Prediction of Chorionic Kidney Disease Using Different Algorithm (May 2022)

Farjahan Akter Boby[1], Md. Asif Mostafa[2], Khan Raqib Mahmud3
[1] University of Liberal Arts Bangladesh Institute
[2]Department of Computer Science and Engineering, 688 Baridhara Road, Dhaka 1207

Corresponding author: Farjahan Akter Boby (e-mail: farjahanakterboby@gmail.com).

ABSTRACT Chronic kidney disease (CKD) is a global health issue that causes a high rate of morbidity and mortality, as well as the onset of additional diseases. Because there are no clear symptoms in the early stages of CKD, people frequently miss it. Early identification of CKD allows patients to obtain timely treatment to slow the disease's progression. Due to their rapid and precise recognition capabilities, machine learning models can successfully assist doctors in achieving this goal. So, this is necessary to know among all this machine learning algorithm which algorithm work well on CKD dataset. In our study, we proposed a solution to predict chronic kidney disease (CKD) using 21 different machine learning algorithms. The major goal of our research is to quantify and compare the performance of various algorithms and find out the best algorithm for CKD dataset. The algorithm used in this study are Logistic Regression, KNN Classifications, SVM, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Extra Tree Classifier, LGBClassifier, CatBoost, Bagging Classifier, Ridge Classifier CV, Linear SVC, Gaussian Process Algorithm, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm. These prediction models are built using a chronic kidney disease UCI repository dataset, and their performance is evaluated in order to determine the best classifier for predicting chronic kidney disease.

Keyword are Logistic Regression, KNN Classifications, SVM, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Extra Tree Classifier, LGBClassifier, CatBoost, Bagging Classifier, Ridge Classifier CV, Linear SVC, Gaussian Process Algorithm, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm.

I. INTRODUCTION

The failure of the kidney functionality is called chronic kidney disease (CKD). The core organ kidney basically filters the blood and transform the wastes into urine. But kidney failures, results in the full renal dysfunction, which can cause hormone disbalance, cystic ovaries, hearts problems, bone diseases etc. And sometimes it requires kidney dialysis or kidney transplants. For the more than 2 billion people who live in Southeast Asia's poor countries, CKD has become a major health issue [2]. Specially in Bangladesh's context CKD is increasing day by day. The symptoms of kidney disease are barely visible at the early stage and when it loses its 25% functionality people can identify the disease [3]. To mitigate these problems early identification and treatments of CKD is important. This study aimed to determine optimal model for prediction CKD with a high accuracy at the early stage. To do that here almost 21 different machine learning algorithms were used which includes Logistic Regression, KNN Classifications, SVM, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Extra Tree Classifier, LGBClassifier, CatBoost, Bagging Classifier, Ridge Classifier CV, Linear SVC, Gaussian Process Algorithm, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm using UCI repository dataset. Then the performance parameters were compared like as precision, recall, accuracy, and the ROC Curve.

II. Related work

In S.Revathy [1] presents a method to predict CKD in early stage by using data preprocessing, data transformation and various classifiers. In this research classifier are constructed using different algorithm like Decision Tree, Support Vector Machine and Random Forest and accuracy was above 94% for

each category. In this paper best classifier model was random forest with 99.16 percent accuracy. Anusorn Charleonnan [2] investigated the performance of four different classifier to predict CKD. In this paper they used Support Vector Machine, Logistic Regression, Decision Tree and KNN. According to the experimental data, the SVM classifier has the highest accuracy of the four with 98.3 percent, whereas the Logistic, Decision Tree, and KNN classifiers have lower accuracy. yield accuracy of 96.55 percent, 94.8 percent, and 98.1 percent on average respectively. JIONGMING QIN [3] introduced a new methodology for diagnosing CKD where they used KNN Imputation to fill in the missing value. Missing values are usually seen in real-life medical situations because patients may miss some measurements for various reasons. And this missing value can reduce model performance. In this paper they used different algorithm like logistic regression, random forest, support vector machine, k-nearest neighbor, naive Bayes classifier and feed forward neural network. UCI machine learning repository dataset was used. They used 24 different feature data contains 400 samples. In this dataset 11 variable or feature were numerical and 13 categorical variables. They also integrated model LOG and RF by using performance perception to improve the performance. And all model accuracy was above 95 percent. AKM Shahariar Azad Rabby [4] showed a comparison between 10 different algorithms like K Nearest Neighbor (known as KNN), Support Vector Machine (known as SVM), Random Forest algorithm (known as RF), Gaussian Naive Bayes, ADA Boosting Classifier, Linear Discriminate Analysis (known as LDA), Logistic Regression Classifier, Decision Tree Classifier (known as DT), Gradient Boosting Classifier and the Artificial Neural Network (known as ANN) to predict kidney disease. The findings of the analysis demonstrate that Decision Tree Classifier and Gaussian Naive Bayes was best performance classifier model among the other classifiers, with an accuracy score of 100 percent and a recall (Sensitivity) score of one. Amruta Rajeev Shetty [5] presents a method to predict CKD using data mining Technique as SVM And KNN With Pycharm. The all-models' performance is evaluated to determine which is the best classifier for predicting CKD for a particular dataset. SVM accuracy was 90.09 percent and KNN accuracy was 83.32 percent. Minhaz Uddin Emon [6] analysis the performance of CKD through machine learning. In this paper they 8 different machine learning classifiers were used like Naive Bayes (NB), Logistic Regression (LG), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (Adaboost), Bagging, Decision Tree (DT), Random Forest (RF) classifier. And they gain the highest accuracy from the Random Forest (RF) and it is 99% and ROC (receiver operating characteristic) curve value is also highest from other algorithms. Syed Danish [7] predict CKD using Naïve Bayes and Data Mining. There were 24 attribute and 400 instance in the dataset or classes or groups. And performance was high. Jdoud Abdullah Al Monsur[8 in this paper , they wanted to prevent CKD by utilizing machine learning technique by detecting CKD at an early age. They have taken datasets which consisted
24 attributes in the paper "Neural network and support vector machine for the prediction of chronic kidney and 400 instances. In this paper we saw that ANN technique is better predictive model in CKD the SVM. They have found that the accuracy of ANN is 99.75% and SVM model is 97.75%. The comparison between these two models have been done here in the terms of detecting CKD. S.Revathy [9] and fellow authors used decision tree ,Random forest and support vector machine learning model to to construct and diagnosis of CKD .They compared the results and found that Random forest classifier model is better to predict CKD then Decision trees and Support Vector machines .They found the accuracy rate of Decision tree is 94.16%, accuracy of SVM is 98.33% and Random Forest is 99.16% . Vijayarani Mohan [10] in this paper, they used Naïve Bayes and Support Vector machine algorithms for classification and predict CKD. MATLAB tool was used to implement this work. They have found that Naïve Bayes execution time was 1.29 s, and SVM execution time was 3.22 s. Here the algorithm which has higher accuracy and minimum execution that was chosen. They had come to a conclusion that SVM achieves increased classification performance.

