18 prominent features into it. The description about the features in the dataset is given in the Table 1.

A dataset may contain many unnecessary values such as noisy data or outliers [11]. Therefore, data pre-processing utilizes the data cleaning step which is essential to remove noisy or inconsistent data to achieve the target dataset. In this step, the data is collected from kaggle data library and then the dataset is cleaned by removing the inconsistent data from the dataset.

After pre-processing of data, first and foremost task is to identify the prominent features which will be suitable for the classification of risks of the candidates. Selection of prominent features is important for development of an efficient model. The most prominent features are selected using feature selection methods [11, 24]. Several feature ranking algorithms are used for this purpose. The ranking of the features, help to which features should be selected and which should be rejected. Out of 112 features of the dataset only 18 features are selected for the construction of the models. Following models are used on the dataset such as Decision Tree (DT), K-Nearest Neighbour (KNN), Naive

Bayes (NB), Logistic Regression (LR) and Random Forest Classifiers (RFC) and finally the model showing the most accurate result is considered.

Table 1. COVID-19 patient's dataset with attributes, types and description

Attributes	Type	Description
Gender	Categorical	Male or Female
Intubated	Categorical	insert a tube into a person's body
Pneumonia	Categorical	Pneumonia is an inflammatory condition of the lung
Age	Numerical	Age of person
Pregnancy	Yes/No	Is Pregnant or Not
Diabetes	Yes/No	Is Diabetic or Not
COPD	Yes/No	This disease, is related to lung disease that blocks airflow
Asthma	Yes/No	A condition in which the airways narrow and swells
Immuno-suppression	Categorical	The partial or complete suppression of the immune response of an individual
Hypertension	Categorical	High pressure in the arteries
Other Diseases	Categorical	Non Major Diseases
Cardiovascular	Categorical	Cardiovascular is related to the heart or blood vessels.
Obesity	Categorical	A condition characterised by abnormal or excessive fat accumulation
Renal chronic	Categorical	Related kidney disease.
Tobacco	Categorical	Habit of a person taking tobacco
Contact Other Person	Yes/No	Contacting the COVID positive person.
Covid Result	+ve/-ve	Result of COVID-19 Test of a Person
ICU	Categorical	Whether a Patient is needed to keep in ICU (Intensive-Care-Unit)

4. METHODOLOGY

A diagram for the models development is shown in the Figure 1. After post validation process, the model predicting the most accurate result is being considered as the result of mortality prediction for the COVID-19 patients.

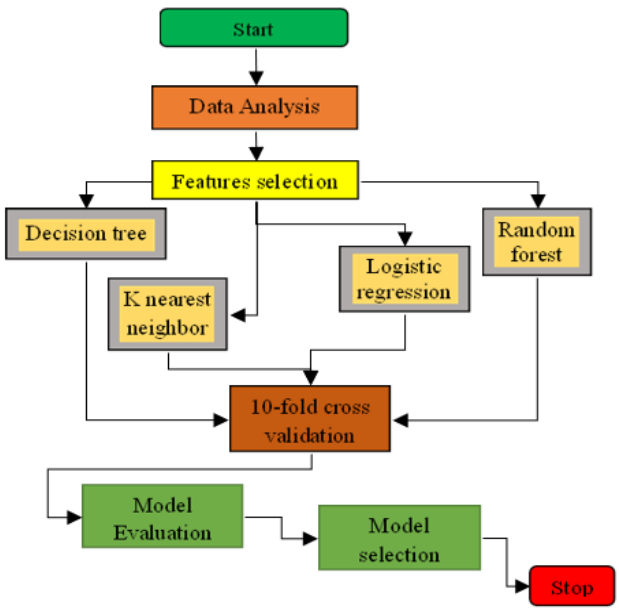

Figure 1. A proposed framework for risk prediction

In this research, several supervised machine learning techniques are selected on the basis of their accuracy rate, execution time, confusion matrix and ability to work with large data in the existing research works carried out by various researchers. The brief description of the algorithms considered for risk prediction of COVID-19 patients given below. DT algorithm is tuned with hyperparameters such as entropy and Gini index for better class prediction. RF algorithm determines the number of trees with the feature subset for better model training and results.

4.1 Decision Tree (DT)

It is mostly considered for the data classification purposes [26]. Internal nodes represent the features in a dataset, branches are used to represent the decision rules and leaf nodes are used to represent the outputs.

