Predictive Analytics Of Chronic Kidney Disease By Using Machine Learning

K.Samatha
Department of CSE
K L Education Foundation, Green Fields,Vaddeswaram, Guntur Dist, Andhra Pradesh, India - 522502
170030531@kluniversity.in

M.Rohitha Reddy
Department of CSE
K L Education Foundation, Green Fields,Vaddeswaram, Guntur Dist, Andhra Pradesh, India - 522502
170030842@kluniversity.in

P.Faizal Khan
Department of CSE
K L Education Foundation, Green Fields,Vaddeswaram, Guntur Dist, Andhra Pradesh, India - 522502
170031003@kluniversity.in

R. Akhil Chowdary
Department of CSE
K L Education Foundation, Green Fields,Vaddeswaram, Guntur Dist, Andhra Pradesh, India - 522502
170031124@kluniversity.in

P.V.R.D Prasada Rao
Department of CSE
K L Education Foundation, Green Fields ,Vaddeswaram, Guntur Dist, Andhra Pradesh, India – 522502
pvrdprasad@kluniversity.in

***Abstract*— Kidney diseases are increasing day by day among people. It is becoming a major health issue around the world. Not maintaining proper food habits and drinking less amount of water are one of the major reasons that contribute this condition. With this, it has become necessary to build up a system to foresee Chronic Kidney Diseases precisely. Here, we have proposed an approach for real time kidney disease prediction. Our aim is to find the best and efficient machine learning (ML) application that can effectively recognize and predict the condition of chronic kidney disease. We have used the data from UCI machine learning repository. In this work, five important machine learning classification techniques were considered for predicting chronic kidney disease which are KNN, Logistic Regression, Random Forest Classifier, SVM and Decision Tree Classifier. In this process, the data has been divided into two sections. In one section train dataset got trained and another section got evaluated by test dataset. The analysis results show that Decision Tree Classifier and Logistic Regression algorithms achieved highest performance than the other classifiers, obtaining the accuracy of 98.75% followed by random Forest, which stands at 97.5%**

Keywords - SVM, KNN, Logistic Regression, Decision Tree, Random Forest, Kidney Diseases

I. INTRODUCTION

Kidney Disease is a condition in which the functioning of the kidneys decreases gradually. These diseases will cease their ability to keep a person healthy. Kidney's filter the blood by removing excess waste from them and they are excreted from the body in the form of urine. When kidney disease reaches advanced stage, it will become fatal for a person unless he undergoes kidney transplant or dialysis.

Factors like diabetes, hypertension and heart diseases contribute to the development of chronic kidney disease. If a family line has a history of kidney failure, it also may contribute to kidney disease. Symptoms of kidney disease include vomiting, weight loss, vomiting and loss of appetite. Prediction of kidney diseases at early stages can help in avoiding major damage. To predict this, we need to obtain details on some indices which can relate well to kidney disease. Our aim is to predict the kidney disease, by analysing the data on those indices and using five classification techniques of machine learning for prediction and selecting the one which give us the maximum rate of accuracy to predict the disease. The five classification techniques are K- Nearest Neighbours Classifier, Support Vector Classifier (SVC), Decision Tree Classifier (DT), Random Forest Classifier (RF), and Logistic Regression.

MACHINE LEARNING CLASSIFIERS:

These are used to predict the class/target/labels/categories of a given data points. Classification belongs to the category of supervised learning in which the targets are provided with input data. They are used in many applications like medical diagnosis, spam detection, target marketing etc. They use a mapping function (f) from input variables (X) to discrete output variables(Y).

DECISION TREE CLASSIFIER:

Decision Tree Algorithm: This is a type of predictive modelling algorithm employed mainly for statistics, data mining and for classification and regression problems in machine learning. It has a flowchart type structure containing internal nodes, leaf nodes and branches. The internal nodes, leaf nodes represent and branches represent test on features, class label, conjunctions of features which lead to the class labels respectively. The classification rules are represented by the path from the root to the leaf.

RANDOM FOREST CLASSIFIER:

It is an extension of decision tree algorithm. It is a supervised learning algorithm which is mainly used for classification problems. This algorithm employs different decision trees on the dataset and chooses the best prediction among the outputs produced by those trees. The process of choosing the result is done by voting. The prediction with the most votes is the output of the algorithm.

KNN (K-Nearest Neighbors) CLASSIFIER:

It is a simple non – parametric algorithm. It is also known as lazy learner algorithm as it does not learn anything from the training set during the training phase, it just stores all the training data instead. During the testing phase or during classification, it assumes a similarity of a data point with a group of the stored data, i.e., it will categorize the new datapoint into a class of data points which are most similar to the present data point. When a new data point is given, it selects K number of neighbours and calculates the Euclidian distance between those neighbours and the point. It counts the number of datapoints in each category and assigns the new data points that category for which the number of the neighbour is maximum.

LOGISTIC REGRESSION:

It is the most popular supervised learning technique. Given a set of independent variables, it predicts the output of a categorical dependent variable. If a data point is given, the output will be a number which represents the probability of that data point belonging to a specific class. It is similar to Linear regression, but instead of a line, an 'S' shaped curve is fitted here and it is used for classification problems.

