Prediction of COVID-19 Possibilities using KNN Classification Algorithm

Prasannavenkatesan Theerthagiri (vprasann@gitam.edu)
GITAM University Bengaluru https://orcid.org/0000-0003-3420-598X

I.Jeena Jacob
GITAM University Bengaluru

A.Usha Ruby
GITAM University Bengaluru

Y.Vamsidhar
GITAM University Bengaluru

Research Article

Keywords: COVID-19, Prediction, Classification, Machine learning algorithms, KNN

Posted Date: November 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-70985/v2

Version of Record: A version of this preprint was published on March 30th, 2021. See the published version at https://doi.org/10.31782/IJCRR.2021.SP173.

Prediction of COVID-19 Possibilities using KNN Classification Algorithm

Abstract: This paper studies the different machine learning classification algorithms to predict the COVID-19 recovered and deceased cases. The k-fold cross-validation resampling technique is used to validate the prediction model. The prediction scores of each algorithm are evaluated with performance metrics such as prediction accuracy, precision, recall, mean square error, confusion matrix, and kappa score. For the preprocessed dataset, the k-nearest neighbor (KNN) classification algorithm produces 80.4 % of predication accuracy and 1.5 to 3.3 % of improved accuracy over other algorithms. The KNN algorithm predicts 92 % (true positive rate) of the deceased cases correctly, with 0.077 % of misclassification. Further, the KNN algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. Also, it produces the receiver operator characteristic curve with an output value of 82 %. Based on the prediction results of various machine learning classification algorithms on the COVID-19 dataset, this paper shows that the KNN algorithm predicts COVID-19 possibilities well for the smaller (730 records) dataset than other algorithms.

Keywords: COVID-19, Prediction, Classification, Machine learning algorithms, KNN.

1. Introduction

Covid-19 [1-3] a disease that was caused due to a virus called coronavirus. It became a global epidemic disease, according to the World Health Organisation (WHO). It was started at the Wuhan of China at the end of 2019. The symptoms of this disease at an early stage are cough, fever, fatigue, and myalgias [1]. Later the patients suffer from heart damages, respiratory problems, and secondary infection situations. Spreading of COVID-19 happens very fast because it spreads through contact, contaminated surfaces, and infected fluids. When the condition of the patient becomes worse with respiratory issues, the patient needs to be treated in an intensive care unit with ventilation.

The mortality of this disease increases day by day, and this disease becomes as a big threat to humankind of the entire world. Along with the clinical researches, the analysis of related data will the help mankind. Many researches have already been done on the Computed Tomography (CT) images of the patients [4-5], their symptom-based analysis, and the influencing factors. Researches on CT images were done for identifying the characteristics of the disease and also diagnosing the disease early. CT images of COVID-19 cases have similarities in terms of inward and circular diffusion [4].

The classifications of Covid-19 are Influenza-A viral pneumonia, Covid-19, and healthy one [4]. The research is done based on CT images of 618 images with 224 images of Influenza patients, 219 images of COVID-19 patients, and 175 healthy humans, and they achieved 87.6% accuracy. Another study [5] was done for segmenting and quantifying the infection of CT images. They used the CT images of the chest and lung, and they implemented it using deep learning techniques. They used 249 images for training and 300 images for testing and achieved an accuracy of 91.6%. Pathological tests and analysis of CT images take some time. So researches are done based on the possibility of disease prediction based on the symptoms. This work uses some classification techniques for predicting the possibility of occurrence of COVID-19 based on their characteristics.

Most of the existing works concentrate on COVID-19 prediction using images. This work proposes the patient data-based prediction of COVID-19 possibilities (recovered or deceased) using the KNN classification algorithm. The prediction performance over 730 records of the COVID-19 patient dataset is evaluated using the KNN algorithm. Further, in this work, the MSE rate, kappa scores, and classification report are analyzed for the proposed KNN algorithm. The results of this research work, suggest that the proposed KNN

algorithm produces better results for the smaller dataset than other algorithms.

The organization of this paper is as follows. Section 2 gives the related work of classification techniques. Section 3 discusses the different machine learning models. Section 4 analyses the performance metrics, experimental analyses, and results, and Section 5 gives the concluding remarks with future work.

2. Related Work

The emergence of Artificial Intelligence (AI) transformed the world in all the fields. Machine learning (ML), a subset of AI helps the human to find solutions for highly complex problems and also plays a vital role in making human life sophisticated. The application areas of ML include business applications, intelligent robots, autonomous vehicle (AV), healthcare, climate modeling, image processing, natural language processing (NLP), and gaming. The learning of ML mimics human intelligence, and it is implemented based on trial and error method. The instructions to the algorithm were given mainly using control statements such as conditional if [6]. Many prediction based algorithms are available in ML [7]. The ML techniques are used for classification and prediction in various fields like disease prediction, stock market, weather forecasting, and business.

In the medical field also, many ML algorithms are used for disease prediction [8] like coronary artery disease [9], predicting cardiovascular disease [10], and prediction of breast cancer [11]. Several researches are also done for COVID-19 confirmed case live forecasting [12] and for predicting the COVID-19 outbreak [13]. These works will aid the higher authorities of the country for taking decisions to handle the situation by foreseeing [14]. At first, the COVID-19 was misinterpreted as pneumonia [15]. But the failure of multi-organs and high mortality rates [16] made it a pandemic in the whole world.

Classification techniques are broadly categorized into semi-supervised [17, 18], supervised [19] and unsupervised [20-23]. Supervised learning takes information about the classes and learns based on that information. Based on this knowledge, this technique can predict the classes for new data. In unsupervised learning, the information about the classes is unknown. The clustering of similar data is done by identifying the similarity among themselves. Semi-supervised techniques know some information about data, and the classification is done based on it. Logistic Regression [24-25] is used for relationship analysis between various dependent variables. Basically, it was used for identifying the existence of a class or event. This was further extended to classify more objects. Artificial neural network (ANN) [26-31] is based on learning and classifies effectively. Here, the nodes are arranged in the input layer, hidden layer, and output layer. Based on the objective function and the number of hidden layers will vary. Support Vector Machine (SVM) [32-35] is another classification technique that separates the variables using a hyperplane.

Many classification and prediction algorithms are applied to study about the possibility of spreading COVID-19. The research was done on the occurrence of asymptomatic infection, and they found it is higher (15.8%) in children under 10 years [36]. Some studies have done in identifying the symptoms and identified having lesser senses of taste and smell are the signs of Covid-19 [37]. Another work [38] also studied about the transmission process of this disease.

3. Methodology

In recent years, predictive medical analysis using machine learning techniques has tremendous growth with promising results. The machine learning algorithms are effectively applied in numerous types of applications in diverse fields. Many kinds of research have proved that the machine learning predictive algorithms had provided better assistance for clinical supports as well as for decision making based on the patient data [39]. In the healthcare field, the disease predictive analysis is one of the useful and supportive applications of machine learning prediction algorithms. This research work applies the predictive disease analysis using KNN machine learning prediction algorithms for the novel COVID-19 disease.

The contribution of the proposed work is listed as follows:

- This research work investigates COVID-19 patient data in order to assess the outcome possibilities of the patient.
- The KNN classification algorithm is proposed in this work to predict the outcome possibilities of patients such as recovered or deceased.
- The prediction results of the proposed KNN algorithm is evaluated for the accuracy rate,

mean square error rate, Kappa score, area under curve, indices, sensitivity, specificity, and f1 score values.