4.1.1 Features selection mechanism

For implementing a Decision Tree, selecting the best attribute for nodes of the tree is important. The attribute selection measure (ASM) can be used to select the most promising attribute from the dataset. The three popular ASM techniques are: Entropy, Gini Index and Information Gain.

4.1.2 Entropy

The uncertainty of a random variable is measured by entropy. Entropy describes the uncleanness of an arbitrary group of instances or the degree of volatility of the data in a dataset. The greater the entropy means higher is the disorder and vice versa. The proper definition of entropy is obtained from Shannon's entropy for calculation of the information gain:

$$\text{Entropy}(x) = -\sum (P(x = k) * \log_2 P(x = k)) \quad (1)$$

here, P(x = k) means the probability of target feature (x) for a specific value, k. Since, logarithm of fractions gives a negative value, therefore, hence a '-' sign is used in entropy formula to negate these negative values.

4.1.3 Gini index

It is also known as Gini impurity is the probability that how a randomly selected feature would classify the instances in a dataset incorrectly. This feature would be suitable to split the dataset can be identified through the Gini index. Which feature should be in the root node, which should be in the internal nodes as well as in the leaf nodes are known from Gini index. A feature with lower Gini index should be preferred. It is calculated as follows:

$$\text{Gini Index} = 1 - \sum (P(x=k))^2 \quad (2)$$

where, the feature k is classified into a distinct class with a probability p.

4.1.4 Information gain

Entropy plays an important part in calculating the information gain. The amount of change in entropy is known as the information gain. Information gain gives the most suitable feature with the maximum information about a class for the purpose of classification.

$$\text{Information Gain (feature)} = \text{Entropy(Dataset)} - \text{Entropy(feature)} \quad (3)$$

where, the root node of the tree is splited on the feature whose information gain is the highest.

4.2 K-Nearest Neighbor (KNN)

The objective of KNN is to placing a new data into a group to which the new one has the similarity [26]. Thus KNN algorithm classifies any object based on the similarity with the existing ones.

4.3 Naive Bayes (NB)

NB algorithm is mostly suitable for classification problems [26]. These supervised algorithms are very simple and most effective as well as suitable for building predictive models for making predictions quickly. It does predictions based on the probability of a feature. Naive Bayes model is very useful for very large data sets. The posterior probability of Bayes theorem $P(c_i|f)$ is calculated from $P(c_i)$, $P(f)$ and $P(f|c_i)$ as follows:

$$P(c_i|f) = P(f|c_i)*P(c_i)/ P(f) \quad (4)$$

Thus,

$$P(c_i/F)=[P(f_1/c_i)*P(f_2/c_i)*\ldots\ldots\ldots.. p(f_n/c_i)]*P(c_i)/[P(f_1)*P(f_2)*\ldots\ldots\ldots P(f_n)] \quad (5)$$

where, $P(f|c_i)$ is the likelihood probability of the class defined by the predictor. Given the predictor $P(f|c_i)$, the posterior probability $P(c_i|f)$ of the class (target) can be computed. The prior probability of the predictor is $P(f)$, $P(c_i)$ is the prior probability of the class and $f = \{f_1,f_2,\ldots,f_n\}$ is the feature set. Here, the target class c is either 'serious' or 'mild'.

4.4 Logistic Regression (LR)

It is most popular machine learning algorithm [26]. It is suitable for predicting the outcome of a dependent variable of categorical nature, in relation to a set of independent variables. Therefore, the result will be a value of categorical or discrete in nature which is either yes or no, 0 or 1, true or false, etc. But instead of assigning the exact value as 0 and 1, it assigns the probabilistic values which lie between 0 and 1. The Logistic regression equation can be represented as follows:

$$y = b_0 + b_1x_1 + b_3x_3 + ... + b_nx_n. \quad (6)$$

The value of y lies between 0 and 1 only, so it is necessary to divide the above equation by (1-y). So, y/(1 – y) will be 0 if y is 0 and will be infinity if y is 1. Thus, the equation will become:

