Chronic Kidney Disease Prediction using Data Mining and Machine Learning

Adeeba Azmi[1], Amiksha Hingu[2], Ruchi Dholaria[3], Ms. Alvina Alphonso[4]

[1,2,3]*Student, Information Technology, St. Francis Institute of Technology*
[4]*Assistant Professor, Information Technology, St. Francis Institute of Technology*

***Abstract*— The aim of this paper is to predict the chronic kidney disease by entering the symptoms. The proposed paper uses Data Mining and Machine Learning techniques to predict results. The techniques used to predict the disease are KNN, SVM Ensemble.
For Data Mining SVM with rbf kernel was used to predict the result. And in Machine Learning KNN with hyperparameter was used to predict the result. We used ensembling technique for greater accuracy for Machine Learning. The proposed solution gives accuracy of 87% in Data Mining and above 92% in Machine Learning. For which the dataset "CKD" is provided which has 400 columns and 24 attributes.

***Keywords*— Data Mining, Machine Learning, Chronic Kidney Disease, KNN, SVM, Ensemble.**

1. INTRODUCTION

Chronic kidney disease (CKD) is the serious medical condition where the kidneys are damaged and blood cannot be filtered. In the end-stage of the disease the renal disease(CKD), the renal function is severely damaged. The starting date of kidney failure may not be known, it may not recognize as an illness of the patient because it cannot show any symptoms initially. And this chronic kidney disease is also called chronic renal failure, which has become quite a serious problem in the world where the kidneys are damaged and it has become the cause of improper function of kidney organ. The main factors contributing to the cause for the disease are: High blood pressure, Diabetes, Cardiovascular (heart and blood vessel) disease, hereditary records on kidney failure. Data mining and Machine Learning is doing significant research in the field of medical science as there is a requirement of well-organized methodologies for analyzing, predicting and detecting diseases. According to the Global Burden of Disease study in 2010, chronic kidney disease was ranked 27th in the list for the cause of total number of deaths worldwide in 1990, but rose to 18th in 2010[8]. Nearly 75% of cases for chronic kidney disease was noted in the year 2014-2016.

So there are various techniques present in the data mining and machine learning such as KNN, SVM, ANN, Naive Bayes, J48, etc. In the work, when tried using Naive Bayes and ANN algorithm didn't gave proper accuracy. In our proposed work KNN, SVM and Ensemble technique is used to classify data, for feature selection to give accuracy percentage of a specific dataset.

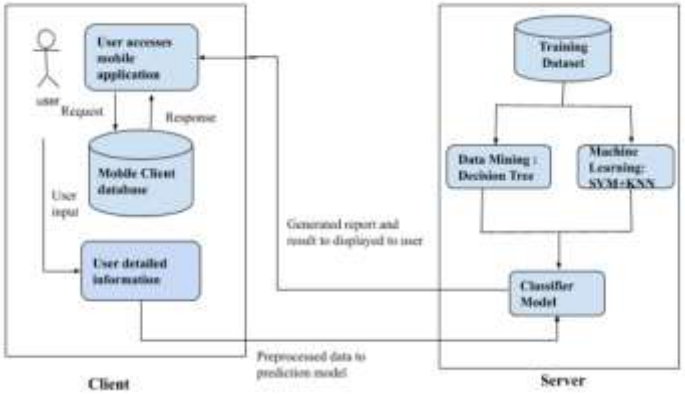

Fig 1. Client Server Design of CKD System

Figure 1 specifies the Client Server Design of the CKD System in which user feeds all the details to the system, after which all these details are stored in the database, where the dataset fed is pre-processed in which data cleaning is done, feature selection is done and after doing all this process classifier model is applied using Data Mining and Machine Learning. In Data Mining, decision tree is used and in Machine Learning SVM+KNN is used. The output from both i.e. Data Mining and Machine Learning is given as report.

2. LITERATURE REVIEW

U.N.Dulhare et.al [1] their work stated that Chronic Kidney Disease is a disease whose symptoms are not easily noticeable if present. They stated that GFR i.e. Glomerular Filtration Rate when calculated provides best standard for predicting Chronic Kidney Disease (CKD). And for extracting action rules for Chronic Kidney Disease (CKD) they have divided the work in 5 different phases. The action rules describes the action to be taken depending the on Glomerular Filtration Rate (GFR) calculated and the Chronic Kidney Disease (CKD) stage patient belongs to. The results they concluded is that by using Wrapper Suset Eval attribute evaluator with Naive Bayes classifier which selects only 6 attributes from 25 attributes with 52% of attributes reduction. One R attribute evaluator with Naive Bayes classifier selects only 5 attributes from 25 attributes with 80% attribute reduction. 85% accuracy was achieved when Naïve Bayes classifier was used on original dataset where as 97.5%was achieved using reduced dataset.

