Prediction of Kidney Disease Using Ensemble Learning Techniques

Shaik Nazreen[1], G.V.S.Joshna[1], G.Mahesh[1], K.Khyathilekha[1], Bhanu Prakash Doppala[1]

[1]Department of CSE, Vignan's Institute of Information Technology(A),
Visakhaptanam, India

Abstract: Chronic renal disease also known as chronic kidney disease, is slowly increasing as a disease with high mortality rate. A person can manage to survive for 18 days without kidneys and being affected by kidney related diseases that can be cured only by kidney transplants and dialysis. It is critical to have good strategies for predicting CKD early on. Machine learning algorithms are useful for predicting CKD. By collection of various clinical data that includes preparation of data, a mechanism to manage missing values, selection of attributes and collaborative filtering, a strategy is presented through this paper. This takes into account the practical elements of data gathering and emphasises the necessity of domain knowledge when applying prediction of CKD status by machine learning and using 26 characteristics and data from CKD patients.Hence,in this proposed system we use ensemble learning techniques to predict the CKD with better accuracy.

Key words: CKD, Bagging, Logistic Regression, Decision Tree, SVM.

1. Introduction:

The Chronic Kidney Disease (CKD) in the modern era is regarded as a threat to the society's health. Chronic kidney disease can be detected by regular laboratory tests. The causes of CKD can be caused by a myriad of reasons such as smoking, lack of sleep and also various other factors. The diagnostic system in present day is made by examining urine with the help of serum creatinine levels. Medical methods such as ultrasound and screening are used for the above mentioned purpose. Screening is done on patients with hypertension, history of diseases mainly cardiovascular disease and also patients with relatives who have suffered from kidney disease in the past. The estimated value of GFR is calculated from the level of serum creatinine and the urine albumin-to-creatinine ratio also known as ACR is measured in a urine specimen taken in the first morning[1].

The gradual loss of kidney function is known to be CKD, also known as chronic kidney failure. The main function of the kidneys is filter the excess fluids from the blood and also filter wastes, which are then excreted in the form of urine. In the case of advanced chronic kidney disease, electrolytes, wastes and dangerous fluids accumulate inside the body[2].

To create multiple versions of a predictor and combining them to produce an aggregated predictor, the technique of bagging predictors is used. In the case of prediction of a numerical outcome, the average of the versions is taken by the aggregation and a plurality vote is used while predicting a class. Multiple versions can be created by making bootstrap replicas of the learning set and by using them as new learning sets. The method of bagging can increase accuracy significantly in tests on simulated and real data sets by using regression trees and classification, as well as the method of subset selection in linear regression[3].

The instability caused by this prediction method is critical. The method of bagging can be used to improve accuracy if significant changes can be caused in the predictor constructed by perturbing the learning set [4].

In order to increase the performance of the model, various ensemble methods are designed by statistical learning or fitting algorithms. To build a linear combination of a few model fitting methods, the general principle of ensemble methods is used instead of a single fit [5].

2. Literature Review:

Following section narrates about different works carried out for the detection and prediction of kidney disease. Various studies investigate & analyse kidney disease by employing several methods of prior detection. Patil [6] examines the detection accuracy of various data mining techniques, such as Artificial Neural Networks, decision table, naive Bayes, multilayer perception, LR, KNN and radical basic function. The accuracy of those models varies depending on dataset type, and no guideline for achieving the greatest outcomes.

Many studies are working on the prediction of CKD using various classification algorithms. And those researchers receive the predicted results of their model. To obtain results, they employ neural networks, KNN and random forest. They apply wrapper approach for reduction of features, which detects CKD with great accuracy.

A comparison analysis was achieved in this suggested work employing a sampling algorithm. The results of this case study show that employing algorithms which increase the classification algorithms performance .Finally, they determined the Decision Forest approach surpasses other models, with a precision of roughly 99 percent for the reduced dataset of 14 characteristics. The suggested effort is concerned with categorizing distinct phases of CKD based on their gravity. By examining several methods such as the Basic Propagation Neural Networks.

Few investigations while categorizing or diagnosing the CKD have been researched in recent years. Di Noia .T. et al.[7] published the software application which employees the ANN to identify status of the patient, that may result in the End-Stage Renal Disease. They were evaluated using recall, F-measure and precision. The offered product is accessible as an Android mobile application as well as an online web application.