$$\log [\, y /(1 - y)] = b_0 + b_1x_1 + b_3x_3 + ... + b_nx_n. \quad (7)$$

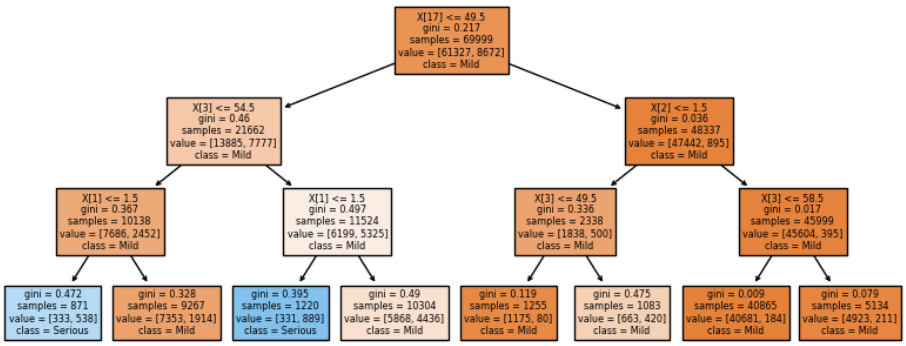

Figure 2. Decision Tree classifying the risk of the patients as mild or serious

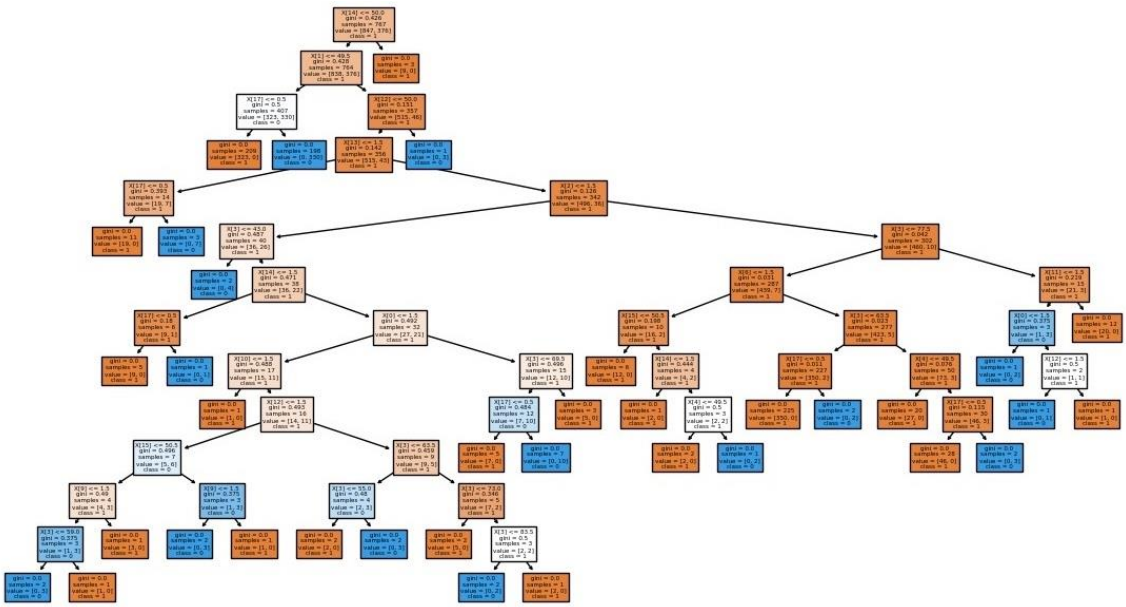

Figure 3. RF forest tree classifying the risk of the patients using the Gini index (sub-tree view1)