- This work considers the only two parameters of the patients with 730 records, and KNN algorithm based outcome prediction results are compared to other algorithms.

3.1 Data Preprocessing and Cleaning

The COVID-19 dataset from the Kaggle [40] is taken for the predictive analysis in this research work. The considered dataset was cleaned using the data preprocessing and data cleaning methodologies, then the resulted dataset has been considered for several number of experiments over different classification algorithms. The COVID-19 dataset contains the patient's details with recovered and deceased status. The vital patient's information is used to diagnose and predict the COVID-19 disease among the infected population.

The considered COVID-19 dataset contains 100284 records. The dataset contains features of patients such as patient number, state patient number, date announced, estimated onset date, age bracket, gender, detected city, detected district, detected state, state code, current status, notes, contracted from which patient (suspected), nationality, type of transmission, status change date, source_1, source_2, source_3, backup notes, num cases, entry_id [40].

The data preprocessing and cleaning process removes the missing and outliers data values from the dataset. The resulted dataset after preprocessing is reduced to 730 records with three required relevant

Table.1 Sample record of cleaned dataset

Age	Gender	Outcome
13	Female	Recovered
96	Male	Recovered
89	Female	Recovered
85	Male	Recovered
27	Male	Recovered
69	Female	Deceased
26	Male	Recovered
65	Male	Deceased
76	Male	Deceased
45	Female	Recovered

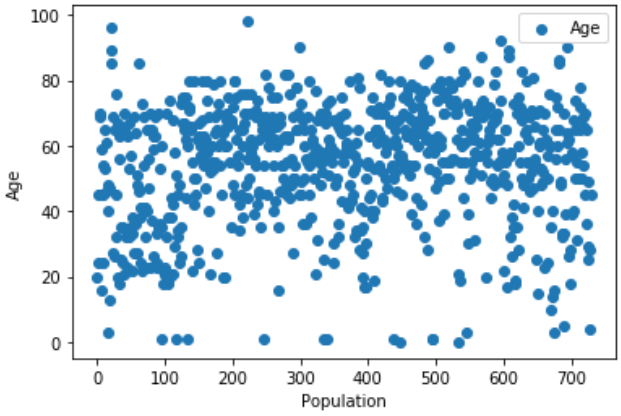

Fig.1 (a) Population vs Age

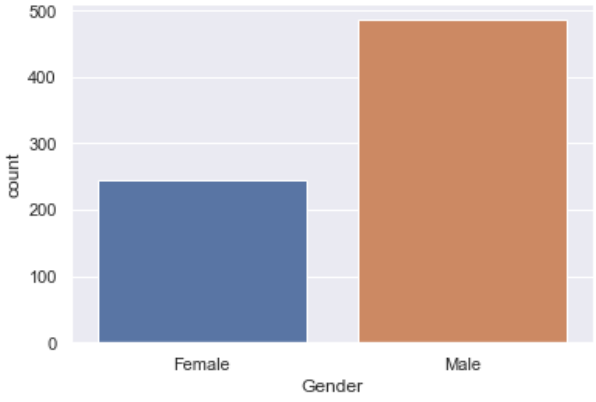

Fig.1 (b) Gender

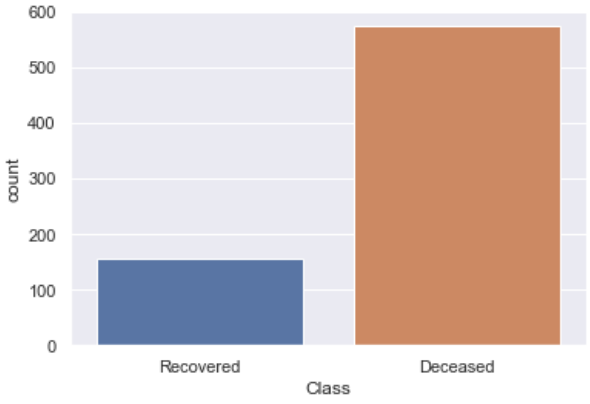

Fig.1 (c) Outcome

features of patient details. In the dataset, there are 730 patient details, out of which 156 cases are in the class of 'recovered from COVID-19 disease' and 574 cases are in the class of 'deceased by the COVID-19 disease' with 99554 records are missing required essential values. Two numerical features from the dataset are taken as the input attributes, and one feature is considered as the output attribute. The COVID-19 patient's information is presented in Table.1.

The patient features such as age and gender is considered as input variables, and the outcome is taken as the output variable—the features such as 1. Age- denotes the patient at the time of infection by the COVID-19 virus; 2. Gender- classifies whether the patient is male or female; 3. Outcome-denotes whether the patient has been recovered from COVID-19 disease or deceased due to COVID-19 disease. Figure.1(a) illustrates the population infected by COVID-19 with respect to age. Figure.1(b) and Figure.1(c) depict the count plot of gender and outcome of COVID-19, respectively.

This research work analysis the prediction of recovered and deceased patients infected by COVID-19. Different classification models are applied to the COVID-19 dataset, and its performance in terms of accuracy, error rates, etc. are evaluated. The classifiers evaluated in this research work are Logistic Regression (LR), K-Neighbors Classifier (KNN), Decision Tree (DT), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP).

3.2 Logistic Regression (LR)

One of the simple and powerful prediction algorithms is logistic Regression. The logistic Regression uses the sigmoid function for predictive modeling of the given problem. It models the dataset maps them into a value between 0 and 1. The logistic Regression performs the predictive analysis based on the relationship between the binary dependent variable and the other one or more independent variables from the given dataset. In order to predict the output value *(Y),* the input values *(X1, X2,...Xn)* are linearly combined using the coefficient values [41]. Let us consider, *'Y'* as the output prediction variable and *X1* and *X2* are input variables, then the logistic regression equation is given as (1),

$$Y = \frac{1}{2}\left[\frac{e^{(mX1+c)}}{1+e^{(mX1+c)}} + \frac{e^{(mX2+c)}}{1+e^{(mX2+c)}}\right] \qquad (1)$$

Where *'c'* represents the intercept, *'m'* is the coefficient of input value *X1 and X2* (in our case, *X1, X2* are age, gender). The coefficient value *'m'* can learn from the training dataset for each input value *(X1, X2)* [41]. This work to classify the deceased and recovered cases of the COVID-19 disease using the equation (1).

3.3 K-Nearest Neighbors (KNN) Classifier

K-Nearest Neighbors algorithm is the non-parametric algorithm. The learning and prediction analysis is performed based on the given problem or dataset. The KNN classification model, the prediction is purely based on neighbor data values without any assumption on the dataset. In KNN, the *'K'* represents the number of nearest neighbor data values. Based on *'K'*, i.e., the number of nearest neighbors, the decision is made by the KNN algorithm on classifying the given dataset [42].

The KNN model directly classifies the training dataset. It means the prediction of a new instance is made by searching the similar *'K'* neighbor instances in the entire training set and classifying based on the class of highest instances. A similar instance is determined using the Euclidean distance formula. Euclidean distance is the square root of the sum of squared differences between the new instance (x_i) and the existing instance (y_j) [42].