Tabassum BG et.al [2] aims the way toward changing over a lot of unstructured crude information, recovered from various sources to an information item valuable for associations shapes the core of Big Data Analytics Big Data

Healthcare in Healthcare using Data Mining techniques is mainly useful in medical discipline where no availability is there for the proof of favouring a selected treatment alternative is located. A Huge volume of complicated information is created constituting patients data, disease record, hospitals bills, medical equipments, insurance claims, treatment price and so on. The system only uses two algorithm of class and one for clustering which provides accuracy upto 75%.The system is not able to predict stages of disease based on given data.

Sirage Zeynu et.al [3] the objective is to analyse and predict chronic kidney disease (CKD) by discovering the hidden pattern of the relationship that is directly related with CKD by using feature selection and data mining classification techniques like k-nearest neighbor (KNN), artificial neural network (ANN), decision tree. The work shows that classification and feature selection based methods for enhancing the performance accuracy of the algorithm for effective analysis and prediction of chronic kidney disease. The artificial neural network is a complex algorithm and requires long time to train the dataset. The performance of the prediction system can be enhanced by ensembling different classifier algorithms.

Salekin and J.Stankovic [4], authors have developed an automated machine learning solution to detect Chronic Kidney Disease and explore 24 parameters related to kidney disease. The dataset used for evaluation consists of 400 patient data and the dataset suffers from noisy and missing data. We need a robust classifier that can deal with these issues. Hence, we evaluate solutions with three different classifiers k-nearest neighbor, random forest and neural nets. The advantage of this research work is that it uses random forest classifier and wrapper method which gives more accurate result of identify whether a person is having CKD or not. New factors can be used by the classifiers for more accurate detection chronic disease than the state of art using formulas.

Guneet Kaur et.al [5] the aim of this paper is to predict or detect chronic kidney disease, KNN (K-nearest neighbour) and SVM (Support Vector Machine) data mining algorithms are used.

Elapsed time: It is observed that KNN and SVM takes more time to predict the chronic kidney dataset having value 0.9817 whereas SVM has 0.8045 elapsed time and time taken by KNN is 0.12558.

Accuracy: It is observed that SVM data mining classifier gives 99.29 accuracy which is more as compared to KNN and KNN and SVM where KNN acquires 97.83 accuracy and KNN and SVM together acquires 98.93 accuracy.

Luyckx, Valerie A et al. [6], in this paper Apriori is used which is a type of candidate generation algorithm proceeds in a level-wise order. A neural network forms an interconnected group of artificial neurons to processes information and computation is done using associated weights. Also, Hierarchical clustering technique focuses on the principle of decomposing databases either in a top-down or bottom-up fashion and K-Medoids Clustering is used. It does by eliminating the sensitivity using medoid as a measure for similarity by choosing the mean value of objects in a cluster, the centrally located object within a cluster is chosen.

S.Dilli Arasu [7], states that a Naive Bayes classifier is a probabilistic classifier that works on the concept of Bayes theorem, Support vector machine is a machine learning technique that works on statistical learning theory. Also, KNN is used which is a distance based algorithm that is applicable when all the attribute values are continuous. It can be modified according to categorical attributes. Decision Tree Based Algorithm is used. This approach works well in cases where a tree is constructed to model the classification process when the classification becomes complicated. If there are more instances in the training set it applies the same principle to classify the k nearest neighbour. Thus, this makes it efficient and scalable to perform mining. Design to take symptoms as input in order to predict the disease based on old patients record.

3. METHODS USED

K-NEAREST NEIGHBOUR:

K-Nearest Neighbors is one of the important classification technique in Machine Learning. KNN algorithm comes in supervised learning domain and usually finds application for pattern recognition, data mining and intrusion detection. Since KNN is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data. It is the kind of distance based algorithm which means when all attribute values are continuous, it may be modified according to categorical attributes. We are given some prior data (also known as training data), which usually classifies coordinates into groups identified by an attribute.