III. Data Description and Analysis

We obtained the data from an internet source that was publicly accessible through the UCI repository. The UCI repository dataset was collected during a two-month period in India. There are 400 observations in this set of data. There are 250 records of patients with CKD and 150 records of patients without CKD in the data. Those who do not have CKD As a result, the

percentage of people in each class is 62.5 percent had CKD, while 37.5 percent did not. The ages of the observations range from 2 to 90 years old. Table 1 lists the features that were used to create our model.

Attribute Name	Attribute Code
Age	age
Blood Pressure	bp
Spesific Gravity	sg
Albumin	Al
Sugar	su
Red blood cells	Rbc
Blood urea	Bu
Serum creatinine	Sc
Sodium	Sod
Potassium	Pot
Hemoglobin	Hemo
White blood cell count	Wc
Red blood cell count	Rc
Hypertension	Htn
Class	class

Table 1. Attribute Description

IV. Methodology

A. Data Preprocessing

Missing, noisy, duplicated, and inconsistent data abound in today's real-world datasets, particularly in healthcare datasets. When you work with poor data, you'll get poor results. As a result, the first step in any machine learning application is to study and comprehend the dataset in order to prepare it for modeling. The term "data pre-processing" refers to this procedure.

1. Outlier

Extreme values that deviate from the majority of other data points in a dataset are known as outliers. Data entry errors cause invalid outliers, which are referred to as "noise in the data. “You must locate them and, if necessary, delete them. So, in this study we have detected outlier and remove them in 4 steps.

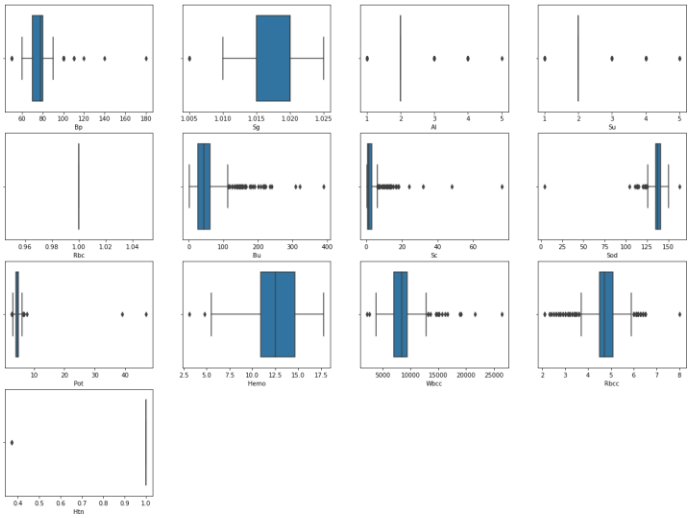


Figure 1: Detecting Outlier 1st step

First, we have an outlier for Bp, Sg, Al, Su, Bu, Sc, Sod, Pot, Hemo, Wbcc, Rbcc,Htn, as shown in Figure 1.Then,to remove these outliers, we have replaced them with the 5th and 95th percentiles, respectively. Then we got figure 2.

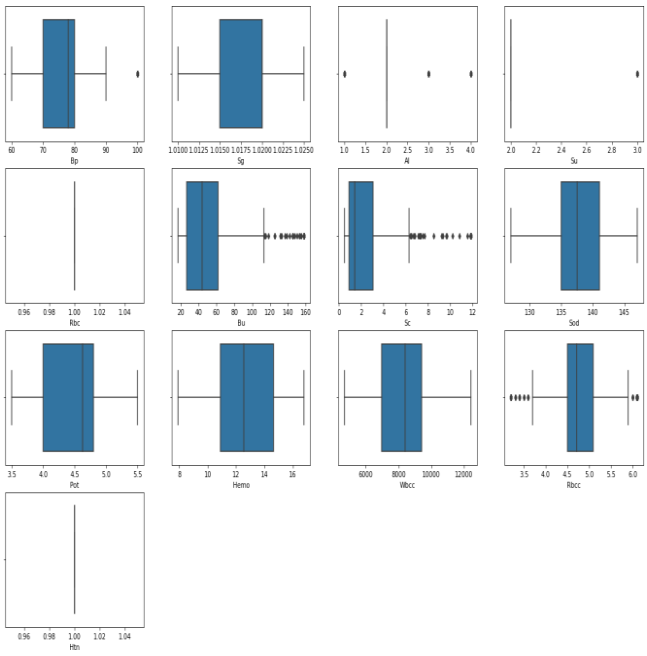


Figure 2: Detecting Outlier 2nd step

We can see that there were still outliers for Bp,Al,Su, Bu,Sc,Rbcc in Figure 2. After that, we replaced these outliers with different percentiles to remove them. An d finally got figure 3. Then we replaced the technique till the outlier was minimized at the 0 level.

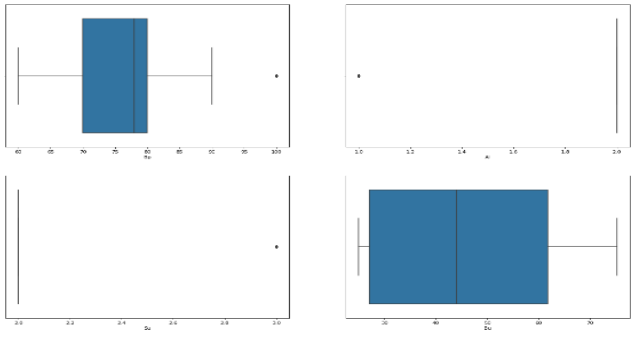


Figure 3: Detecting Outlier 3rd step

2. Missing Value

Missing data is a significant problem in real-world datasets, especially in the medical field. Every patient record and attribute, on average, has a few missing values**.** In this study Missing values are imputed with the feature's mean value, or with nan for continuous data, as part of data pre-processing.

3. Feature correlation

The linear relationship between two or more variables is measured via correlation. We can predict one variable from the other using correlation. The good variables are highly linked with the target, therefore employing correlation for feature selection is important. In this study we have shown data correlation using Heatmap as we can see in figure 4.