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{jk})^2} \qquad (2)$$

3.4 Decision Tree (DT)

The decision tree algorithms are the powerful prediction model used for both classification and regression problems. The decision tree models are represented in the form of a binary tree. It means the given problem/dataset is solved by splitting or classifying them as a binary tree. In the decision tree, the prediction is made by taking the root node of the binary tree with a single input variable *(x),* splitting the dataset based on the variable, and its leaf nodes of the binary tree have resulted as the output variable *(y).* That is, from the root node, the tree is traversed through each branch with their divisions, and prediction is made based on the leaf nodes. It uses the greedy method for splitting the dataset in a binary manner [43].

In this research work, the COVID-19 dataset with two inputs *(x)* is taken as age, gender, and output is whether the patient is recovered or deceased. The decision tree classification algorithm uses the Gini index function to determine the impurity level of the leaf nodes for the predictions. The Gini index function *(G)* is given in equation (3).

$$G = \sum_{i=1}^{n} xk(1-xk) \qquad (3)$$

Where *'x'* is the proportion of training instances in the input class *'k'*. Binary tree representation of the dataset makes the prediction as straightforward [43].

3.5 Support Vector Machines (SVM)

The support vector machine can handle categorical and continuous variables. Also, the SVM model works well on classification and regression problems. The support vector machine is a classification algorithm that creates the hyperplanes for each class labels in the multidimensional space by employing the margin values. The SVM intends to maximize the margins among different classes by optimally separating hyperplanes [44]. The hyperplane is a data instance of the given dataset used by the support vectors. The margin is the maximum distance between the support vector and the hyperplane [44]. If the given dataset is linear bounded, then linear SVM can be adopted, and the dataset is non-linear bounded, then Non-linear SVM can be adopted for the classification tasks [45].

Let us consider a dataset $(A_1, B_1, \ldots A_n, B_n)$; where $(A_{1,\ldots} A_n)$ is the set of the input variable, $(B_1, \ldots, B_n)$ is the output variable, and *'C'* is the intercept, then the SVM classifier [44] is given as like equation (4).

$$SVM = \sum_{i=1}^{n} \beta_i - \frac{1}{2} \sum_{i,j=1}^{n} b_i b_j C(a_i, a_j) \beta_i \beta_j \quad (4)$$

In the equation (4), $i=1,2,,3\ldots.n$; and $C=b_i\beta_i + b_j\beta_j$. The SVM equation (4), is used in this research work to classify the deceased and recovered cases of the COVID-19 disease.

3.6 Multilayer Perceptron (MLP)

The Multilayer Perceptron algorithm is suitable for classification problems and predictive analysis. The MLP is the classical neural network with one or more layers of hidden neurons. It comprises the input layer (where the data variables are fed), the hidden layer (with function to operate on the data), and the output layer (contains the predicted values). MLP uses the back-propagation to learn from the given input and output dataset. The activation function $A_j(X, W)$ of the MLP is the summation of the inputs (X_i) multiplied with respective weights (W_{ij}) as represented in equation (5). The output function (O_j) with the sigmoid activation function of the MLP back-propagation [46] algorithm is given in equation (6).

$$A_j(X, W) = \sum_{i=0}^{n} (X_i, W_{ij}) \quad (5)$$

$$O_j(X, W) = \frac{1}{1 + e^{-A_j(X,W)}} \quad (6)$$

4 Results and Discussions

This section summarizes the prediction results of the logistic Regression, k-neighbors classifier, decision tree, support vector machines, and multilayer perceptron algorithms.

4.1 Cross-Validation

In order to evaluate and validate the performance of the machine learning model, resampling methods are adopted. This method estimates the prediction ability on the machine learning algorithm on new unseen input data. The k-fold cross-validation is one of the resampling procedure used in this work to validate the machine learning models on the limited data sample. The *'k'* represents the number of times the data model is to split. Each split of the data sample is called as a subsample or sampling group. These subsamples are used to validate the training dataset. In this work, the *'k'* value is chosen as 10. Therefore, it can be called as a 10-fold cross-validation resampling method. The 10-fold cross-validation method intends to reduce the bias of the prediction model [47].

4.2 Performance Metrics

Typically, the performance of the machine learning prediction algorithms measured by using some metrics based on the classification algorithm. In this work, the prediction results are evaluated by using the metrics such as accuracy, mean square error (MSE), root mean square error (RMSE), Kappa score, confusion matrix, the area under curve (ROC_AUC), classification performance indices, sensitivity, specificity, and f1 score values.

Mean Square Error (MSE): It is the average of the squared difference between predicted results (P_i) and actual results (A_i). It is calculated by using the equation is given in (7), where n is the number of samples [48].

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (T_i - A_i)^2 \quad (7)$$

Root Mean Squared Error (RMSE): The RMSE is the square root of the average of squared differences between predicted and actual results, likewise given in (8). It depicts the inconsistencies among the observed and predicted values [49].

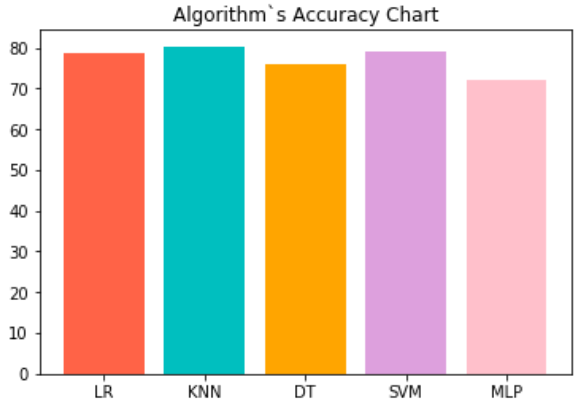

Fig.2 Prediction Accuracy

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - P_i)^2} \qquad (8)$$

Accuracy: Accuracy of the prediction algorithm is the ratio of the total number of correct predictions of class to the actual class of the dataset. The equation (9) calculates the accuracy of the model. Typically, any prediction model produces four different results, such as true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) [42].

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (9)$$

Precision: the precision of the prediction algorithm is the number of correctly predicted recovered COVID-19 cases that is belonging to the actual recovered COVID-19 cases [42, 47].

$$Precision = \frac{TP}{TP + FP} = \frac{True\ Positive}{Total\ predicted\ positive}$$

Recall: recall of the prediction algorithm is the number of correctly predicted recovered COVID-19 cases made out of all recovered COVID-19 cases in the dataset. It is a true positive rate [42, 47].

$$Recall = \frac{TP}{TP + FN} = \frac{True\ Positive}{Total\ predicted\ positive}$$

F1 Score: it is the measure of the balanced score (harmonic mean) of both precision and recall [42].

$$F1\ Score = \frac{Precision \times Recall}{Precision + Recall}$$

Cohen's kappa Score: Cohen's kappa score estimates the consistency of the prediction model. It compares the result of the predicted model with actual results. It is a statistic value between 0 and 1. The value near to 1 might have the great consistency [47].

$$K = \frac{[TP + TN/N] - [(TP + FN)\ (TP + FP)(TN + FN)/N^2]}{1 - [(TP + FN)\ (TP + FP)\ (TN + FN)]/N^2]}$$

Confusion matrix: The confusion matrix provides a complete insight into the performance of a prediction model. It produces prediction results in the matrix form with the information of the number of correctly predicted cases, incorrectly predicted cases, errors of incorrect, and correct prediction cases [47].