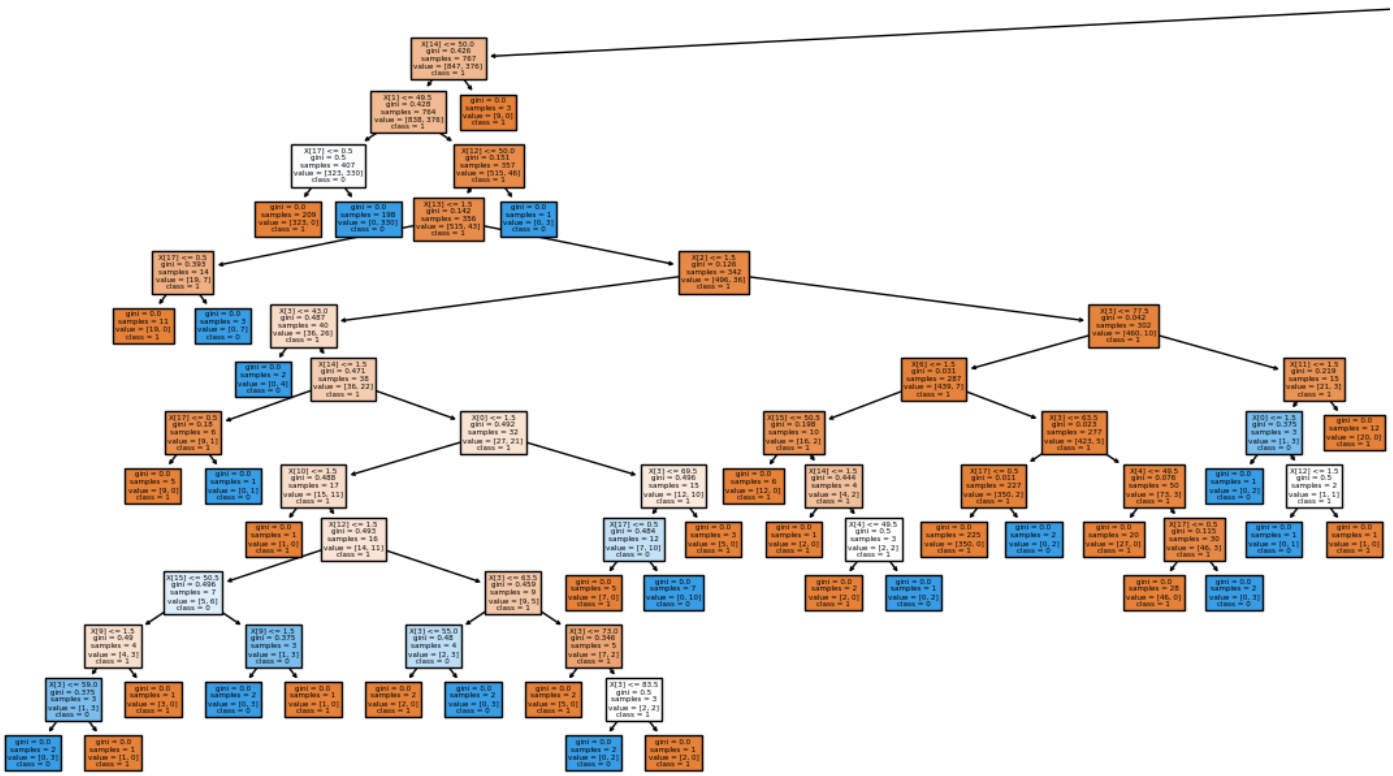

Figure 4. RF forest tree classifying the risk of the patients using the Gini index (sub-tree view1)

Figure 2 shows the decision of the tree while classifying the risk of the COVID-19 patients as serious or mild based on the dataset.

4.5 Random Forest (RF)

Another popular supervised machine learning algorithm RF is used in this data analysis [27]. The concept of ensemble learning is the basis for random forest where to solve a complex problem. Multiple classifiers are used through which the performance of the model can be improved. RF classifier contains many number of Decision Trees (see Figures 3 and 4). Each tree makes decision on various subsets of the given dataset and considers the average of all the decisions as overall decision on the dataset. RF takes the prediction of all the trees and considers the prediction made by the majority of trees as the final output.

5. EXPERIMENTAL SETUP AND RESULTS

In this research work, a dataset of 10,000 data of confirmed COVID-19 patients of both male and female is used. The dataset contains 18 attributes about the patients after dimensionality reduction, including demographic and physiological data. Our experiments were carried out using Numpy and Pandas packages as well as Scikit-learn (Sklearn) library of Python programming language. Pre-processing is done after collecting the dataset. The purpose of pre-processing is to remove the noises and missing values available therein. After pre-processing, the dataset is categorized into two distinct parts one for training and the other for testing purposes. To split the dataset we have used the SKlearn library of Python language. Initially, we have separated the target variable from the features in the dataset. 10-fold cross validation process is used for all the models and approximately, in the ratio of 70:30 the dataset is divided into training and testing sets. Lastly, the machine is trained with the training dataset using different algorithms and thereafter the test dataset is used to measure the accuracy of the algorithms. The accuracy of the predictions of DT, KNN, NB, LR, RFC algorithms with the test dataset is calculated in a confusion matrix.

Various evaluation metrics such as confusion matrix, accuracy rate as well receiver operating characteristic (ROC) curve are considered for assessing the performances predicted by the algorithms. These metrics are discussed in the following sections.