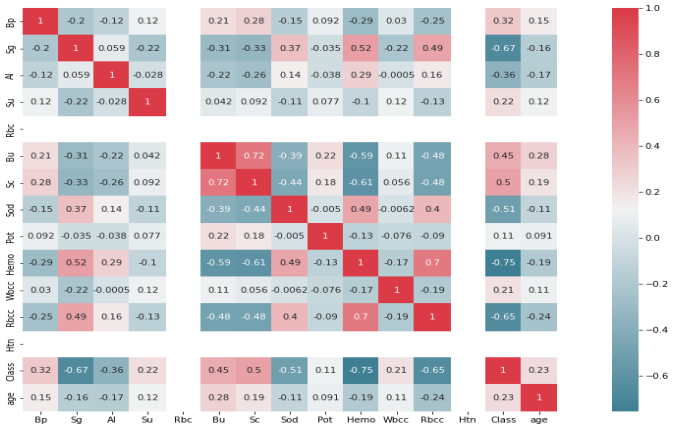


Figure 4: Heatmap data Correlation

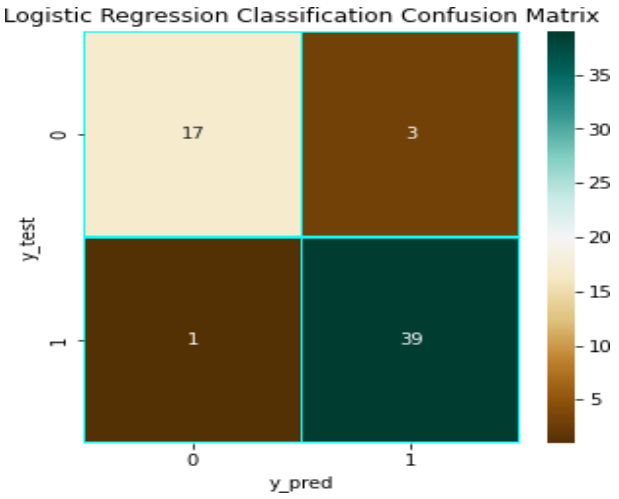
B. Modeling

In this study we have used 21 different machine learning algorithm Logistic Regression, KNN Classifications, SVM, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Extra Tree Classifier, LGBClassifier, CatBoost, Bagging Classifier, Ridge Classifier CV, Linear SVC, Gaussian Process Algorithm, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm. 85% of the data is utilized to train the model, with the remaining 15% used for testing. To evaluate various approaches, performance measurements like as accuracy, recall, percussion for Receiver Operating Characteristic (ROC) Curve are utilized.

1. Logistic Regression

Logistic regression is a machine learning classification technique. The dependent variable is modeled using a logistic function. The dependent variable is dichotomous, which means that only two classes are conceivable. So, in this study we have used Logistic Regression and got 0.93 percent accuracy. In Figure 5 we have shown the visual

confusion matrix of logistic regression from our experiment in below.



I.

Figure 5: Logistic Regression Confusion Matrix

2. K-NEAREST NEIGHBORS (KNN)

The K-Nearest Neighbors algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms. The K-NN method stores all available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a well-defined category using the K-NN method. So, in this study we have used K-Nearest Neighbor (KNN) and got 0.73 percent accuracy. In Figure 6 we have shown the visual confusion matrix of K-Nearest Neighbor (KNN) from our experiment in below.

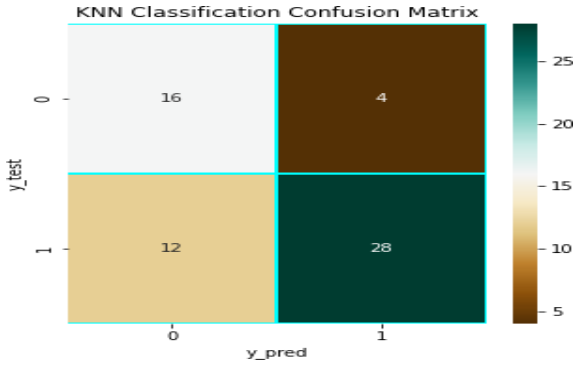


FIGURE 6: KNN CONFUSION MATRIX

3. SUPPORT VECTOR MACHINE ALGORITHM:

SVM (Support Vector Machine) is a common Supervised Learning technique for Classification and Regression. However, it is most commonly employed in Machine Learning for Classification issues. The SVM algorithm's purpose is to find the best line or decision boundary that can divide n-dimensional space into classes so that fresh data points can be readily placed in the correct category in the future. A hyperplane denotes the optimal choice boundary. The hyperplane is created using SVM, which selects the extreme points/vectors. Support vectors are the extreme situations, and the Support Vector Machine algorithm is named after them. So, in this study we have used SVM (Support Vector Machine) and got 0.66 percent accuracy. In Figure 7 we have shown the visual confusion matrix of SVM (Support Vector Machine) from our experiment in below.

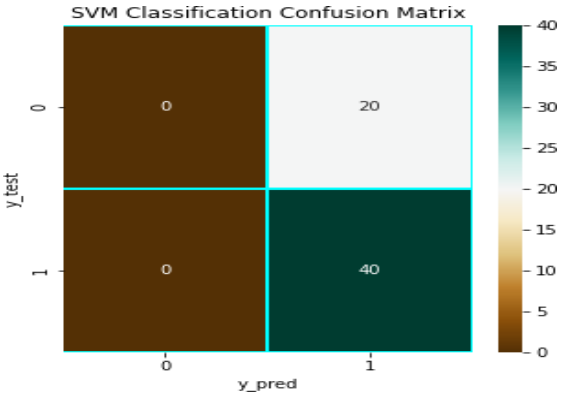


Figure 7: SVM (Support Vector Machine) Confusion Matrix

4. NAÏVE BAYES CLASSIFIER ALGORITHM:

The Nave Bayes method is a supervised learning technique for addressing classification issues that is based on the Bayes theorem. The Nave Bayes Classifier is a simple and effective classification method that aids in the development of fast machine learning models capable of making quick predictions. It's a probabilistic classifier, which means it makes predictions based on an object's probability. Classifier and got 0.81 percent accuracy. In Figure 8 we have shown the visual confusion matrix of Naïve Bayes Classifier from our experiment in below.

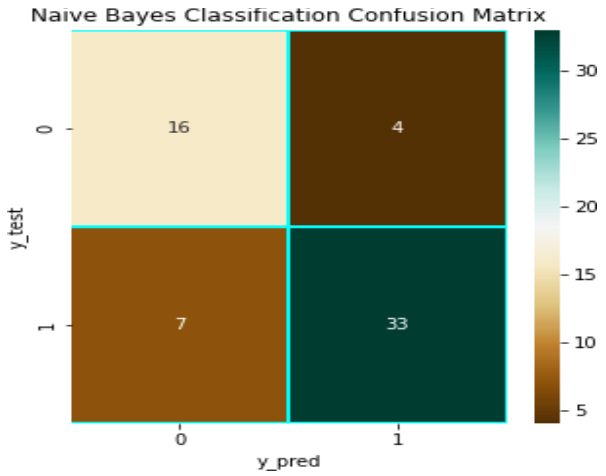


FIGURE 8: NAÏVE BAYES CLASSIFIER CONFUSION MATRIX

5. Decision Tree

The supervised learning category includes the decision tree method. They can be used to address problems involving regression and classification. The problem is solved using the tree representation, in which each leaf node corresponds to a class label and characteristics are represented on the tree's interior node. So, in this study we have used decision tree and got 0.95 percent accuracy. In Figure 9 we have shown the visual confusion matrix of decision tree from our experiment in below.

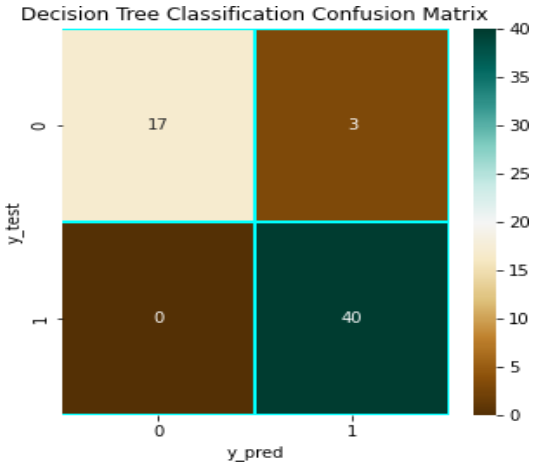


FIGURE 9: DECISION TREE CONFUSION MATRIX

6. Random Forest Algorithm

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it can be utilized for both classification and regression issues. It is based on ensemble learning, which is a method of integrating numerous classifiers to solve a complex problem and increase the model's performance. "Random Forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset," according to the name. Instead, then relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. So, in this study we have used Random Forest and got 100 percent accuracy. In Figure 10 we have shown the visual confusion matrix of Random Forest from our experiment in below.