Receiver Operating Characteristic (ROC)-Area Under Curve (ROC_AUC): The ROC_AUC curve is a graphical illustration of the performance of the prediction model [47]. The ROC curve is the relationship between the recall and precision over varying threshold values. The threshold is the positive predictions of the model. The ROC_AUC curve plotted by keeping the x-axis a false positive rate and the y-axis as a true positive rate. Its value ranges from 0 to 1 [47].

4.3 Performance Evaluation

In most of the research works, the accuracy of the prediction model has been taken as one of the common performance metrics while working on a prediction algorithm [42]. In this work, the prediction accuracy (that is, whether the COVID-19 infected patient is recovered or deceased) of different machine algorithms (logistic Regression, k- nearest neighbor, decision tree, support vector machines, and multilayer perceptron) are determined. Each classification model has a different prediction accuracy based on its hyperparameters and a certain level of improvement over other prediction models. This work considers 70 % dataset for training and 30 % of the data samples for testing in classification algorithms. In this work, each model's accuracy is compared, and its prediction results are summarized in Table.2.

Table.2 Accuracy score of classifiers

S. No	Classifier	Accuracy
1.	Logistic Regression (LR)	78.5388
2.	K Neighbors Classifier (KNN)	80.3653
3.	Decision Tree (DT)	75.3425
4.	Support Vector Machines (SVM)	78.9954
5.	Multi-Layer Perceptron (MLP)	77.1689

In Table.2, the classification algorithms such as logistic Regression, k-nearest neighbor, decision tree, support vector machines, and multilayer perceptron have the prediction accuracy of 78.5388, 80.3653, 75.3425, 78.9954, and 77.1689 respectively.

Table.3 Error metrics of classifiers

S. No	Classifier	MSE	RMSE	Kappa
1.	Logistic Regression (LR)	0.2146	0.4633	0.4109
2.	K Neighbors Classifier (KNN)	0.1963	0.4431	0.469
3.	Decision Tree (DT)	0.2466	0.4966	0.3043
4.	Support Vector Machines (SVM)	0.21	0.4583	0.4266
5.	Multi-Layer Perceptron (MLP)	0.2283	0.4778	0.3411

Whereas, the k-nearest neighbor algorithm predicts the outcome of the COVID-19 cases (based on age and gender) more accurately than the other algorithms. Here, the *'k'* value is chosen as 2, since the KNN algorithm classifies into two clusters as recovered and deceased based on the training dataset.

multilayer perceptron have the MSE error rate as 0.2146, 0.1963, 0.2466, 0.21, and 0.2283, respectively. As per Figure.3(a), the KNN classification algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. The KNN algorithm

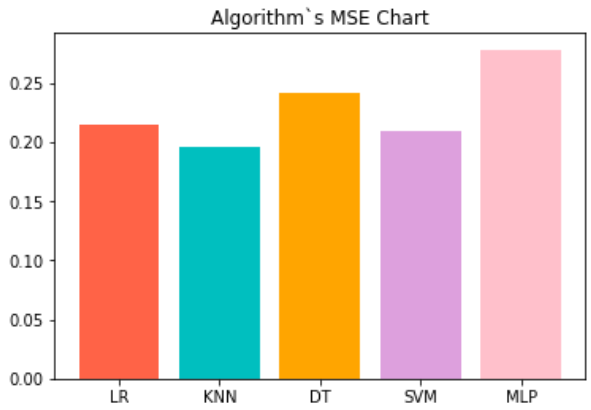

Fig.3(a) MSE rates

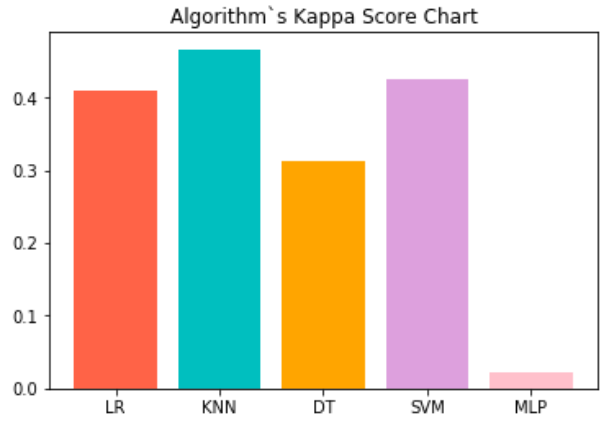

Fig.3(b) Cohen's kappa scores

It works by calculating the distance between the test data and training data. For each data points based on the distance values, the testing datasets are classified. Such that the KNN algorithm produces a higher classification rate than the other algorithms.

Figure.2 depicts the accuracy scores of different classification algorithms. From Figure.2, we can clearly see that the k-nearest neighbor algorithm has the highest accuracy of 80.4. The KNN algorithm has 1.5 to 3.3 % of improved accuracy as compared to LR, DT, SVM, and MLP algorithms. The KNN algorithm works by finding the similarity feature on the COVID-19 dataset in order to classify the features. The closely matching features are grouped together. Thus, it increases the accuracy rate of the KNN algorithm.

Table.3 presents the performance error metrics of the various machine learning algorithms. The error metrics mean square error, root mean square error, and Cohen's kappa score values for each algorithm is evaluated. The logistic Regression, k-nearest neighbor, decision tree, support vector machines, and

classifies the testing dataset by calculating the Euclidean distance between the new (testing) instance (x_i) and the existing (training) instance (y_j). Therefore, it results in lower error rates.

Similarly, the KNN's RMSE error rate also very low (0.44) as compared to the error rates of LR (0.46), DT (0.50), SVM (0.45), and MLP algorithms. Cohen's kappa score estimates the consistency of the classification algorithm based on its predictions. As depicted in Figure.3(b), the KNN classification algorithm produces the highest consistency among the evaluated algorithms as 0.47. The SVM algorithm offers the next highest consistency (0.42) on correctly predicting the COVID-19 cases. Moreover, the prediction of the decision tree algorithm has the lowest consistency value as 0.30. The LR and MLP have consistency values of 0.41 and 0.34, respectively.

Figure.4(a) illustrates the normalized confusion matrix of the k-nearest neighbor classification algorithm. In all classification algorithm, 30 % of the data samples are taken for testing with the 70 %

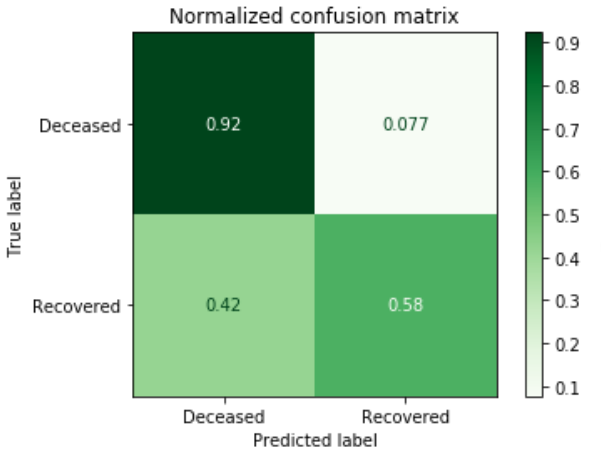

Fig.4 (a) Normalized Confusion matrix

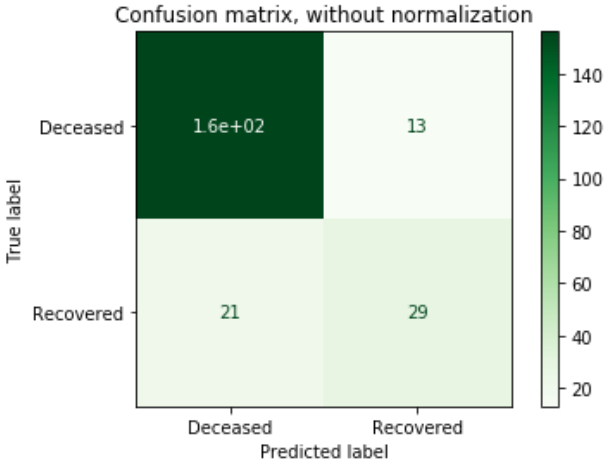

Fig.4 (b) Confusion matrix (no normalization)

training dataset. In this Figure, the x-axis represents the percentage of predicted values and the y-axis represents the percentage of true values. It can be seen that the KNN algorithm predicts 92 % (true positive) of the deceased cases correctly, with 0.077 % (false positive) of misclassification.