5.1 Confusion matrix

The outcomes of the predictions of different machine learning algorithms can be summarized in a table as a matrix. This matrix is called the confusion matrix consists of two dimensions, one is actual and the other one is predicted. This matrix is also used to measure the accuracy of the predictions of machine learning algorithms for a given test dataset. In this case, the test dataset considered consists of 3032 instances. The details of a confusion matrix are as follows:

True Positive (TP). TP represents the value of number of positive cases which are predicted as positive on the given test dataset.

False Positive (FP). FP is the value of number of negatives which are falsely predicted as positives.

True Negative (TN). TN is the value of number of negative cases in the test dataset which are predicted as negative.

False Negative (FN). FN is the number of positives which are falsely predicted as negative.

5.2 Accuracy rate

Accuracy rate means the capability of any model to predict both the true positive and true negative cases from all the predictions [26]. That is, it represents the ratio of the sum of

true positive and true negative predictions to the all the predictions. Using SKlearn library in python, the accuracy of the algorithms has been measured. Thus, the accuracy rate of prediction of an algorithm is (TP+TN)/(TP+FN+TN+FP). The accuracy rate of the algorithms for predicting the test dataset are shown in the following confusion matrix. The accuracy of the algorithms obtained after 10-fold cross-validation.

The confusion matrix results show a comparative analysis among all five ML algorithms, which are shown in Table 2. RF is able to predict well for COVID-19 patients on the basis of the selected attributes. In Table 3, the accuracy of all the models is achieving a good level, and KNN is slightly lacking compared to others. LR method is outperforming compared to other methods and signifies better prediction of mortality rate. RF method shows the highest recall value, which signifies accurately identifying the patients in the prediction.

Table 2. Comparison of DT, KNN, NB, LR and RF with confusion matrix

Algorithms	True Positive	True Negative	False Positive	False Negative
Decision Tree	2516	163	227	126
K-Nearest Neighbors	2462	158	232	180
Naive Bayes	2570	136	95	231
Logistic Regression	2575	141	91	225
Random Forest	2605	69	274	84

Table 3. Accuracy of mortality prediction of COVID-19 patients with various ML algorithms

Algorithms	Accuracy	Precision	Recall	F1-Score
Decision Tree	88.36 %	0.9172	0.9523	0.9344
K-Nearest Neighbors	86.41 %	0.9138	0.9318	0.9227
Naive Bayes	89.25 %	0.9643	0.9175	0.9403
Logistic Regression	89.58 %	0.9658	0.9196	0.9421
Random Forest	88.19 %	0.9048	0.9687	0.9145

5.3 Roc curve

This ROC curve is used to show the performance of the classification models at all classification thresholds [28]. This curve is plotted in terms of two parameters which are true positive rate and false positive rate. That is the curve is plotted as rate of true positive against the rate of true negative cases.

TP rate (TPR) and FP rate (FPR) are defined as TP/(TP+FN) and FP/(FP+TN) respectively. The ROC curve for the five algorithms tested with the data set is shown in Figure 5. Although the results of the confusion matrix show that the accuracy of prediction of the Decision Tree, naïve Bayes, KNN, random forest, and logistic regression is close to each other, logistic regression still has slightly better accuracy, which is 89.58%. However, the ROC curve shows that logistic regression is the clear winner among the five algorithms.

These metrics are helpful to evaluate the performance of models. The ROC curve shows trade-offs between sensitivity (TP) and specificity (FP). Area under the curve (AUC) provides a single metric to summarize the ROC curve performance. It also signifies the ability of the model to distinguish the classes. A threshold can be selected using a ROC curve for a classifier, which maximizes the value of true positives by minimizing the value of false positives.

However, the optimal threshold may be different for different algorithms. The objective of the ROC curve is to measure the performance of the algorithms over its entire operating range. The area under a curve is known as AUC and is a measure of the area in an ROC, which can be used to compare the performance of the algorithms under a single threshold. The value of AUC ranges between 0.0 to 1.0. If the value of the AUC of a model is 1.0, then 100% prediction of the model is correct, whereas if AUC is 0.0, then 100% prediction of the model is wrong. Figure 5 depicts that the value of AUC is 0.921 for the logistic regression as the highest in the ROC. This suggests that the logistic regression classifier performance is best among the five classifiers.

