Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Chronic Kidney Disease Prediction Using Classifiers

Dr.Mary Magdalene Jane [1], Mr.Selva Vignesh.M [2],Miss.Tharani.R[3]

Department of Computer Science with Data Analytics

Dr.N.G.P. Arts and Science College

***Abstract*—** In today's era people are health conscious and pay more attention to health in spite of their workload and busy schedule. The field of biosciences have advanced to a larger extent and have generated large amounts of information from Electronic Health Records. This has given rise to the acute need of knowledge generation from this enormous amount of data. Data mining methods and machine learning play a major role in this aspect of biosciences. Chronic Kidney Disease (CKD) is a condition in which the kidneys are damaged and cannot filter blood. A family history of kidney diseases or failure, high blood pressure, type 2 diabetes may lead to CKD. This is a lasting damage to the kidney and chances of getting worse by time is high. The very common complications that results due to a kidney failure are heart diseases, anemia, bone diseases, high potassium and calcium. The worst-case situation leads to complete kidney failure and necessitates kidney transplant to live. An early detection of CKD can improve the quality of life to a greater extent. This calls for good prediction algorithms to predict CKD at an earlier stage. A wide range of machine learning algorithms are employed for the prediction of CKD. This work implements techniques like data pre-processing, exploratory analysis of data, model optimization, and by using various classifiers such as Logistic Regression, Naive Bayes and K-Nearest Neighbors (KNN) to predict CKD, the best classifier is chosen. Various experiments were carried out and the results confirm that Logistic Regression and Naive Bayes are good at predicting CKD in an early stage.

Keywords: *CKD, Machine Learning, Algorithms, Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN)*

I. Introduction

There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, poor nutrition health and nerve damage, Decreased Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied. immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Glomerular Filtration Rate (GFR) is the best test to measure the level of kidney function and determine stage of chronic kidney disease. It can be calculated from the results of blood creatinine, age, race, gender, and other factors. The earlier disease is detected the better chance of showing or stopping its progression.

II. Literature Review

Early efforts in CKD prediction involved traditional statistical methods. Gunarathne et al. (2017) conducted a study that evaluated the performance of machine learning classification techniques for disease prediction, including CKD.[1] S. Ramya and Dr. N. Radha (2016) explored the use of machine learning algorithms for the diagnosis of CKD, emphasizing the potential of these techniques in enhancing diagnostic accuracy. [2] Dilli Arasu and Dr. R. Thirumalaiselvi (2017) reviewed the application of data mining techniques in understanding and

managing CKD. They highlighted the importance of advanced analytics for early detection and prognosis. The dataset used in this paper is sourced from the UCI machine learning repository, indicating a reliance on publicly available datasets for CKD research. [3] Sharma and Rizvi (2018) provided a broader perspective on the application of machine learning algorithms in predicting heart disease. Their survey encompassed various diseases and emphasized the importance of predictive models for healthcare management.[4] Shinde and Rajeswari (2018) conducted a comprehensive review of intelligent health risk prediction systems, emphasizing the potential of machine learning in predicting various health conditions.

III. Methodology

Data Preprocessing

Initially null values are checked in the dataset. The "?" symbols and null values are converted to Nan. As the first column age consists of Nan values, it is replaced with the mean of that particular column. Similarly it is done for the blood pressure, specific gravity, albumin, Blood glucose random, Blood urea, Serum creatinine, Sodium, Potassium, Haemoglobin, Packed cell volume, White blood cells count and Red blood cell count. The sugar column consists of values from 0 to 5, where there are missing values too. The mode method replaces the Nan values of the sugar column with the values that are repeated number of times. It is repeated for variables red blood corpuscles, pus cells, pus cell counts, bacteria, hypertension, diabetes, coronary artery disease, appetite, pedal edema and anaemia. The Red blood corpuscle is a nominal column which consists of normal and abnormal results which is converted into numerical variables (normal – 1, abnormal – 0). The pus cell and puss count attributes are also converted similarly. Likewise, the bacteria variable is replaced to 0s and 1s for present and non-present respectively. The hypertension variable consisting of yes and no is replaced with 0s and 1s. The same is done for anaemia too. The appetite variable consisting of no, poor and good is replaced with 0, 1 and 2. Now data pre-processing is over and now the cleaned data is ready for data analysis.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach for analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can provide before the modelling task. For better understanding of chronic dataset well, EDA is performed in such a way that the dependencies and weightage of values are clearly shown. After understanding the variables, it becomes easy to implement them in the model.