Similarly, in Figure.4(b) the confusion matrix without normalization is depicted, where 160 cases are correctly predicted as deceased cases, and 13 cases are misclassified. Further, 31 cases are correctly predicted as deceased cases, and 31 cases are misclassified. Also, it correctly predicts 29 patients (true negative) as the recovered cases, and 21 cases are misclassified (false negative).

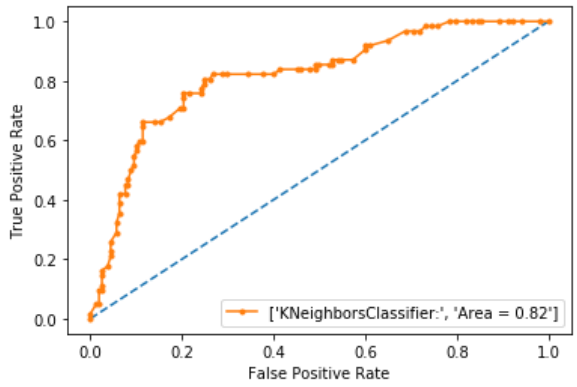

Fig.5 ROC_AUC Curve

Figure.5 is the pictorial representation between the false positive rate and true positive rate in the form ROC area under the curve. The k-nearest neighbor classification algorithm produces the highest value of 0.82 as compared with LR, DT, SVM, and MLP algorithms.

Figure.6 summarizes the performance metrics such as precision, recall, and confusion matrix of the k-nearest neighbor classification algorithm. The KNN algorithm produces the precision (true positive rate) value of 0.82 for the recovered cases and 0.72 for the deceased cases. The recall values for the recovered and deceased cases are 0.92 and 0.50, respectively. Further, the F1 score for recovered and deceased cases is 0.87 and 0.59, respectively.

```
KNeighborsClassifier:
--------------------

MSE: 0.1963470319634703
RMSE: 0.4431106317427628
Kappa_score: 0.46685161071165715
Accuracy: 80.36529680365297

Classification Report:

              precision    recall  f1-score   support

   Recovered       0.82      0.92      0.87       157
    Deceased       0.72      0.50      0.59        62

    accuracy                           0.80       219
   macro avg       0.77      0.71      0.73       219
weighted avg       0.79      0.80      0.79       219
```

Fig.6 Summary of Performance metrics scores of KNN algorithm

5. Conclusion and Future Enhancements

The predictive disease analysis is the major application area. This work has implemented Logistic Regression, k-nearest neighbor, decision tree, support vector machines, and multilayer perceptron to classify the COVID-19 dataset. The KNN classification algorithm has the 1.5 to 3.3 % of improved accuracy over other machine learning algorithms reported in work. Moreover, the KNN classification algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. In order to improve the accuracy of predictions, the future work will

concentrate on predicting the COVID-19 cases using classification and optimization algorithm

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China", *The Lancet,* Vol.395, No. 10223, pp. 497-506, 2020.

[2] P. Huang, S. Park, R. Yan, J. Lee, LC. Chu, CT. Lin, "Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study", *Radiology*, Vol. 286, No.1, pp. 286-95.

[3] A. Esteva, B. Kuprel, RA. Novoa, J. Ko, SM. Swetter, HM. Blau, "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115-118, 2017.

[4] X. Xie, X. Li, S. Wan, Y. Gong, "Mining X-ray images of SARS patients. Data Mining: Theory, Methodology, Techniques, and Applications", *Williams, Graham J., Simoff, Simeon J. (Eds.)*, pp. 282-294, ISBN: 3540325476 , Springer-Verlag, Berlin, Heidelberg, 2006.

[5] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, "Lung Infection Quantification of COVID-19 in CT Images with Deep Learning" *arXiv preprint* arXiv:200304655. 2020.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PloS one*, vol. 13, no. 3, 2018.

[7] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," *in European business intelligence summer school. Springer*, 2012, pp. 62–77.

[8] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: advantages, problems, and suggested solutions." *Cancer treatment reports*, vol. 69, no. 10, pp. 1071–1077, 1985.

[9] P. Lapuerta, S. P. Azen, and L. LaBree, "Use of neural networks in predicting the risk of coronary artery disease," *Computers and Biomedical Research*, vol. 28, no. 1, pp. 38–52, 1995.

[10] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *American heart journal*, vol. 121, no. 1, pp. 293–298, 1991.

[11] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Scienc*e, vol. 83, pp. 1064–1069, 2016.

[12] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *Plos one*, vol. 15, no. 3, p. e0231236, 2020.

[13] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response," *Jama*, 2020.

[14] WHO. Naming the coronavirus disease (covid-19) and the virus that causes it. [Online]. Available: https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

[15] Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china," *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi*, vol. 41, no. 2, p. 145, 2020.

[16] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, “Identification of a new human coronavirus,” *Nature medicine*, vol. 10, no. 4, pp. 368–373, 2004.

[17] HG. Lee, M. Piao, YH. Shin, "Wind Power Pattern Forecasting Based on Projected Clustering and Classification Methods," *ETRI Journal*, Vol.37, pp.283–294, 2015.

[18] L. Gomez-Chova, G. Camps-Valls, L. Bruzzone, J. Calpe-Maravilla, "Mean Map Kernel Methods for Semisupervised Cloud Classification". *IEEE Trans. Geosci. Remote Sens*, Vol.48, pp.207–220, 2010.

[19] C. Bishop, "Improving the Generalization Properties of Radial Basis Function Neural Networks", *Neural Comput*. Vol.3, pp.579–588, 1991.

[20] J. Lee, LJ. Du, "Unsupervised classification using polarimetric decomposition and the complex Wishart classifier" *IEEE Trans. Geosci. Remote Sens,* Vol. 37, pp.2249–2258, 1999.

[21] T. Pahikkala, A. Airola, F. Gieseke, O. Kramer, "Unsupervised Multi-Class Regularized Least-Squares Classification", *In Proceedings of the 12th IEEE International Conference on Data Mining (ICDM),* Brussels, Belgium, 10–13 December 2012, pp. 585–594.

[22] J. Hang, J. Zhang, M. Cheng, "Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine", *Fuzzy Sets Syst.,* Vol. 297, pp.128–140, 2016.

[23] KI. Kim, CH. Jin, YK. Lee, KD. Kim, KH. Ryu, "Forecasting wind power generation patterns based on SOM clustering", *In Proceedings of the 3rd International Conference on Awareness Science and Technology,* Dalian, China, 27–30 September 2011, pp. 508–511.