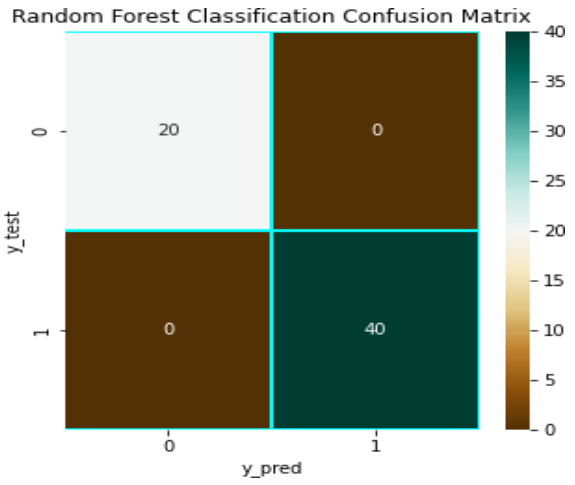


Figure 10: Random Forest Confusion Matrix

7. Gradient Boosting

One of the most powerful algorithms in the field of machine learning is the gradient boosting technique. As we all know, machine learning algorithm faults can be divided into two categories: bias error and variance error. Gradient boosting is one of the boosting strategies that is used to reduce the model's bias error. So, in this study we have used Gradient boosting and got 100 percent accuracy. In Figure 11 we have shown the visual confusion matrix of Gradient boosting from our experiment in below.

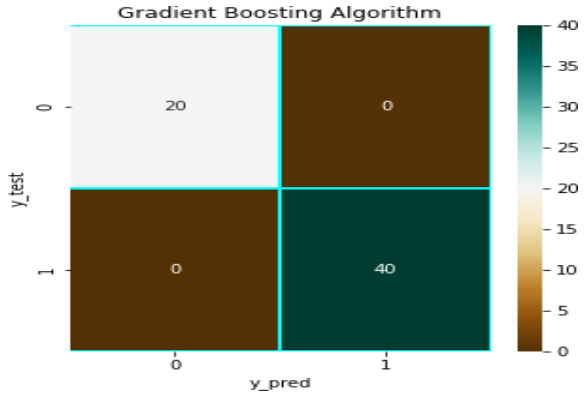


Figure 11: Gradient boosting Confusion Matrix

8. AdaBoost

One of the first boosting algorithms to be used in solving practices was AdaBoost. AdaBoost makes it possible to merge several "weak classifiers" into a single "strong classifier." Decision trees with a single split, also known as decision stumps, are the weak learners in AdaBoost. AdaBoost works by giving more weight to cases that are difficult to categorize and less to those that are already well-classified. Both classification and regression problems can be solved with AdaBoost algorithms. So, in this study we have used AdaBoost and got 100 percent accuracy. In Figure 12 we have shown the visual confusion matrix of AdaBoost from our experiment in below.

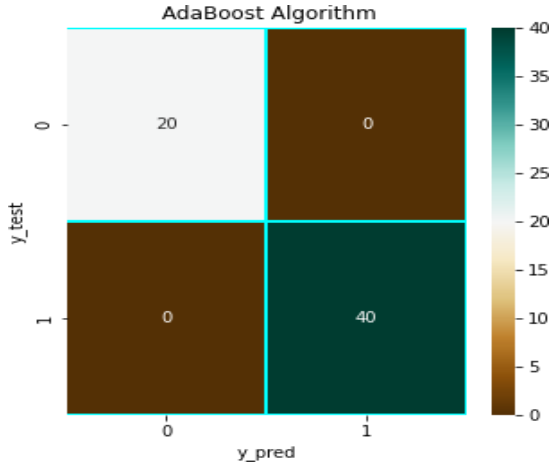


Figure 12: AdaBoost Confusion Matrix

9. XGBoost Classifier

The XGBoost classifier is a machine learning technique that may be used to classify both structured and tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. XGBoost is a gradient boost technique with high gradients. So, in this study we have used XGBoost and got 100 percent accuracy. In Figure 13 we have shown the visual confusion matrix of XGBoost from our experiment in below.

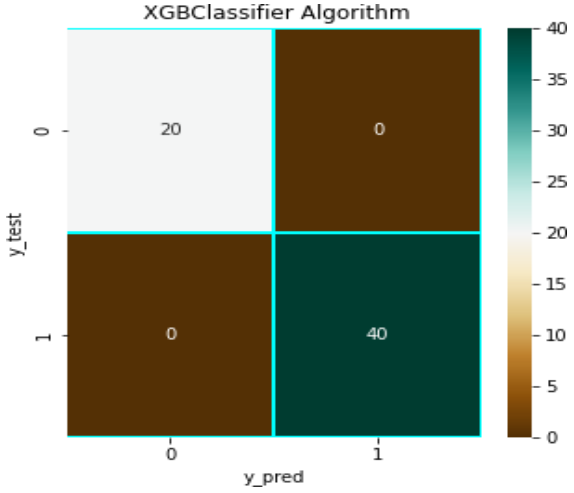


Figure 13: XGBoost Confusion Matrix

10. Extra Trees Classifier

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a form of ensemble learning technique that outputs a classification result by aggregating the outcomes of several de-correlated decision trees collected in a "forest." It is conceptually identical to a Random Forest Classifier, with the exception of how the decision trees in the forest are constructed. So, in this study we have used Extra Trees Classifier and got 0.96 percent accuracy. In Figure 12 we have shown the visual confusion matrix of Extra Trees Classifier from our experiment in below.

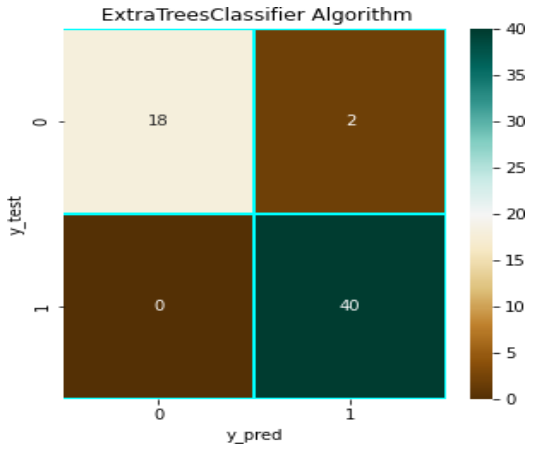


Figure 13: Extra Trees Classifier Confusion Matrix

11. LightGBM Classifier

In this kernel, I'll go through one of the most popular machine learning algorithms, LightGBM Classifier. LightGBM is a gradient boosting framework based on decision tree techniques that may be used for ranking, classification, and a variety of other machine learning problems**.** So, in this study we have used LightGBM Classifier and got 100 percent accuracy. In Figure 14 we have shown the visual confusion matrix LightGBM Classifier from our experiment in below.