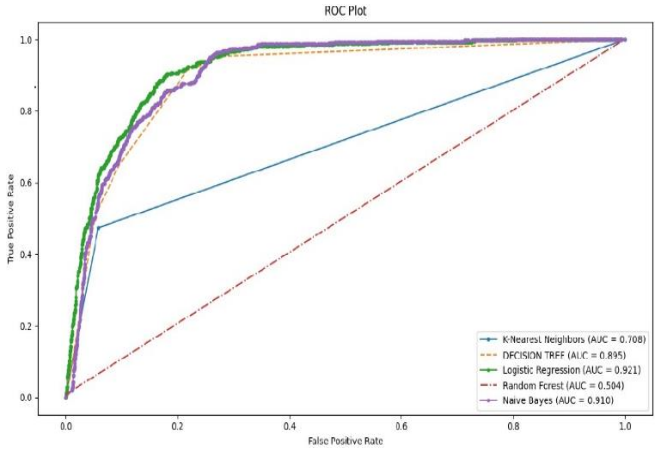

Figure 5. Comparison of the performance of the five algorithms through the ROC curves

6. CONCLUSIONS

In this study, five different machine learning algorithms are implemented on a dataset consisting of 10000 COVID-19 cases throughout the world to predict the risk of mortality of the patients. Various metrics were used to evaluate the prediction of the developed models. The accuracy of prediction of every model is very high, which shows the efficacy of the models. We emphasized pre-existing symptoms and conditions apart from demographic and physiological data of the COVID-19 patients in this study. It is seen that the logistic regression-based prediction accuracy is better among all the ML algorithms, which is 89.58%. Moreover, the ROC curve shows that the logistic regression achieves correct prediction with an AUC of 0.912, which also reflects a good performance.

This paper proposed, based on the exhaustive analysis, that a predictive model is able to identify COVID-19 patients with high mortality risk. This work will not only be helpful for underwriters to determine the risk of COVID-19 patients for the calculation of premium optimally. Moreover, it will be useful for the healthcare systems to provide better medical facilities to risky patients without any delay. All five algorithms were able to produce good accuracy in the risk prediction of mortality of the COVID-19 patients in the dataset. However, some additional features can also be considered for further analysis related to this work, even though it may increase the complexity of the model.

REFERENCES

[1] Fang, K., Jiang, Y., Song, M. (2016). Customer profitability forecasting using big data analytics: A case study of the insurance industry. Computers & Industrial Engineering, 101: 554-564. https://doi.org/10.1016/j.cie.2016.09.011

[2] Joly, Y., Burton, H., Irani, Z., Knoppers, B., Feze, I., Dent, T., Pashayan, N., Chowdhury, S., Foulkes, W., Hall, A., Hamet, P., Kirwan, N., Macdonald, A., Simard, J., Hoyweghen, I. (2014). Life insurance: Genomics stratification and risk classification. European Journal of Human Genetics, 22(5): 575-579. https://doi.org/10.1038/ejhg.2013.228

[3] Bell, M. (2016). Is analytics the underwriting we know? https://insurance-journal.ca/article/is-analytics-changing-theunderwriting-we-know, accessed on Dec. 6, 2023.

[4] Sivarajah, U., Kamal, M., Irani, Z., Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. Journal of Business Research, 70: 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[5] Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1): 58-75. https://doi.org/10.1016/j.jfds.2016.03.001

[6] Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., Arab, M. (2016). Improving fraud and abuse detection in general physician claims: A data mining study. International Journal of Health Policy and Management, 5(3): 165-172. https://doi.org/10.15171%2Fijhpm.2015.196

[7] Goleiji, L., Tarokh, M. (2015). Identification of influential features and fraud detection in the Insurance Industry using the data mining techniques (Case study: automobile's body insurance). Majlesi Journal of Multimedia Processing, 4(3): 1-5.

[8] Bhalla, A. (2012). Enhancement in predictive model for insurance underwriting. International Journal of Computer Science & Engineering Technology, 3(5): 160-165.

[9] Cummins, J., Smith, B., Vance, R., Vanderhel, J. (2013). Risk Classification in Life Insurance. 1st ed. Springer, Dordrecht, New York.