The below figure i.e. Figure 2 KNN example[10] shows that similar data points are close to each other. This algorithm captures the idea of proximity with some mathematics calculating the distance between the points in the graph. This algorithm is very easy to implement, also there is no need to build a model tuning several parameters or making assumptions.

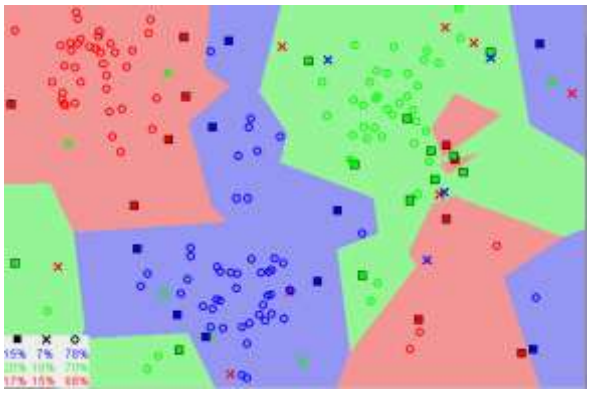

Fig 2. KNN example

SUPPORT VECTOR MACHINE

Support vector machines (SVMs, also support vector networks) is one of the most popular supervised learning algorithm which can be used for both classification as well as regression. analysis. It is a discriminative classifier defined by a separating hyperplane.

This algorithm can solve linear and non-linear problems in the practical world. The main aim of the SVM model is to create a best line or decision boundary which can segregate n-dimensional space into class so that we can easily put the new data point into the correct category in future. And the best decision is known as hyperplane. When the SVM chooses the extreme points/vectors for creating the hyperplane, then those extremes are called support vectors. For the below diagram i.e. Figure 3 Possible Hyperplane[11], there are many possible hyperplanes that could be chosen for separating the classes of datapoints. But our main objective is to find a plane having maximum margin. As maximizing the margin makes it possible for adding future data points so that those can be classified with more confidence.

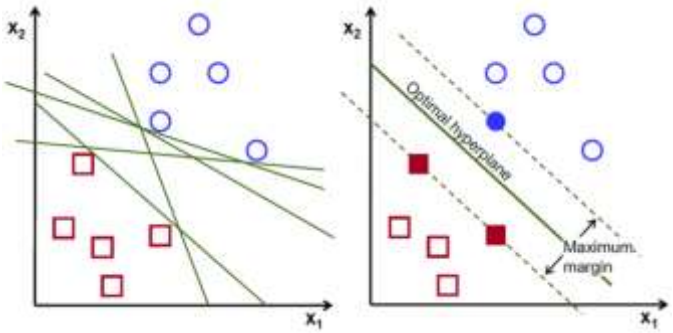

Fig 3. Possible Hyperplane

HYPERPLANE

Hyperplane are the decision making boundaries to help classify the data points. And the data points falling on the either side of the hyperplane can be categorized as different classes. The dimension of hyperplane is based on the number of features. So, if the number of input feature is 2, then hyperplane is just a line. If the number of input feature is 3, then hyperplane is two-dimensional. It can become very difficult to imagine if the number of features exceeds above three.

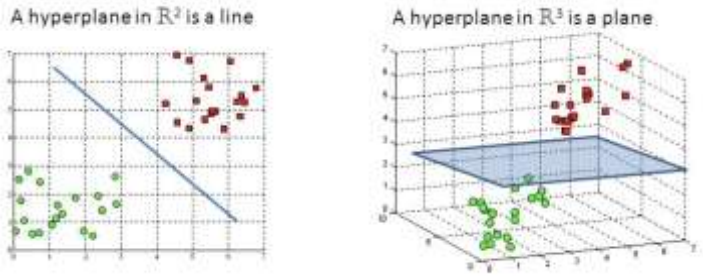

Fig 4. Hyperplane in SVM

The above diagram i.e. Figure 4 Hyperplane in SVM[12]. For two dimensions there is the separating line was the hyperplane. Similarly, for three dimensions a plane with two dimensions divides the 3d space into two parts and thus act as a hyperplane.

OPTIMAL HYPERPLANE

The below figure i.e. Optimal Hyperplane example[12], shows the hyperplane guaranteeing the best generalisation performance is the one with maximal margin of separation between two classes. This is known as the optimal or maximal margin hyperplane and is unique. For calculating the margin, two parallel hyperplanes are to be constructed on either side of the hyperplane, which are "pushed up against" the two data sets. So a good separation is achieved that has the largest distance to the neighbouring datapoints of both classes by the hyperplane.