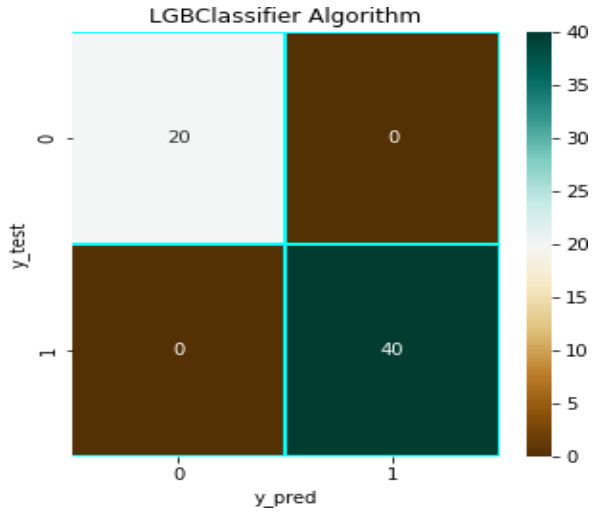


Figure 14: LightGBM Classifier Confusion Matrix

12. CatBoost Algorithm

CatBoost is a machine learning ensemble technique that belongs to the GBDT family. CatBoost has been utilized successfully for machine learning projects incorporating Big Data since its launch in late 2018. So, in this study we have used CatBoost and got 100 percent accuracy. In Figure 15 we have shown the visual confusion matrix CatBoost from our experiment in below.

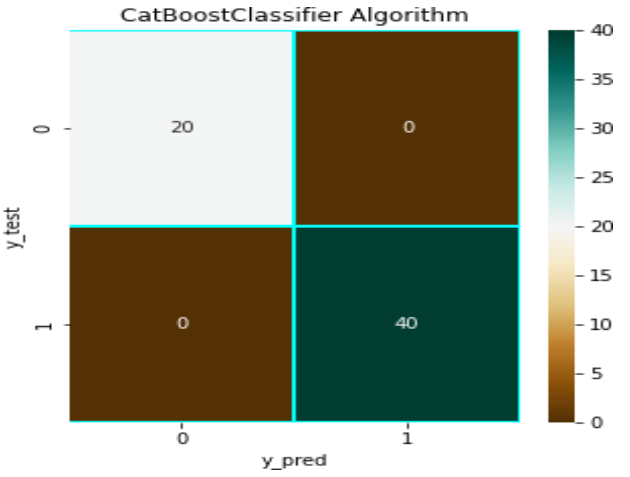


Figure 15: CatBoost Confusion Matrix

13. Bagging classifiers

Bagging classifiers are ensemble meta-estimators that fit base classifiers on random subsets of the original dataset and then aggregate their individual predictions (either by voting or average) to generate a final prediction. By averaging or voting, bagging lowers overfitting (variance), but it also increases bias, which is offset by the reduction in variance. So, in this study we have used Bagging classifiers and got 100 percent accuracy. In Figure 16 we have shown the visual confusion matrix Bagging classifiers from our experiment in below.

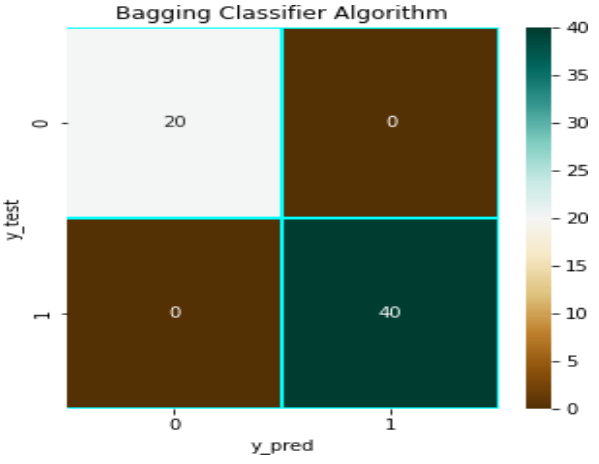


Figure 16: Bagging classifiers Confusion Matrix

14. RidgeCV

RidgeCV is a ridge regression approach that uses cross validation. RidgeCV is a ridge regression approach that uses cross validation. Ridge Regression is a type of regression that is commonly employed in multicollinear datasets. So, in this study we have used RidgeCV and got 0.98 percent accuracy. In Figure 17 we have shown the visual confusion matrix RidgeCV classifiers from our experiment in below.

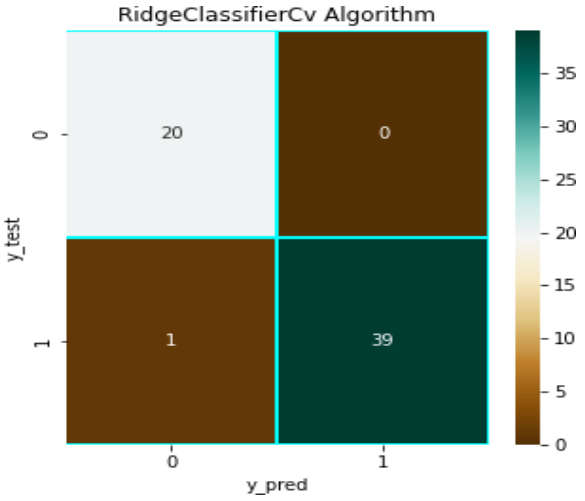


Figure 17: RidgeCV Confusion Matrix

15. Linear SVC

A Linear SVC (Support Vector Classifier) is designed to fit to the data you provide and provide a "best fit" hyperplane that divides or categorizes your data. Following that, you may input some features to your classifier to check what the "predicted" class is after you've obtained the hyperplane. This makes this algorithm particularly ideal for our purposes, however it can be used in a variety of circumstances. So, in this study we have used Linear SVC and got 0.35 percent accuracy. In Figure 18 we have shown the visual confusion matrix Linear SVC classifiers from our experiment in below.

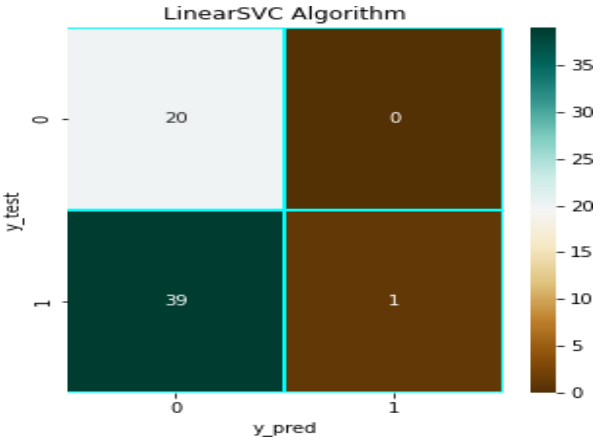


Figure 18: Linear SVC Confusion Matrix

16. Gaussian Processes

Gaussian Processes (GP) are a type of supervised learning method that can be used to address problems like regression and probabilistic categorization. The prediction extrapolates the data from the observations (at least for regular kernels). The prediction is probabilistic (Gaussian) so that empirical confidence intervals can be computed and used to determine if the prediction should be refitted (online fitting, adaptive fitting) in a particular region of interest. So, in this study we have used Gaussian Processes and got 0.68 percent accuracy. In Figure 19 we have shown the visual confusion matrix Gaussian Processes from our experiment in below.