Chase.H.S. et al. [8] recognized 2 categories of stage 3 patients uses the data in 117 individuals progressed (eGFR decreased by more than 3 points). Even if beginning eGFR levels were comparable, progressors had lower eGFR values than non-progressors.

Finally, they discovered that the Patients identified as progressors have a higher chance of progression was 81 percent (73 percent 86 percent), with non-progressors accounting for 17 percent. To predict the Chronic Kidney Disease, J. Stankovicn et.al [9] used 3 algorithms: neural network, random forest and KNN. They utilized dataset of 400 UCI patients with 24 attributes. The feature reduction process was done by the wrapper technique to identify the qualities that diagnose the illness with high accuracy.
Microsoft Azore was utilized to predict CKD patient status in a research conducted by K. Perera et. Al [10]. They compared four distinct algorithms by taking 14 out of 25 qualities into account: Multiclass Decision Jungle, Multiclass Decision Forest, Multiclass Neural Network and Multiclass Decision Regression. When they compared the outcomes, they discovered the Decision Forest performed the best with 99.1 percent precision. In their study, H. D. Mehr et.al [11] employed the SVM algorithm combined with 2 attribute selection methods: wrapper and filter which is to minimize the CKD dataset dimensionality, with 2 distinct assessments for all techniques.
The dataset utilized in the present case study is a tiny dataset which contains the minor imbalance problem. As a result, there are various worries about this dataset, including the overfitting problem or generalization problem, noisy data and imbalance. In their review, Yang .P et al.[12] found that the ensemble approach had the benefit of easing problem of data which consists of small size by taking average and combining over numerous models to lower the overfitting problem. Another investigation conducted by M. Pasha and M. Fatima [13] discovered that Support Vector Machine gave better precision in predicting CKD. In the previous studies, the researchers use single model to predict the CKD. But in the proposed system we used bagging ensemble techniques which are hybrid algorithm to predict the CKD with better accuracy than the existing system.

3. PROPOSED SYSTEM

Figure 1 depicts the suggested flow chart of prediction of kidney disease, which consists of several sub-modules such as model construction, ensemble model prediction, preprocessing of data, performance evaluation and testing. All modules demonstrate how it contributed to total precision of prediction algorithm. The rest of this study is organized around the recommended the architecture of chronic renal prediction.

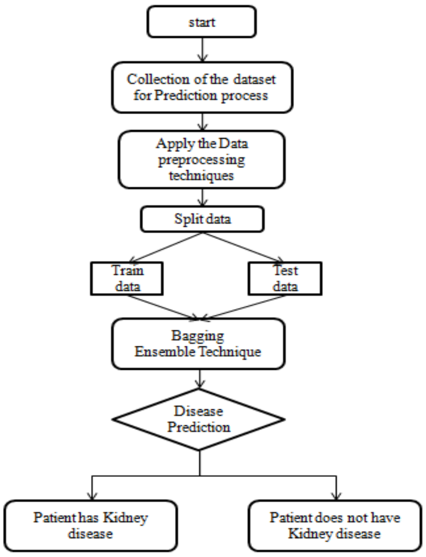

start
Collection of the dataset for Prediction process
Apply the Data preprocessing techniques
Split data
Train data
Test data
Bagging Ensemble Technique
Disease Prediction
Patient has Kidney disease
Patient does not have Kidney disease

Fig.1.Work Flow of Proposed system

3.1 Description of dataset

This study made use of a CKD dataset taken from the UCI repository[14] As indicated in Table 1, the dataset contains 400 cases of Chronic Kidney Disease victims with various indications & 25 characteristics with 14 nominal and 11 numeric values. The dataset is divided into two classes: non-CKD and CKD, which refer to individuals who do not have Chronic Kidney Disease and those who do have kidney disease.