[10] Maier, M., Carlotto, H., Sanchez, F., Balogun, S., Merritt, S. (2019). Transforming underwriting in the life insurance industry. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 9373-9380. https://doi.org/10.1609/aaai.v33i01.33019373

[11] Noorhannah, B., Jayabalan, M. (2019). Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4: 145-154. https://doi.org/10.1007/s40747-018-0072-1

[12] Mustika, W.F., Murfi, H., Widyaningsih, Y. (2019). Analysis accuracy of XGboost model for multiclass classification - A case study of applicant level risk prediction for life insurance. In Proceedings of 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, pp. 71-77. https://doi.org/10.1109/ICSITech46713.2019.8987474

[13] Jain, S., Ramin, M., Byron, C.W. (2019). An analysis of attention over clinical notes for predictive tasks. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, pp. 15-21. https://doi.org/10.18653/v1/W19-1902

[14] Batty, M., Tripathi, A., Kroll, A., Sheng, Peter, Wu, C., Moore, D., Stehno, C., Lau, L., Guszcza, J., Katcher, M. (2010). Predictive modeling for life insurance. Project Report, Deloitte Consulting LLP.

[15] Dwivedi, S., Mishra, A., Gupta, A. (2020). Risk prediction assessment in life insurance company through dimensionality reduction method. International Journal of Scientific & Technology Research, 9(1): 1528-1532.

[16] Prabhu, T., Darshana, J., Dharani, K.M., Hansaa, N.M. (2019). Health risk prediction by machine learning over data analytics. International Research Journal of Engineering and Tesinchnology, 6: 606-611.

[17] Franch-Pardo, I.M., Napoletano, B., Rosete-Verges, F., Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of the Total Environment, 739: 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

[18] Kryshtanovych, S., Lyubomudrova, N., Tymofeev, S., Shmygel, O., Komisarenko, A. (2022). Modeling ways of counteraction to external threats to corporate security of engineering enterprises in the context of COVID-19. International Journal of Safety and Security Engineering, 12(2): 217-222. https://doi.org/10.18280/ijsse.120210

[19] Saran, S., Singh, P., Kumar, V., Chauhan, P. (2020). Review of geospatial technology for infectious disease surveillance: Use case on COVID-19. Journal of the Indian Society of Remote Sensing, 48(8): 1121-1138. https://doi.org/10.1007/s12524-020-01140-5

[20] Baruah, P., Singh, P.P. (2023). Risk prediction in life insurance industry using machine learning techniques—A review. In International Conference on Advances in IoT and Security with AI, pp. 323-332. https://doi.org/10.1007/978-981-99-5085-0_31

[21] Kamel Boulos, M.N., Geraghty, E.M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. International Journal of Health Geographics, 19(1): 1-12. https://doi.org/10.1186/s12942-020-00202-8

[22] Singh, P.P., Das, B., Poddar, U., Choudhury D.R., Prasad, S. (2017). Classification of diabetic's patient data using machine learning techniques. In: Perez, G., Tiwari, S., Trivedi, M., Mishra, K. (eds.) Ambient Communications and Computer Systems. Advances in Intelligent Systems and Computing, Springer Berlin Heidelberg Singapore, pp. 427-436. https://doi.org/10.1007/978-981-10-7386-1_37

[23] Baruah, P., Singh, P.P., Ojah, S.K. (2023). A novel framework for risk prediction in the health insurance sector using GIS and machine learning. International Journal of Advanced Computer Science and Applications, 14(12): 469-476. https://doi.org/10.14569/ijacsa.2023.0141249

[24] Hussien, A.A., Alzboon, K.K., Matalqa, W., AlEssa, A.H.M. (2023). Health risk assessment due to indoor air pollution in air conditioning manufacturing plants. International Journal of Safety and Security Engineering,

13(6): 1083-1090. https://doi.org/10.18280/ijsse.130611
[25] Open Data General Directorate of Epidemiology. https://www.gob.mx/salud/documentos/datos-abiertos-152127/, accessed on Dec. 6, 2023.
[26] Pourhomayoun, M., Shakibi, M. (2020). Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. Smart Health, 20: 100178. https://doi.org/10.1016/j.smhl.2020.100178
[27] Alam, M.Z., Rahman, M.S., Rahman, M.S. (2019). A Random forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked, 15: 100180. https://doi.org/10.1016/j.imu.2019.100180
[28] Brownlee, J. (2020). Classification accuracy is not enough: More performance measures you can use. https://machinelearningmastery.com/classification-accuracy is-not-enough-more-performance-measures-you-can-use/, accessed on Dec. 6, 2023.