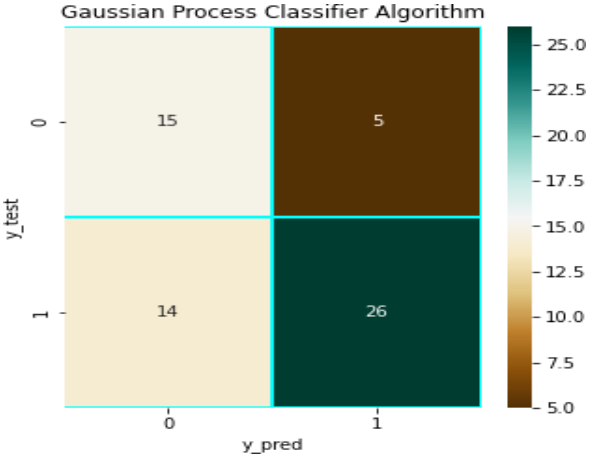


Figure 19: Gaussian Processes Confusion Matrix

17. Nu SVC

Nu-Support Vector Classification is an acronym for Nu-Support Vector Classification. Similar to SVC, however the number of support vectors is controlled by a parameter. So, in this study we have used Nu SVC and got 0.68 percent accuracy. In Figure 20 we have shown the visual confusion matrix Nu SVC from our experiment in below.

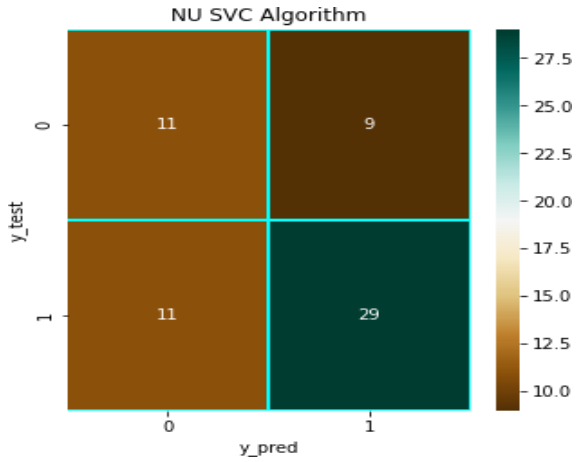


Figure 20: Nu SVC Confusion Matrix

18. BernoulliNB

For multivariate Bernoulli models, the Naive Bayes classifier is used. This classifier, like MultinomialNB, is good for discrete data. MultinomialNB works with occurrence counts, whereas BernoulliNB is for binary/Boolean features. So, in this study we have used BernoulliNB and got 0.68 percent accuracy. In Figure 21 we have shown the visual confusion matrix BernoulliNB from our experiment in below.

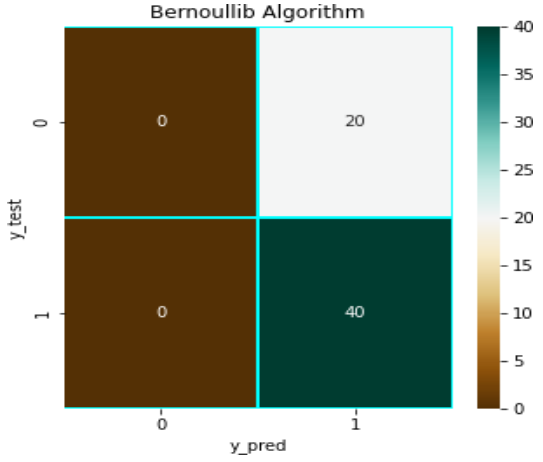


Figure 21: BernoulliNB Confusion Matrix

19. SGD Classifier

This estimator uses stochastic gradient descent (SGD) learning to create regularized linear models: the gradient of the loss is estimated one sample at a time, and the model is updated along the way with a decreasing strength schedule (aka learning rate). The partial fit method in SGD enables for minibatch (online/out-of-core) learning. The data should have a zero mean and unit variance for the best results when using the default learning rate schedule. So, in this study we have used SGD Classifier and got 0.66 percent accuracy. In Figure 22 we have shown the visual confusion matrix SGD Classifier from our experiment in below.

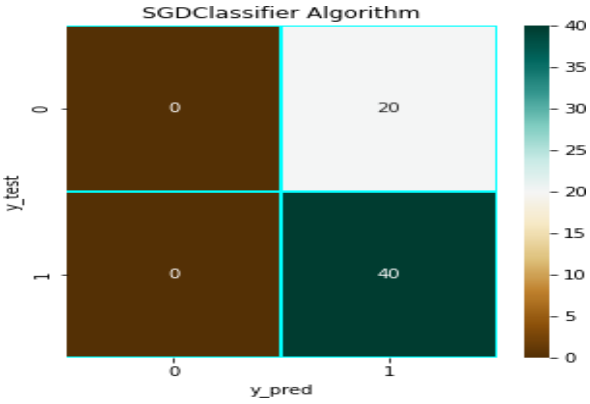


Figure 22: SGD Classifier Confusion Matrix

20. Perceptron algorithm

The perceptron algorithm is used to tackle binary classification problems. As a result, a Perceptron is a binary classifier that can determine if an input belongs to one of two classes. "Spam" or "ham" are two words that come to mind. So, in this study we have used perceptron algorithm and got 0.66 percent accuracy. In Figure 23 we have shown the visual confusion matrix perceptron algorithm from our experiment in below.

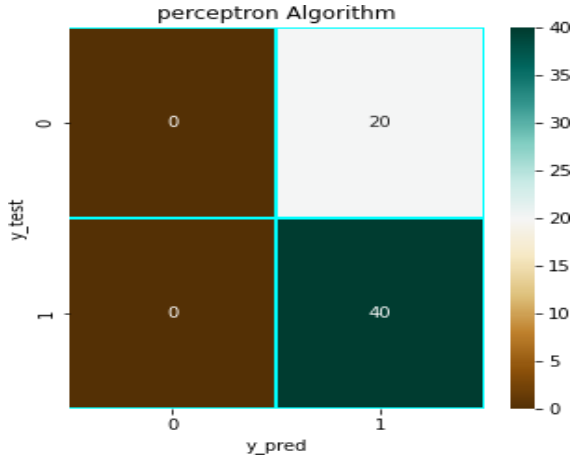


Figure 23: Perceptron algorithm Confusion Matrix

21. Passive-Aggressive algorithms

For large-scale learning, passive-aggressive algorithms are commonly used. It's one of the few 'online-learning algorithms' on the market. In contrast to batch learning, where the full training dataset is used at once, online machine learning algorithms take the input data in a sequential order and update the machine learning model step by step. So, in this study we have used passive-aggressive algorithms and got
0.66 percent accuracy. In Figure 24 we have shown the visual confusion matrix passive-aggressive algorithms from our experiment in below.