S.No	Attributes	Description	Type
1	sg	Specific Gravity	Nominal
2	su	Sugar	Nominal
3	pc	Pus Cell	Nominal
4	ba	Bacteria	Nominal
5	bu	Blood Urea	Numerical
6	sod	Sodium	Numerical
7	hemo	Hemoglobin	Numerical

8	wc	White Blood Cell Count	Numerical
9	htn	Hypertension	Nominal
10	cad	Coronary Artery Disease	Nominal
11	bp	Blood Pressure	Numerical
12	al	Albumin	Nominal
13	rbc	Red Blood Cells	Nominal
14	pcc	Pus Cell Clumps	Nominal
15	bgr	Blood Glucose Random	Numerical
16	sc	Serum Creatinine	Numerical
17	pot	Potassium	Numerical
18	pcv	Packed Cell Volume	Numerical
19	rc	Red Blood Cell Count	Numerical
20	dm	Diabetes Mellitus	Nominal
21	appet	Appetite	Nominal

Table.1.Attributes of data set

3.2 Preparation of Data:

3.2.1 Handling of Noisy Data: The dataset of kidney disease had a significant number of missing values across the different characteristics. It is simply not possible to exclude samples or features that have missing values since significant information may be lost in the process. Backfilling is a method where adjacent value of the non-existing numeral is used to fill the blanks.

3.2.2 Splitting of data: The dataset of CKD is divided into two parts where 30% of the dataset is used to evaluate the machine learning methods and the remaining 70% is used to train the model.

3.2.3 Data Scaling: Some standard procedures were used to bring the features on to the same scales based on the respective intended values since the features present on the dataset because all of the features in the dataset do not reflect items collected and measured on the similar scale.

3.3 Model building

By utilizing seventy percentage of the complete dataset as a training set 3 various classifiers namely Logistic regression, SVM , and Decision Tree where training of data will be in order to form algorithms and their predictions were gathered. A majority voting rule that attempts to reduce the correlation between the estimators in the ensemble model by collecting random samples of features and training them rather than the entire feature set, an ensemble technique namely bagging combines the classifier predictions.

3.3.1 Decision tree

The divide-and-conquer strategy is used by decision trees to build classification decision-making rules issues using the ratio of information gain that overcomes the picking characteristics with multiple values[15] The purpose of prediction using decision tree method in data mining classifications learning is to figure out how to transfer an i/p variable x to an o/p variable c, in a labelled collection as:

The method learns a mapping function from i/p variable x to o/p variable c given a labelled collection of input & output pairs as:

$$D = \{(x,\ c)\}_{i=1}^{N} \qquad \text{Eq (1)}$$

The training set in Equation 1 the number of training data is N. In the most basic case, each x in the training set is a D-dimensional vector of integers known as features or characteristics. The class output variable is represented by c. The decision tree classifies data based on information gain, i.e. the estimation of the variations in the entropy between the previous and the next set. The set S for the entropy that is computed is divided by feature A.

As a result, while performing grouping, the characteristic with maximum information gain, considered as best classifier, and determined as:

$$Entropy\ H(S) = -\sum_{i=1}^{n} p(C) log_2 p(C) \qquad \text{Eq (2)}$$

Eq 2 represents the entropy of the current set of H,
C is the group of classe 0 and 1, and n is the no. of features.

As a result, the info gained by a trained dataset is defined as follow

$$Info\ Gain(S) = H(s) - \sum_{i^n}^{n} \frac{/s_i/}{/s/} H(s)$$

Eq (3)

Equation 3 represnts the information gain formula.
Here S denotes the no. of instances in partition S, S_i denotes the no. of instances in partitions i.

3.3.2 Support vector machine

Support vector machine is a supervised learning technique which is used for both classification and regression algorithms. However, it is mostly used in Machine Learning techniques for Classification problems[16].
The main goal of the SVM algorithm is to identify a hyperplane in an N-dimensional space that clearly classifies the input points. The size of the hyperplane is determined by the number of features[17].The SVM algorithm is illustrated in fig.2

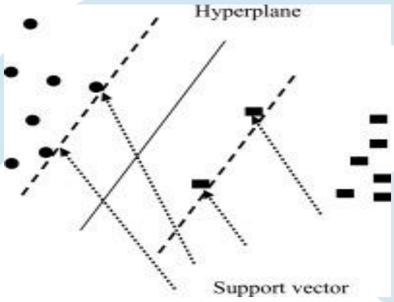


Fig.2.Support Vector Machine

3.3.3 Logistic Regression

Logistic regression is a Machine Learning method that belongs to the Supervised Learning approach. It is used to forecast the categorical dependent variable from a group of independent factors.

$$Logistic\ function = \frac{1}{1+e^{-x}}$$

Eq (4)

In the logistic function equation, x is the input variable. Let's feed in values −20 to 20 into the logistic function. As illustrated in Fig.3, the inputs are in range 0 and 1.