SUPPORT VECTOR CLASSIFIER:

It is one of the most used classifier algorithms. It creates a decision boundary which is also called as the best line to separate data points in an n-dimensional space into different classes. This is done by choosing extreme data points or vectors to generate the hyperplane, that is why, it is known as a support vector machine. Whenever a new data point is given, it will add that data point in the most suitable category in the future. There are two types of SVM classifiers which are Linear and Non-Linear SVM's. Linear SVM's are used when the data belongs to only two classes and they can be separated with a single line. In all the other cases, Non-Linear SVM's are used.

classification rules are represented by the path from the root to the leaf.

II. RELATED WORK

In [1] the authors worked on improved prediction algorithms on the data of chronic cerebral in fraction disease. They found that when the data is incomplete, the accuracy of a model decreases, using structured and unstructured data from hospital, they designed a (CNN)-based multimodal disease risk prediction algorithm. They also used latent factor model to reconstruct the unknown data. Their algorithm was 98.4% accurate in its predictions and it had more convergence speed than existing CNN-based unimodal disease risk prediction algorithm.

In [2] the authors implemented decision tree by using both ID3 which is implemented by the use of information gain and gain ratio and evolutionary algorithm, which is implemented with fitness proportionate and rank as their selection strategies. Their results showed that the ID3 algorithm performed well than the evolutionary algorithm.

The authors in [3] found that while using the KNN classifier, the computational burden on the CPU increases polynomially. As the size of the data increases. They showed that the use of the NVIDIA CUDA API accelerates the search for the KNN up to a factor of 120. In [4], the authors analyzed and surveyed different machine learning models such as SVM, KNN, DT etc. The authors in [5] compared SVM, RF and ELM algorithms for the detection of intrusion in a secured network. Their results showed that ELM outperforms all the techniques used by them.

III. WORKING

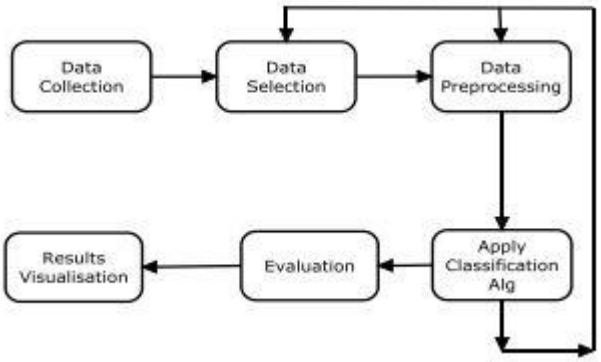

Fig: Work Flow

Our System has five stages In Kidney Disease prediction. Each stage is explained below.

Data collection:

- First, the data was obtained from the UCI Machine Learning Repository. It had data on age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anaemia and class.

- Next, the shape of the dataset was obtained and it was found to be 400X26, i.e. 400 records and 26 features were given.

- In the given dataset, classification was the target and the others were predictors.

```
[4]: data.shape

[4]: (400, 26)
```

Fig: Shape of the Dataset obtained

Data pre-processing:

- It was found that some categorical data was missing in the dataset.
- The index id did not have any effect on the disease, so it was dropped.
- A set of operations were performed to remove the presence of the missing categorical values and then all the values in it were converted into numerical.

Data analysis:

- First, a heat map of the categories was generated.
- It was found that sg, hemo, sod, pcv and rc indices are negatively correlated with classification, i.e. with their decrease, the disease was found to be decreasing.
- Next, boxplots were grouped for al, htn and dm indices by the classification.
- Next, Histograms were plotted for all the indices.
- Then, the data was normalized using minmax function.

```
import seaborn as sns
corr = df.corr()
sns.heatmap(corr,xticklabels = corr.columns,yticklabels = corr.columns)
<matplotlib.axes._subplots.AxesSubplot at 0x1d8e38c2978>
```

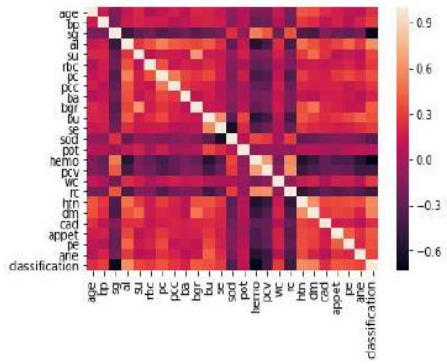

Fig: Heat map of the dataset

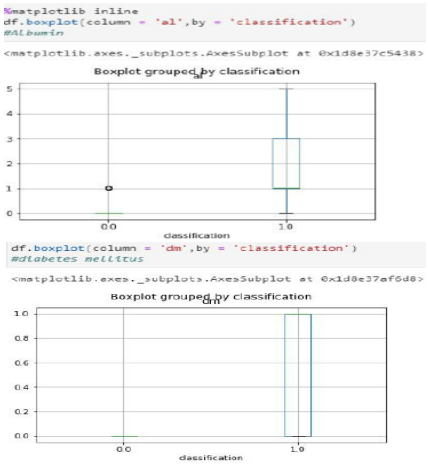

Fig: Boxplot of al and dm grouped by classification

Fig: Histograms of all the indices in the data

Applying Algorithms:

- The dataset is split into Training and Testing Dataset.
- Five models viz Decision Tree, KNN, SVC, Logistic Regression and Random Forest models are trained with the training data.
- The normalized data previously generated, is given as input for the model for training.