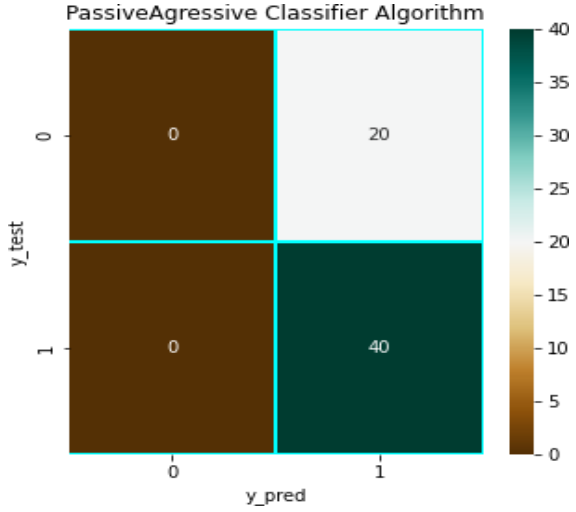


Figure 24: Passive-aggressive algorithms Confusion Matrix

V. Result and Conclusion

The most accurate algorithms are LightGBM, CatBoost, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Bagging Classifier. This algorithm has a 100% accuracy rate. For this dataset, these algorithms are most powerful. This research demonstrates that the CKD of a new patient may be predicted successfully with a 100% acceptable ratio. The accuracy rates of Ridge Classifier CV, Extra Tree Classifier, Decision Tree, Logistic Regression, Naïve Bayes, KNN Classifications are respectively 0.98 percent, 0.96 percent, 0.95 percent, and 0.93 percent, 0.81 percent, 0.73 percent. The Gaussian Process Algorithm as low 0.68 accuracy and SVM, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm as low accuracy 0.66. Linear SVC, algorithms have the lowest accuracy rates 0.35 among all algorithm Table 2 shows a comparison of algorithms in terms of Precision, Recall and accuracy.

MLA Namc	▮ʹíai-ı AUC	▮ʹcst AUC	Píccisio-ı	Rccall	AUC
LGBMClassifieí	1	1	1	1	1
CatBoost	1	1	1	1	1
Random Foíest	1	1	1	1	1
Gíadient Boosting	1	1	1	1	1
AdaBoost	1	1	1	1	1
XGBClassifieí	1	1	1	1	1
BaggingClassifieí	0.9969	0.975	0.980769	0.981	1
Extía ▮ʹíee	1	0.975	0.962963	1	0.96
RidgeClassifieíCV	0.9375	0.9625	0.980392	0.962	0.98
Decision ▮ʹíee	1	0.9375	0.927273	0.981	0.95
LogisticRegíession	0.9281	0.9	0.923077	0.923	0.93
GaussianNB	0.9	0.8625	0.918367	0.865	0.81
Gaussian Píocess	1	0.7125	0.808511	0.731	0.68
Knn	0.8	0.6625	0.755102	0.712	0.73
SVC	0.6188	0.65	0.65	1	0.66
LineaíSVC	0.6156	0.65	0.653846	0.981	0.35
BeínoulliNB	0.6188	0.65	0.65	1	0.66
SGDClassifieí	0.6188	0.65	0.65	1	0.66
Peíceptíon	0.6188	0.65	0.65	1	0.66
Passive Aggíessive	0.6188	0.65	0.65	1	0.66
NuSVC	0.6562	0.55	0.690476	0.558	0.66