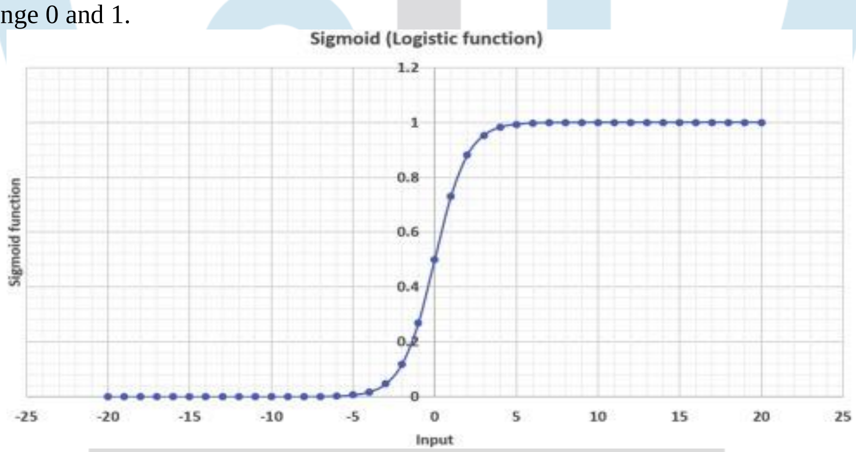


Fig.3.Logistic function

3.3.4 Bagging

The "Bagging" strategy utilized in this work to combine the performance of basic classifiers is a hybrid-algorithm technique which increases the accuracy of the classification algorithms [18].
The main goal behind the bagging is to accumulate predictions from classifiers of data mining (SVM, Logistic regression & Decision Tree) using the training data samples to produce new hypotheses. While the bagging technique, the base learners used in this study, SVM, Logistic regression and Decision Tree, which undergo training and generated from the sample of dataset of CKD, are coupled with bagging techniques using voting approach to build a best algorithm. This may be resolved with the use of Eq.

$$C_f = \underset{i}{\arg\max} \sum_{i=1}^{m} J_{ry_i}$$

$$and\ J_{ry_i} = \begin{matrix} 1, & if\, j_r = y_i \\ 0, & if\, j_r \neq y_i \end{matrix}$$

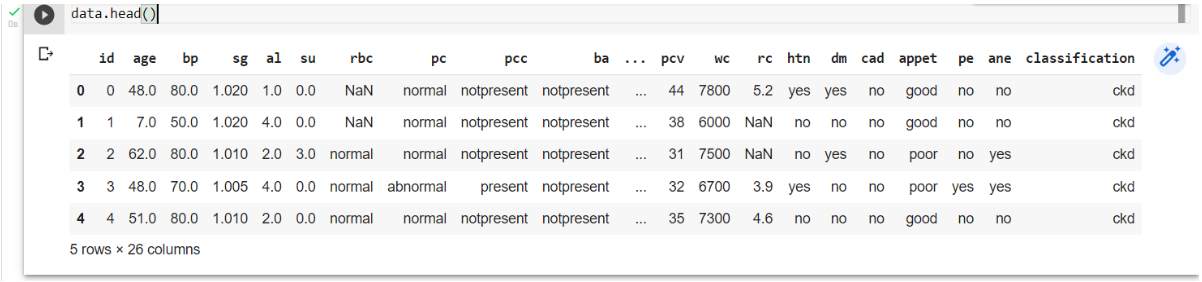
Eq (5)

Here Jr is the choice of rth classifier in class yi, Cf is the final prediction.