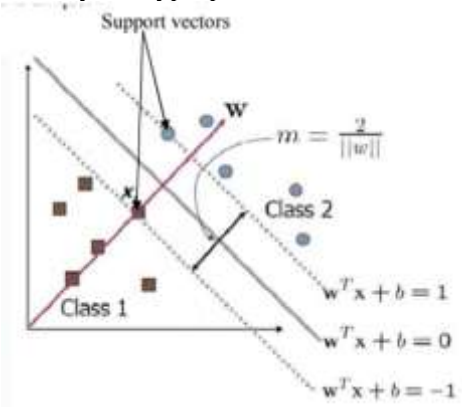

Fig 5. Optimal Hyperplane example

ENSEMBLE

Ensemble method comes under Machine Learning techniques combining several base models in order to produce one optimal predictive model. The main principle of the ensemble technique in machine earning is that a group of weak learners come together to become a strong learner, to increase the accuracy of model. So whenever we try to predict the target variable using any techniques in machine learning the main cause of difference in predictive and noisy values will be noisy , biased and variance. Ensemble method reduces these factors except for noise.

Below figure i.e. Figure 6 Ensemble learning types[9] shows the ensemble learning types (bagging, boosting and stacking).

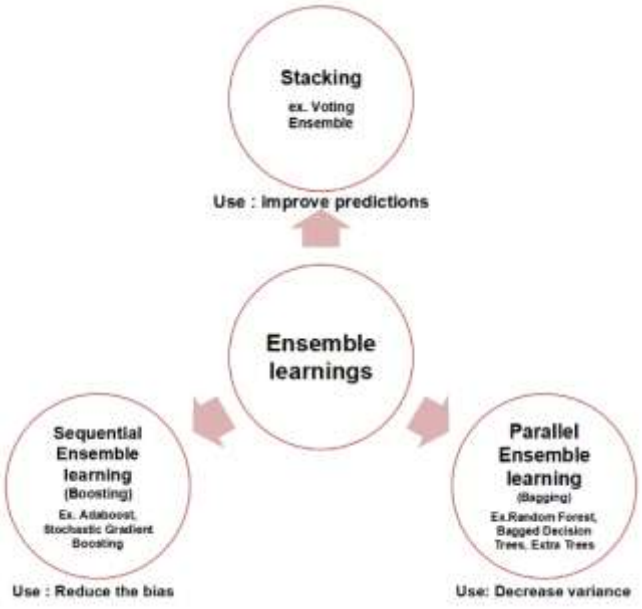

Fig 6. Ensemble learning types

4. IMPLEMENTATION

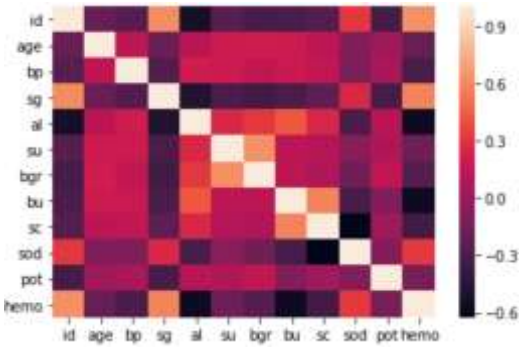

Fig 2. Classification using heatmap

Figure 2 specifies the classification using heatmap. A heatmap is a good visualization technique to compare any 2 features with respect to the values. The output of it represents that lighter the color of that attribute the greater it is affected and here age and haemoglobin are taken against other features.