Predicting the Data:

- All the classifiers are separately trained with the training data.
- The training and testing data are labelled datasets.
- After Training, these algorithms are tested for accuracy score and cross validation score using the test data.

- Then, these scores are compared with that of training data.

- For evaluating the performance of models, we have used the confusion matrix to calculate accuracy, roc score, recall score, sensitivity and specificity.

IV. RESULTS

The Following Results have been obtained from the evaluation of the five algorithms on the test data.

TABLE I

Algorithm	**TP**	**FP**	**TN**	**FN**
KNN	25	4	32	19
LR	28	1	51	0
SVC	0	29	51	0
DTC	29	0	50	1
RF	28	1	50	1

Values Obtained for confusion matrix using different algorithms

TABLE II

Algorithm	**Accuracy**	**Recall Score**	**ROC Score**	**Sensitivity**	**Specificity**
KNN	71.25%	62.74%	74.47%	0.5681	0.8888
LR	98.75%	100%	98.27%	1.0	0.98
SVC	63.75%	100%	50%	Nan	0.6375
DTC	63.75%	98.03%	99.01%	0.97	1.0
RF	97.5%	98.3%	97.29%	0.96	0.98

The Accuracy, recall score, ROC score, Sensitivity and Specificity of the algorithms used

V. CONCLUSION AND FUTURE SCOPE

It had been found that the decision tree algorithm, logistic regression and random forest algorithms are more efficient in the prediction of chronic kidney diseases. Their accuracy was found to be 98.75%, 98.75 and 97.5% respectively. In the future, this work can be upgraded by building up a web application based on these algorithms and using a bigger dataset when contrasted with the one utilized in this examination. This will help in giving better outcomes and help healthcare experts in the prediction of kidney diseases adequately and productively.

REFERENCES:

References
[1]Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE Access 2017.

[2] V. Mohan, “Decision Trees: A comparison of various algorithms for building Decision Trees,” Available at: http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf

[3] Garcia, Vincent & Debreuve, Eric & Barlaud, Michel. (2008). Fast k Nearest Neighbor Search using GPU. CVPR Workshop on Computer Vision on GPU. 10.1109/CVPRW.2008.4563100.

[4] V V. Ramalingam ,Ayantan Dandapath, M Karthik Raja,” Heart Disease Prediction using Machine Learning Techniques: A Survey”, 7(2.8): p. 684-687 ,October 2018

[5] I. Ahmad, M. Basheri, M. J. Iqbal and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," in IEEE Access, vol. 6, pp. 33789-33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[6] Swethalakshmi, H., et al. Online handwritten character recognition of Devanagari and Telugu Characters using support vector machines. 2006.

[7] Al-Talqani, H.M., Dyslipidemia and Cataract in Adult Iraqi Patients. EC Ophthalmology, 2017. 5: p. 162-171.

[8] McKinley, R., et al., Fully automated stroke tissue estimation using random forest classifiers (FASTER). Journal of Cerebral Blood Flow & Metabolism, 2017. 37(8): p. 2728-2741.

[9] Jos Timanta Tarigan, C.L.G., Elviawaty Muisa Zamzami, A REVIEW ON APPLYING MACHINE LEARNING IN GAME INDUSTRY International Journal of Advanced Science and Technology, 2019-09-27 28(2).

[10] Chronic Kidney Disease Data Set, Available from: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[11] J. Davis and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves,” Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006.

[12] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, “A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[13] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The big data revolution in healthcare: Accelerating value and innovation,” 2016.

[14] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[15] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” Nature Reviews Genetics, vol. 13, no. 6, pp. 395–405, 2012.

[16] J. C. Ho, C. H. Lee, and J. Ghosh, “Septic shock prediction for patients with missing data,” ACM Transactions on Management Information Systems (TMIS), vol. 5, no. 1, p. 1, 2014.

[17] “Ictclas,” http://ictclas.nlpir.org/.

[18] “word2vec,” https://code.google.com/p/word2vec/.

[19] Y.-D. Zhang, X.-Q. Chen, T.-M. Zhan, Z.-Q. Jiao, Y. Sun, Z.-M. Chen, Y. Yao, L.-T. Fang, Y.-D. Lv, and S.-H. Wang, “Fractal dimension estimation for developing pathological brain detection system based on minkowski-bouligand method,” IEEE Access, vol. 4, pp. 5937–5947, 2016.

[20] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, and A. Marinez, “Dynamic hierarchical classification for patient risk-ofreadmission,” in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015, pp. 1691–1700.

[21] Sak Haim, Andrew Senior, Franoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition", 2014.

[22] D. W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley Interscience, 2000.

[23] C.J.C. Burges, "Simplified Support Vector Decision Rules", Proc. 13th Int'l Conf. Machine Learning, pp. 71-77, 1996.