[24] J. Tolles, WJ. Meurer, "Logistic Regression Relating Patient Characteristics to Outcomes", *JAMA*. Vol.316, No. 5, pp.533–4, 2016.

[25] Tjur, Tue "Coefficients of determination in logistic regression models". *American Statistician*: pp.366–372, 2009.

[26] SJ. Lee, CL. Hou, "An ART-based construction of RBF networks", *IEEE Trans. Neural Netw.* 2002, 13, 1308–1321.

[27] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control. Signals Syst.*, Vol. 2, pp.303–314, 1989.

[28] R. Durbin, DE Rumelhart, "Product Units: A Computationally Powerful and Biologically Plausible Extension to Backpropagation Networks", *Neural Comput.,* Vol.1, pp.133–142, 1989.

[29] O. Buchtala, M. Klimek, B. Sick, "Evolutionary optimization of radial basis function classifiers for data

mining applications", *IEEE Trans. Syst. Man Cybern.*, Vol.35, pp.928–947, 2005.
[30] X. Yao, "Evolving artificial neural networks*", Proc. IEEE*, Vol.87, pp.1423–1447, 1999.
[31] PA. Gutiérrez, F. López-Granados, JM. Peña-Barragán, M. Jurado-Expósito, MT. Gómez-Casero, C. Hervás-Martínez, "Mapping sunflower yield as affected by Ridolfia segetum patches and elevation by applying evolutionary product unit neural networks to remote sensed data", *Comput. Electron. Agric.*, Vol. 60, pp.122–132, 2008.
[32] BE. Boser, IM. Guyon, VN. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92)*, Pittsburgh, PA, USA, 27–29 July 1992; Association for Computing Machinery: New York, NY, USA, 1992; pp. 144–152.
[33] C. Cortes, V. Vapnik, "Support-vector networks. *Mach. Learn*., Vol.20, pp.273–297, 1995.
[34] S. Salcedo-Sanz, JL. Rojo-Álvarez, M. Martínez-Ramón, G. Camps-Valls, "Support vector machines in engineering: An overview. In Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery", *John Wiley & Sons:* Hoboken, NJ, USA, 2014; Volume 4, pp. 234–267.
[35] CW. Hsu, CJ. Lin, "A comparison of methods for multiclass support vector machines", *IEEE Trans. Neural Netw.* Vol.13, pp.415–425, 2002.
[36] L. Xiaoxia, Z. Liqiong, D. Hui, Z. Jingjing, L. Yuan, Q. Jingyu, "SARS-CoV-2 Infection in Children", *New England Journal of Medicine.* 2020.
[37] R. Beth, M. Charlotte, R. Anne, H. Claire, P. Sophie, V. Hemelrijck, "Mieke Anosmia and ageusia are emerging as symptoms in patients with COVID- 19: What does the current evidence say?2020" *ecancer* Vol.14, ed98.
[38] L. Lixiang, "Propagation analysis and prediction of the COVID-19," *Infectious Disease Modelling*, vol.5,pp. 282-292, 2020.
[39] C. Naganna, "An Improved Method for Disease Prediction using Fuzzy Approach", *2015 Second International Conference on Advances in Computing and Communication Engineering.*
[40] https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset
[41] WY. Chiang, D. Zhang, L. Zhou, "Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression", *Decision Support Systems*, Vol. 41, no.2, pp.514-31, 2006.
[42] O. Altay, M. Ulas, "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children", *In2018 6th International Symposium on Digital Forensic and Security (ISDFS)* 2018 Mar 22, pp. 1-4. IEEE.
[43] J. Elson, A. Tailor, S. Banerjee, R. Salim, K. Hillaby, D. Jurkovic, "Expectant management of tubal ectopic pregnancy: prediction of successful outcome using decision tree analysis", *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, Vol. 23, No.6, pp.552-6, 2006.
[44] B. Schoslkopf, A. Smola, "Learning with Kernels, Support Vector Machines", *London: MIT Press*, 2002.
[45] G. RaviKumar, GA. Ramachandra, K. Nagamani, "An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets", *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 4, N. 2, pp.272-277, 2014.
[46] A. Pal, JP. Singh, P. Dutta, "Path length prediction in MANET under AODV routing: Comparative analysis of ARIMA and MLP model", *Egyptian Informatics Journal*, Vol.16, No. 1, pp.103-11,2005.
[47] H. Wu, S. Yang, Z. Huang, J. He, X. Wang, "Type 2 diabetes mellitus prediction model based on data mining", *Informatics in Medicine Unlocked*, Vol.1, No. 10, pp. 100-7, 2018.
[48] P. Theerthagiri, "FUCEM: futuristic cooperation evaluation model using Markov process for evaluating node reliability and link stability in mobile ad hoc network", *Wireless Networks*. Vol. 15, pp.1-6, 2020.
[49] T. Prasannavenkatesan, "COFEE: Context-aware Futuristic Energy Estimation model for sensor nodes using Markov model and auto-regression", *International Journal of Communication System*, e4248, 2019.