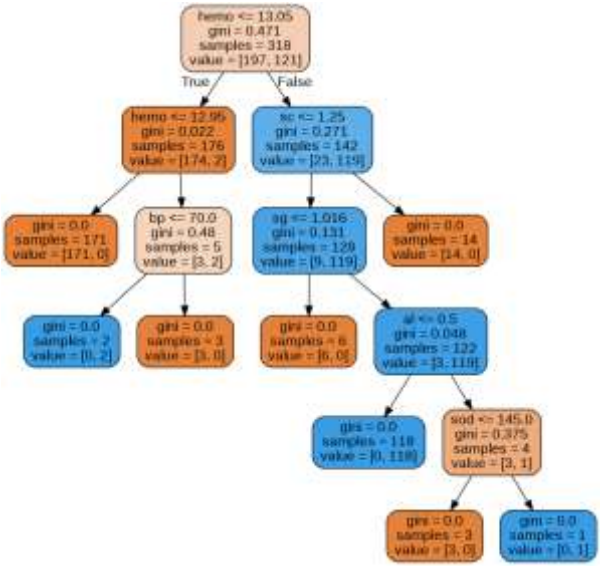

Fig 3. Visualization Graph of CKD System

Figure 3 specifies the Visualization Graph of the CKD System. Here a tree is built where the condition/ feature is selected on each split on the basis of information gain or gini impurity value.

```
clf = SVC(kernel='rbf')
clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
print(accuracy_score(y_test,y_pred))
```

```
0.875
```

Fig 4. SVM using rbf kernel of CKD System

Figure 4 specifies the SVM using rbf Kernel. Radial Basis Function(rbf) is popular method used in SVM Model. It is a function whose value depends on the distance from the origin or some other point. The output that we got from this function is 0.875, higher the value of the output more accurate it is.

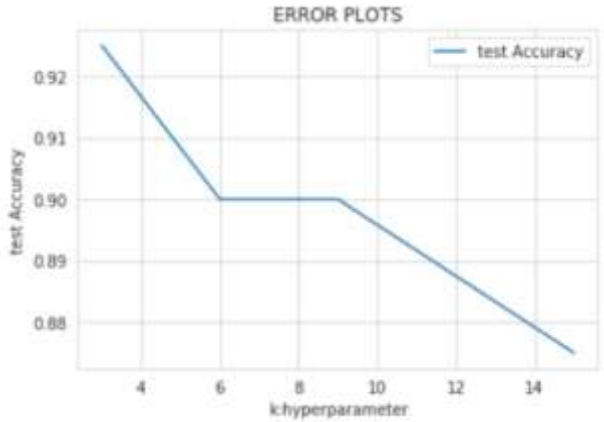

Fig 5. KNN with hyperparameter

Figure 5 specifies the KNN with hyperparameter of CKD System. As the performance of the model may not give accurate result, so with the use of hyperparameter tuning with the KNN model improved the performance by 20%. So the output that we got for chronic Kidney Disease Prediction went above 92%.

5. CONCLUSIONS

The objective of this work is predict the disease using data mining and machine learning techniques. So the chronic kidney disease can be very well in this, we took a dataset of 400 patients, 250 among them have early stage of CKD. This dataset contain some noisy and missing values. We evaluated them with classification using heatmap, visualization graph, SVM using rbf kernel and KNN with hyperparameter. For Data mining SVM using rbf kernel was used which gave result of 87% and for Machine Learning KNN with hyperparameter was used giving result above 92%. Thus, we can state that machine learning can be really helpful in healthcare sector to help predict Chronic Kidney Disease.

REFERENCES

[1] U.N.Dulhare and M.Ayesha, "Extraction of action rules for chronic kidney disease using Naive bayes classifier", IEEE International Conference on Computational Intelligence and Computing Research(ICCIC), Chennai,2016.

[2] Tabassum BG, Mamatha Majumdar, Jharna Majumdar, "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), Vol 4, Issue 9, September 2017, 10.13140/ RG.2.2.26856.72965.

[3] Sirage Zeynu, Shruti Patil, "Survey on Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques and Feature Selection", International Journal of Pure and Applied Mathematics,2018.

[4] Salekin and J.Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," IEEE International Conference on Healthcare Informatics (ICHI), Chicago, 2016, IL, 2016, pp.262-270.

[5] Guneet Kaur, Ajay Sharma, "Predict Chronic Kidney Disease using Data Mining algorithm in Hadoop", International Journal of Advances in Electronics and Computer Science, 2018.

[6] Luyckx, Valerie A et al. "The global burden of kidney disease and the sustainable development goals." Bulletin of the World Health Organization, Vol. 96, 2018.

[7] S.Dilli Arasu, Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining Techniques," International Journal of Applied Engineering Research ISSN09734562,Vol. 12, No. 23, 2017.

[8] Jayalakshmi V, Lipsa Nayak, K.Dharmarajan, "A Survey on Chronic Kidney Disease Detection Using Novel Methods, "International Journal of Pure and Applied Mathematics, Vol.119, No. 10, 2018.

[9] https://medium.com/ml-research-lab/ensemble-learning-the-heart-of-machine-learning-b4f59a5f9777

[10] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[11] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[12] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[13] https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989