4. RESULTS AND DISCUSSION:

Data:

Data.head (n) gives the first n rows of the dataset. It has by default value of n=5 which it gives the top 5 rows of the dataset if the value of n is not given.The data of top 5 rows is represented as shown in Fig.4.



```
data.head()
```

```
   id  age    bp     sg   al   su     rbc        pc         pcc          ba  ...  pcv    wc   rc  htn   dm  cad  appet   pe  ane  classification
0   0  48.0  80.0  1.020  1.0  0.0     NaN    normal  notpresent  notpresent  ...   44  7800  5.2  yes  yes   no   good   no   no             ckd
1   1   7.0  50.0  1.020  4.0  0.0     NaN    normal  notpresent  notpresent  ...   38  6000  NaN   no   no   no   good   no   no             ckd
2   2  62.0  80.0  1.010  2.0  3.0  normal    normal  notpresent  notpresent  ...   31  7500  NaN   no  yes   no   poor   no  yes             ckd
3   3  48.0  70.0  1.005  4.0  0.0  normal  abnormal     present  notpresent  ...   32  6700  3.9  yes   no   no   poor  yes  yes             ckd
4   4  51.0  80.0  1.010  2.0  0.0  normal    normal  notpresent  notpresent  ...   35  7300  4.6   no   no   no   good   no   no             ckd
5 rows × 26 columns
```

Fig.4.Data

Heat Map:

Heatmaps assist in resolving this issue by using colors to represent data in a 2D table format. Values that are similar are assigned same colors, and a significant color change indicates a difference in data values. As a result, heatmaps assist data scientists in understanding which features or attributes, as well as high values or low values, corresponding to several groups. The following figure illustrates the heat map of the proposed model(Bagging Technique).

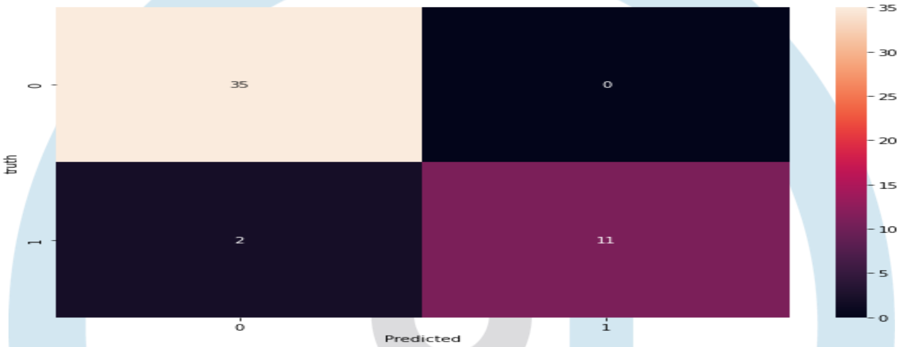
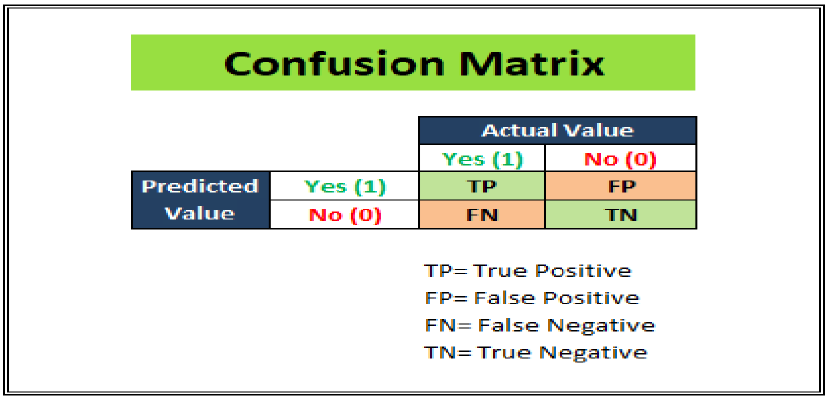


Fig.5.Heatmap of LR,DT,SVM

Fig.5 generates the heat map of the proposed system which it describes the terms Accuracy, Precision, Sensitivity, Specificity.



Confusion Matrix

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	TP	FP
	No (0)	FN	TN

TP= True Positive
FP= False Positive
FN= False Negative
TN= True Negative

Fig.6.Confusion matrix

Fig.6 illustrates the confusion matrix of MLmodel.The terms in the confusion matrix are described as follows:

Precision: It is defined as the no. of correct outputs supplied by the model or the percentage of all positive classes that the model is predicted should be true. The formula below can be used to evaluate it:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: The fraction of true positives correctly predicted by the model is known as sensitivity. True Positive Rate or sensitivity are other terms for it.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-measure: It's difficult to compare two models that have low precision but great recall, or vice versa. F-score can be used for this purpose. This score allows us to assess both recall and precision simultaneously. The F-score is maximum, when the precision and recall are equal. The formula below can be used to compute it:

F-measure=2*recall*precision/(recall+precision)

Sensitivity: The fraction of true positives correctly predicted by the model is known as sensitivity. True Positive Rate or Recall are other terms for it.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity: The fraction of true negatives correctly predicted by the model is known as specificity. True Negative Rate is another name for it (TNR).

$$Specificity = \frac{TN}{TN + FP}$$

Comparision Table:

Algorithm	Sensitivity	Specificity	Precision	Recall	Accuracy
LR	0.9714	0.846	0.944	0.9714	0.9375
NB	1	1	1	1	0.96
SVM	1	0.2307	0.777	1	0.7916
RF	1	0.8461	0.945	1	0.9583
KNN	0.9714	0.3076	0.7906	0.9714	0.791
LNS	1	0.8461	0.9459	1	0.9583
KDS	1	0.3076	0.8139	1	0.8125
KNS	1	0.3076	0.8139	1	0.8125
KRS	1	0.3076	0.8139	1	0.8125
KLD	1	0.8461	0.9459	1	0.9583
KLS	1	0.3076	0.8173	1	0.8125
RLD	1	0.923	0.9721	1	0.9783
RLN	1	0.9221	0.965	1	0.9786
RKN	1	0.9213	0.954	1	0.9773
SLN	1	0.913	0.97	1	0.9719
KLN	1	0.902	0.971	1	0.9789
LDS	1	0.8461	0.9459	1	0.9791

Table.2.Comparision Table of Machine learning Algorithms

Table.2 describes the difference between the Machine Learning Algorithms and the proposed system(LR,DT,SVM).

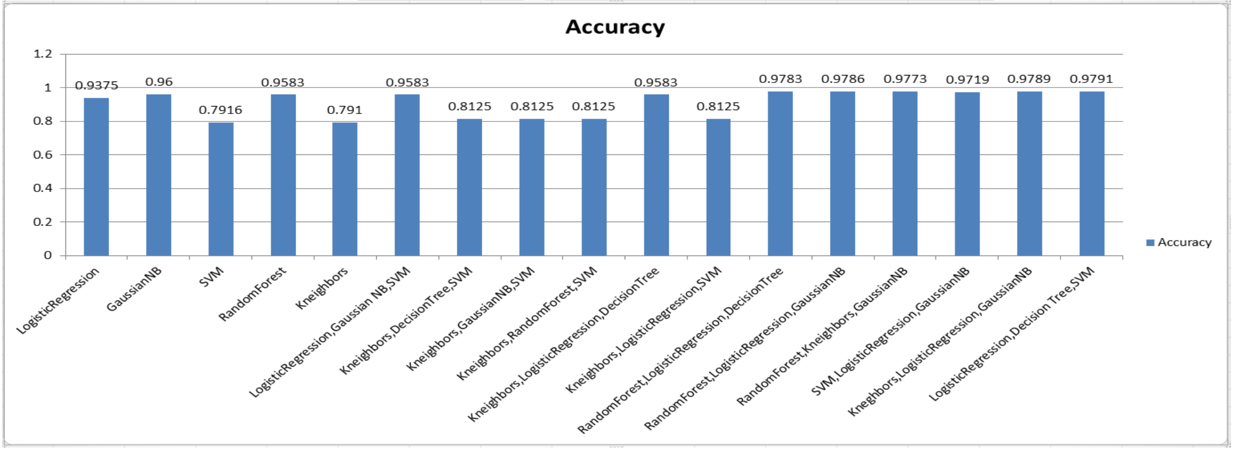


Fig.7.Vary in Accuracy of ML algorithms and bagging technique

Fig.7 shows the vary in accuracy of each model algorithm. The final result is the accuracy of the proposed model(Logistic Regression,SVM,Decision Tree)which attains an accuracy of 97.91%.