Figures

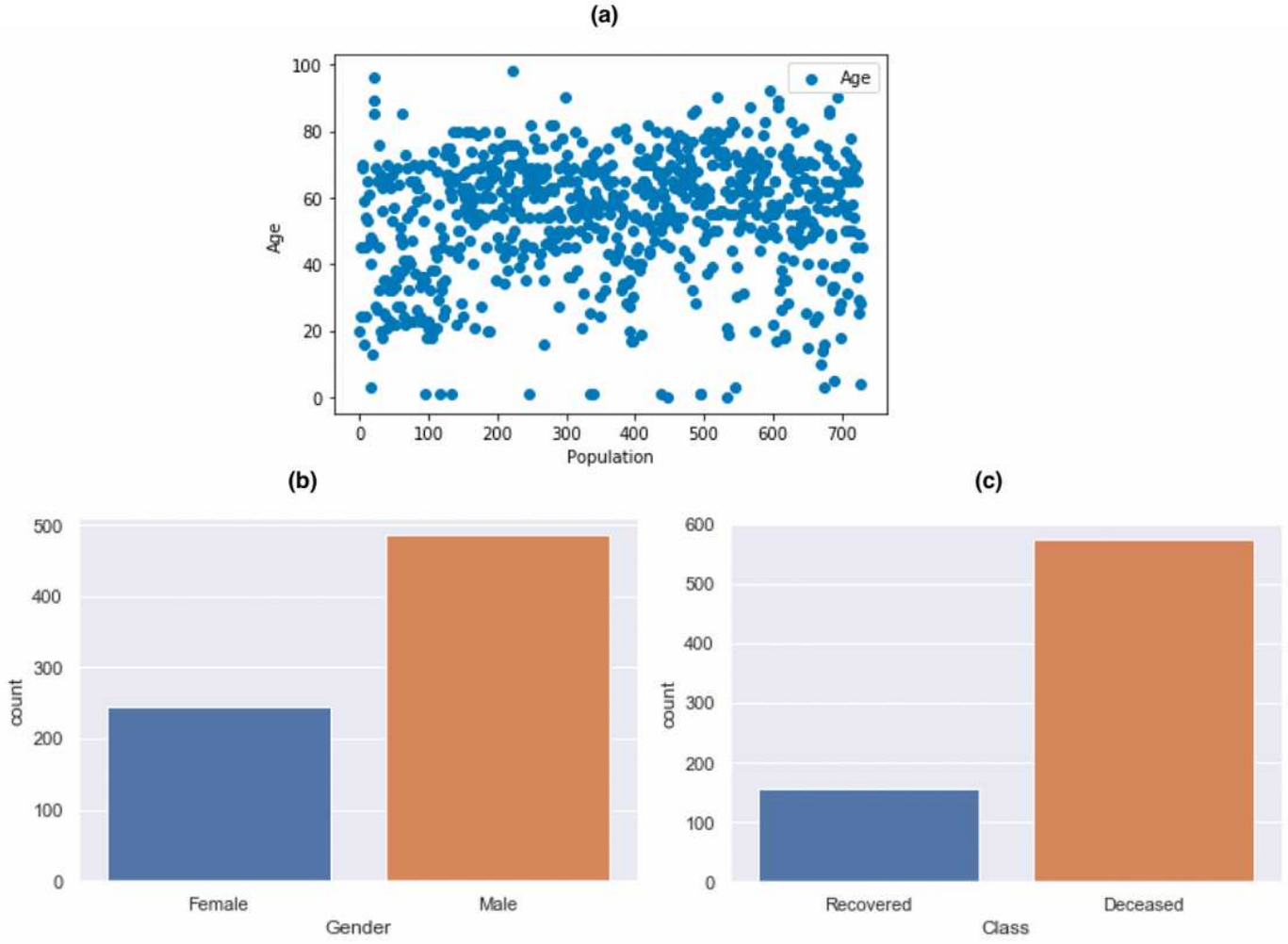

Figure 1

(a) Population vs Age (b) Gender (c) Outcome

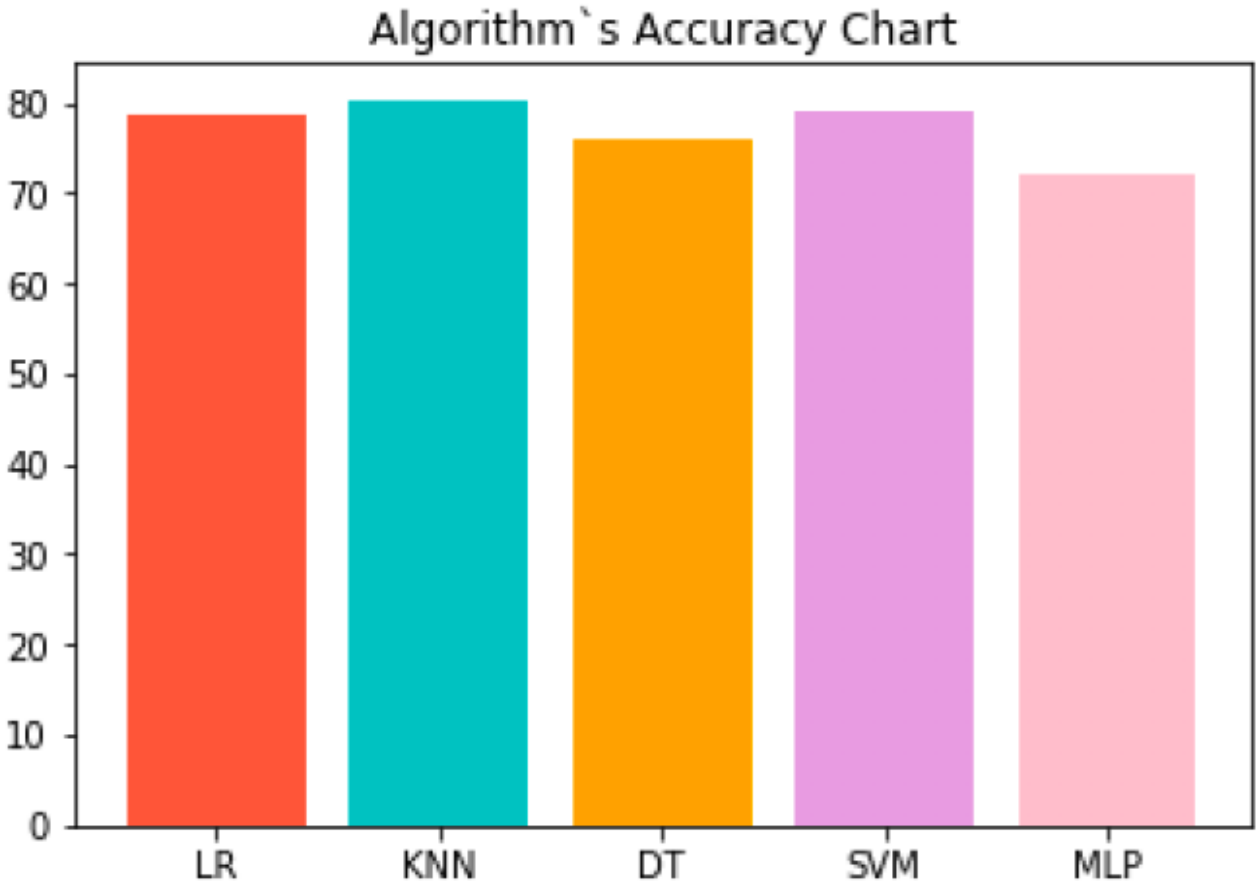

Figure 2

Prediction Accuracy

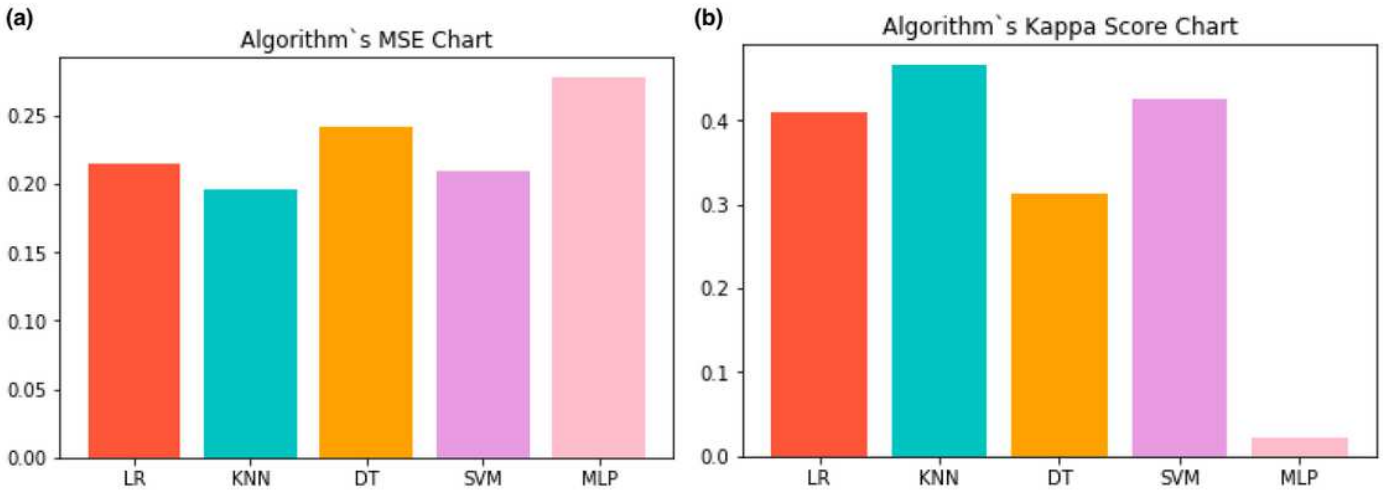

Figure 3

(a) MSE rates (b) Cohen's kappa scores

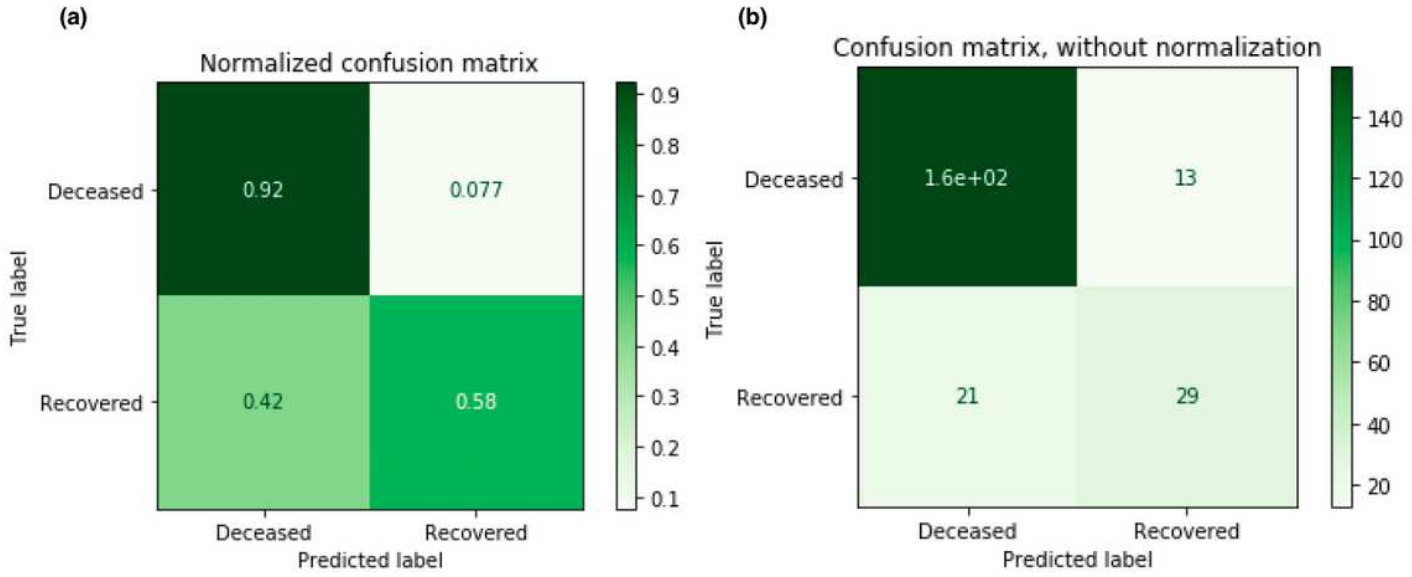

Figure 4

(a) Normalized Confusion matrix (b) Confusion matrix (no normalization)

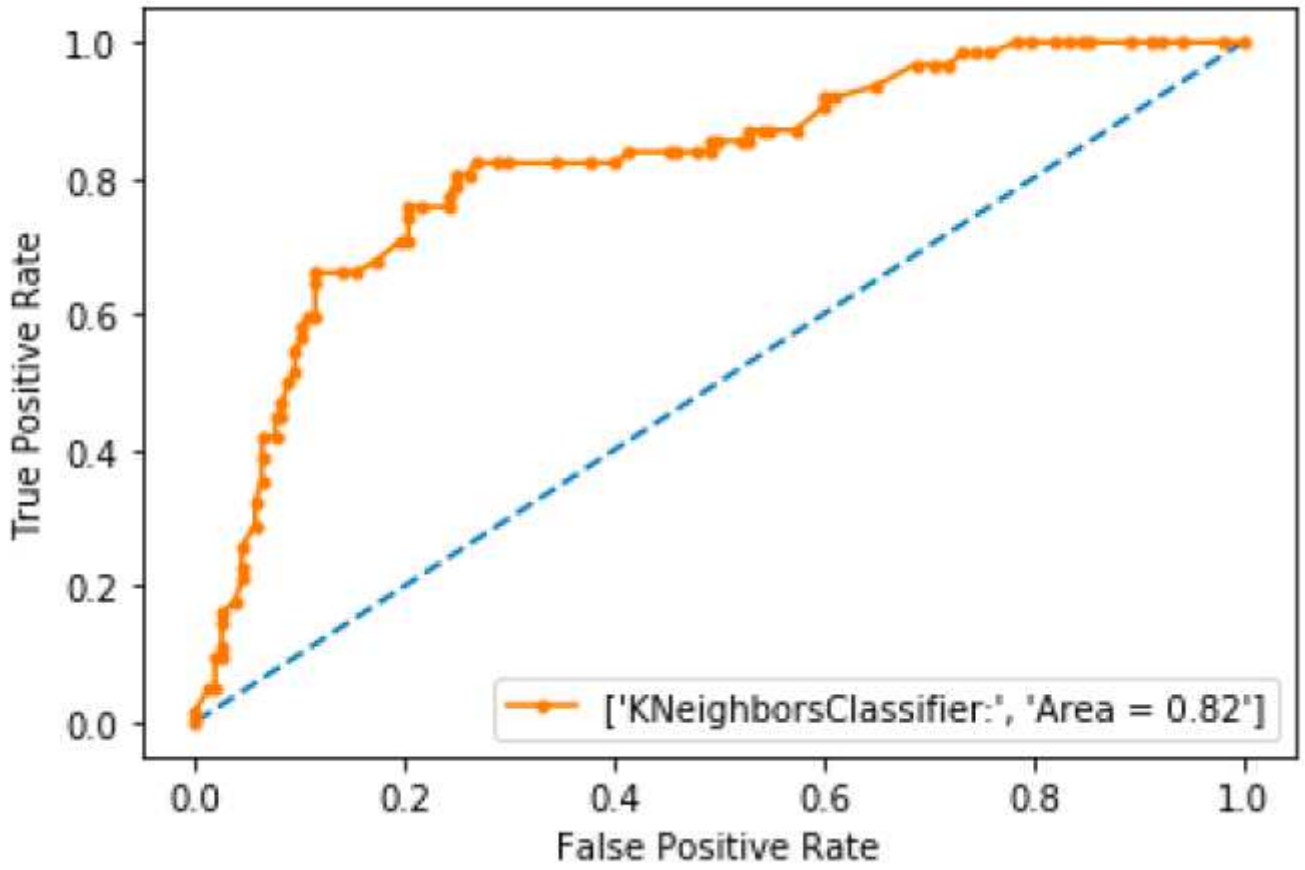

Figure 5

ROC_AUC Curve

KNeighborsClassifier:

MSE: 0.1963470319634703
RMSE: 0.4431106317427628
Kappa_score: 0.4668516107116571 5
Accuracy: 80.36529680365297

Classification Report:

	precision	recall	f1-score	support
Recovered	0.82	0.92	0.87	157
Deceased	0.72	0.50	0.59	62
accuracy			0.80	219
macro avg	0.77	0.71	0.73	219
weighted avg	0.79	0.80	0.79	219

Figure 6

Summary of Performance metrics scores of KNN algorithm