Table 2. Comparing The Various Algorithm Carried Out CKD

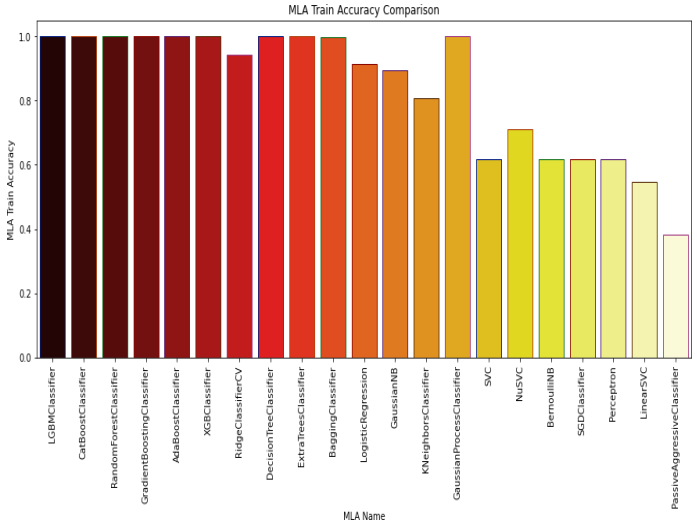


Figure 25: Comparison of all different algorithm train set accuracy

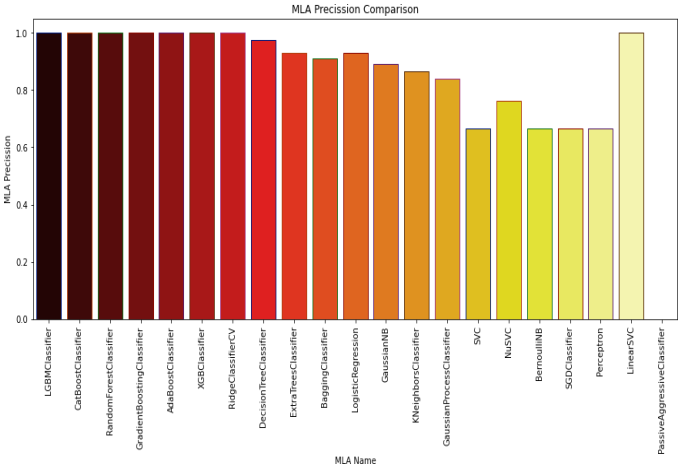


Figure 27: Comparison of all different algorithm precession

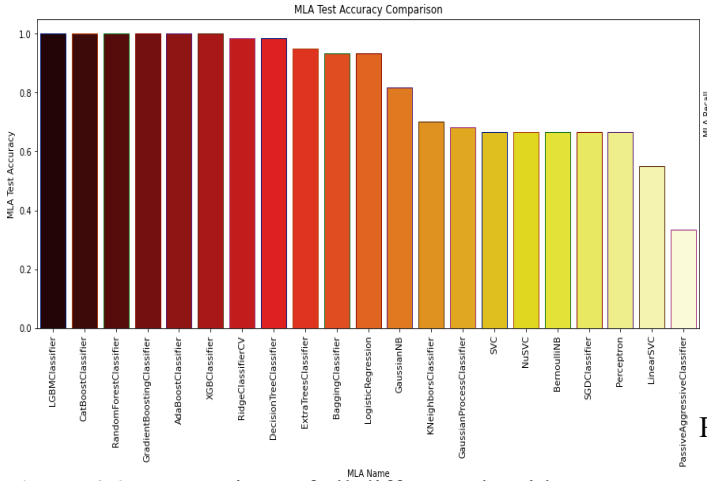


Figure 26: Comparison of all different algorithm test set accuracy

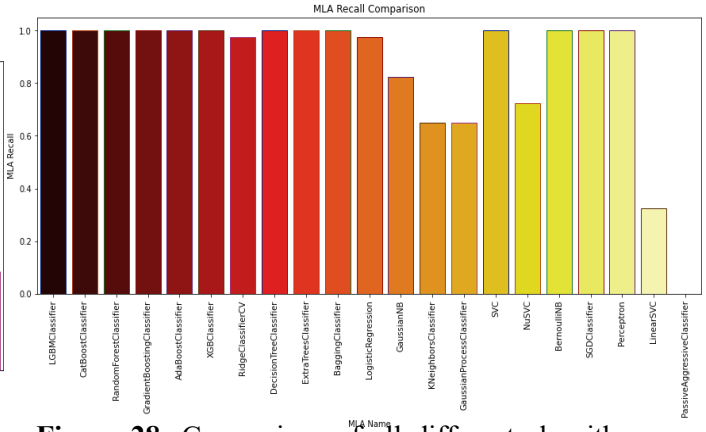


Figure 28: Comparison of all different algorithm Recall

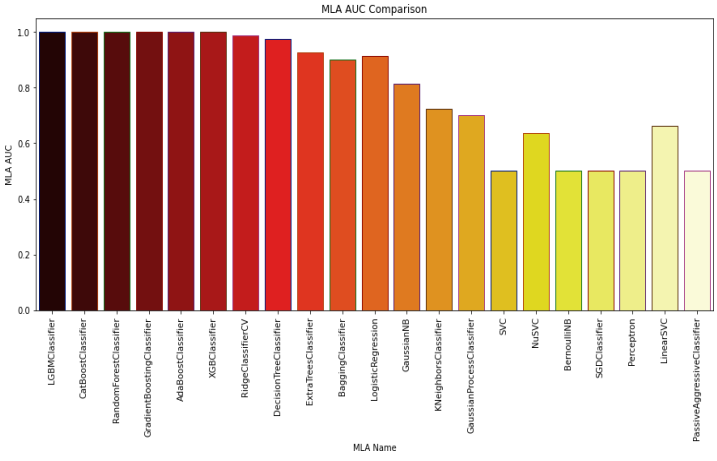


Figure 28: Comparison of all algorithm's Accuracy

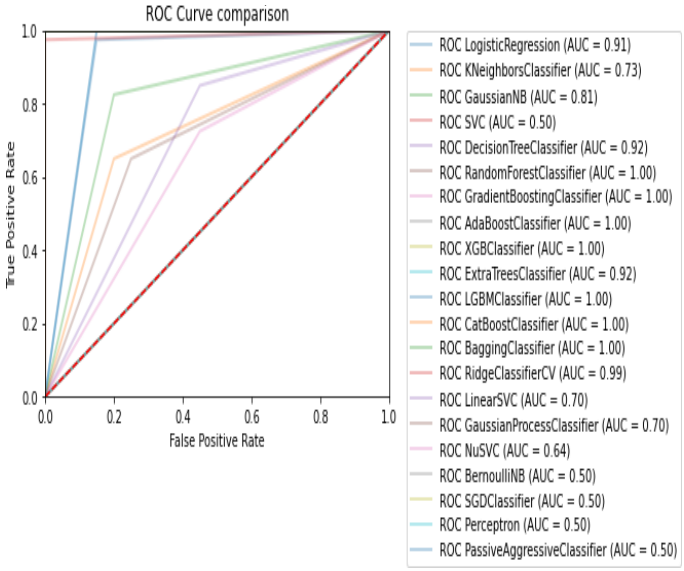


Figure 29. Comparison of The Algorithms with ROC Curve

VI. Conclusion

To predict CKD at an earlier stage, the researchers used algorithms such asLogistic Regression, KNN Classifications, SVM, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Extra Tree Classifier, LGBClassifier, CatBoost, Bagging Classifier, Ridge Classifier CV, Linear SVC, Gaussian Process Algorithm, Nu SVC Algorithm, BernoulliNB algorithm, SGD Classifier, Perceptron Algorithm, Passive Aggressive Algorithm. These algorithms produce a variety of experimental outcomes depending on Accuracy, Precision, Recall, and the ROC Curve. The effectiveness of these strategies was assessed and contrasted. According to the findings, the LightGBM, CatBoost, Random Forest, Gradient Boosting, AdaBoost, XGB Classifier, Bagging Classifier Algorithm beats other algorithms and achieves 100% accuracy. The use of a different algorithm to anticipate CKD will help to understand which algorithm works better in CKD Dataset. However, in the future, we will collect the most recent dataset on CKD diagnosis from various places throughout the world. The findings of this study will motivate us to continue working on additional projects.

VII. Reference

[1] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh , Chronic Kidney Disease Prediction using Machine Learning Models , International Journal of Engineering and Advanced Technology (IJEAT), 2019.

[2] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, The Management and Innovation Technology International Conference (MITiCON-2016) , 2016.

[3] JIONGMING QIN, LIN CHEN, YUHUA LIU, CHUANJUN LIU , CHANGHAO FENG , BIN CHEN , A Machine Learning Methodology for Diagnosing Chronic Kidney Disease ,IEEE,2019.

[4] AKM Shahariar Azad Rabby , Rezwana Mamata, Monira Akter Laboni , Machine Learning Applied to Kidney Disease Prediction: Comparison Study, Conference Pape,2019.

[5] Amruta Rajeev Shetty, Fouziya Basheer Ahmed, Veena Madev Naik, CKD Prediction Using Data Mining Technique as SVM And KNN With Pycharm, nternational Research Journal of Engineering and Technology (IRJET),2019.

[6] Minhaz Uddin Emon, Md. Al Mahmud Imran, Rakibul Islam, Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches, DAFFODIL INTERNATIONAL UNIVERSITY DHAKA, BANGLADESH ,2021.

[7] Syed Danish, Shaikh Aamer, S.P Kharde, S.S Gadekar, REVIEW ON CHRONIC KIDNEY DISEASE USING NAÏVE BAYES ALGORITHM, International Research Journal of Modernization in Engineering Technology and Science,2020.

[8] Njoud Abdullah Almansour, Hajra Fahim Syed, Nuha Radwan Khayat, Rawan Kanaan Altheeb, Renad Emad Juri, Jamal Alhiyafi, Saleh Alrashed, Sunday O, Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study, Computers in Biology and Medicine,2019

[9] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, Chronic Kidney Disease Prediction using Machine Learning Models, International Journal of Engineering and Advanced Technology (IJEAT),2019.

[10] Vijayarani Mohan, data mining classification algorithms for kidney disease prediction, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4

Farjahan Akter Boby

Currently pursuing BSc in Computer Science and Engineering at University of Liberal Arts Bangladesh (ULAB). At the same time holding the position of Finance secretary in ULAB Computer Programming Club (UCPC). Her field of research interest are Machine Learning and Deep learning.

Asif Mostafa

Currently pursuing BSc in Computer Science and Engineering at University of Liberal Arts Bangladesh (ULAB). Research interest Machine Learning and Deep Learning.