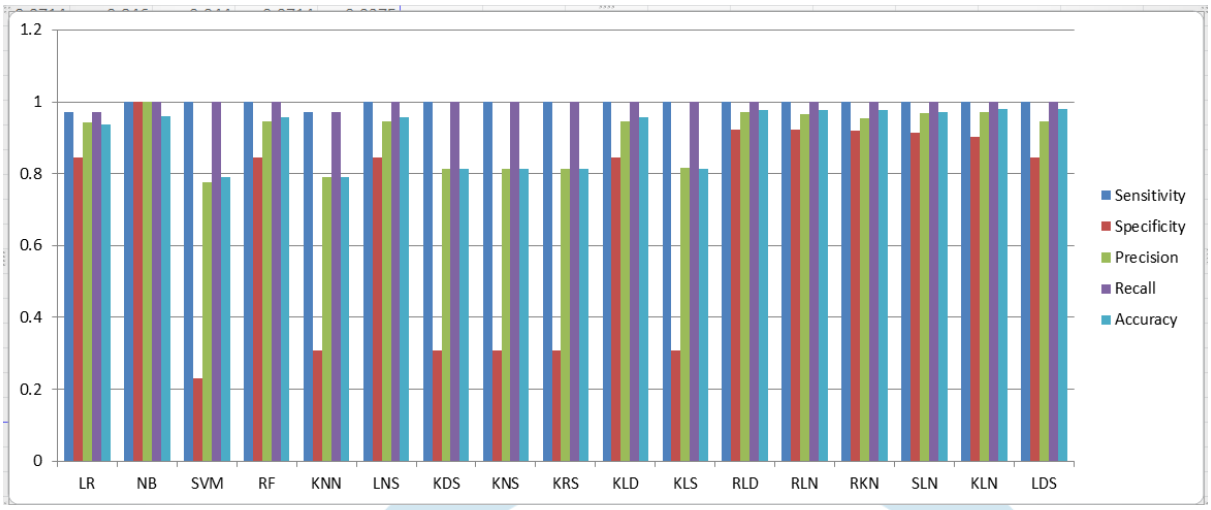


Fig.8.barplot of ML Algorithms

Fig.8 illustrates the bar plot of Machine Leaning Algorithms and bagging technique by using confusion matrix.

5. CONCLUSION:

The global frequency of CKD is alarming, and it is considered a vital hazard to real life. Two ensemble strategies and three base learners are presented in this paper to enlarge the precision of classification for Prediction of CKD. Because the lives of human are at stake, the model's precision is critical. The main criteria of ensemble techniques are to aggregating the results of several data mining algorithms (Logistic regression, Decision tree classifier and SVM). Python programming was used to implement the proposed technique. The proposed model's performance was 97.9% accuracy.

REFERENCES:

1. [Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, Chronic Kidney Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)]
2. [https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521]
3. [https://www.fda.gov/news-events/press-announcements/fda-approves-treatment-chronic-kidney-disease]
4. [Breiman, L. Bagging Predictors. Machine Learning 24, 123–140 (1996)]
5. [Bühlmann, Peter. (2012). Bagging, Boosting and Ensemble Methods. Handbook of Computational Statistics. 10.1007/978-3-642-21551-3_33.]
6. [Patil PM, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol.5, No.5, (2016), pp.135-141]
7. [T. Di Noia et al, "An end stage kidney disease predictor based on an artificial neural networks ensemble," Expert Syst. Appl., vol. 40, (11), pp.4438-4445
8. [H. S. Chase et al, "Presence of early CKD-related metabolic complications predict progression of stage 3 CKD: a case-controlled study," BMC Nephrology, vol. 15, (1), pp. 187]
9. [A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Healthcare Informatics(ICHI),IEEE International Conference]
10. [W. Gunarathne, K. Perera and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Bioinformatics and Bioengineering (BIBE),IEEE 17th International Conference].
11. [H. Polat, H. D. Mehr and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med.Syst., vol. 41, (4), pp. 55]
12. [P. Yang et al, "A review of ensemble methods in bioinformatics," Current Bioinformatics, vol. 5, (4), pp. 296-308]
13. [M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, (01), pp. 1]
14. [UCI machine learning repository. (2017):Chronic kidney disease dataset.Retrieved march 2017, fromhttp://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_disease].
15. [M. Oded, R. Lior Data mining and knowledge discovery handbook. Introduction To Knowledge Discoveries and Data Mining Springer, Isreal (2010), pp. pp.1-pp15].
16. [https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm]
17. [https://www.geeksforgeeks.org/support-vector-machine-algorithm/]
18. [L. Breiman Bagging predictors Mach. Learn., 24 (2) (1996), pp. 123-140].