Fixing the Models

The main part of the model is model fixing. The models are trained in order to make the best prediction. The Chronic kidney disease dataset from the UCI repository consisting of 400 rows and 25 attributes. It comprises of 250 instances of CKD patients and the rest 150 for non-CKD instances.Three types of classifiers are experimented for CKD prediction and based on accuracy the best models for classification is chosen. The classifiers used here are,

- Logistic Regression
- Naive Bayes classification
- KNN Classifier

The classifiers are tested for their precision, recall, accuracy and F1 score. Optimization of model is necessary to increase its efficiency and for its true prediction and accuracy. The K-fold cross validation method is used to optimize the model . In K-fold CrossValidation (CV) a test set is separated from the remaining data in the data set to use for the final evaluation of models. The data that is remaining, i.e. everything apart from the test set, is split into K number of folds (subsets). The Cross Validation then iterates through the folds and at each iteration uses one of the K folds as the validation set while using all remaining folds as the training set. This process is repeated until every fold has been used as a validation set. In a Kfold CV after every iteration the model is scored and the average of all scores is computed to get a better representation of how the model performs compared to only using one training and validation set.

A. *Logistic Regression:*

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is commonly employed when the dependent variable is binary, meaning it has only two possible outcomes (e.g., 0 or 1, True or False, Yes or No).

```
Cross validation scores: [0.96551724 1.         1.         1.         1.
 1.         0.92592593 1.         0.96296296]
Mean of scores: 0.9854406130268201
Variance: 0.024347512633036986
Classification Report              precision    recall  f1-score   support

           0       1.00      1.00      1.00        45
           1       1.00      1.00      1.00        75

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120

Confusion Matrix [[45  0]
 [ 0 75]]
Training Accuracy 99.28571428571429
Testing Accuracy 100.0
The accuracy score achieved using Logistic Regression Model is: 100.0 %
```

Fig A.1 Logistic Regression

From the logistic regression model, The first obtained K fold Cross validation scores for about 10 folds. Then the mean values obtained that is about 0.98544, variance of about 0.02434. The variance is so low so there are not as much of errors occurred in the model. The classification report which displays the precision, recall, F1 and support scores for the model.

- Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 1.00, indicating the preciseness of the model which is so accurate.
- Recall value for class 0 is 1.00 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
- As the precision and recall values are similar, there are no dominance in classes.

Confusion matrix is nothing but an error matrix. From the matrix it shows that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn't even make a slight error and predicted the values perfectly. The training and test accuracy of the model are 99.25% and 100% respectively. To conclude the accuracy achieved using logistic regression model is 100%.

B. *Naïve Bayes Classification:*

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem which calculates the probability of a hypothesis of given evidence. In the context of classification, it helps determine the probability of a particular class given the observed features. and used for solving classification problems.

```
Cross validation scores: [1.         1.         1.         1.         1.
 1.         1.         1.         0.96296296]
Mean of scores: 0.9962962962962963
Variance: 0.011111111111111127
Classification Report              precision    recall  f1-score   support

           0       1.00      1.00      1.00        45
           1       1.00      1.00      1.00        75

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120

Confusion Matrix [[45  0]
 [ 0 75]]
Training Accuracy 100.0
Testing Accuracy 100.0
The accuracy score achieved using Naive Bayes Model is: 100.0 %
```

Fig A.2 Naïve Bayes

From the Naive Bayes model , The K fold Cross validation used has about 10 folds. Majority of the class predicted are 1.The mean of the scores obtained is about 0.99629, variance of about 0.01111. It has a low variance score so there is less errors occurred in the model. The classification report which displays the precision, recall, F1 and support scores for the model.

- Precision Score for both classes is 1.00, indicating the preciseness of the model i.e. so accurate.
- Recall value for is also 1.00, which describes the amount up – to which the model can predict the output.
- As the precision and recall values are similar, there are no imbalanced classes found.

From the confusion matrix it states that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn't even make a slight error and predicted the values perfectly. The training and test accuracy of the model is 100%. The accuracy score obtained using the Navie Byes model is 100%

C. ***KNN Classifier:***

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. In the same way the KNN classifies by considering the attributes involved in classification.

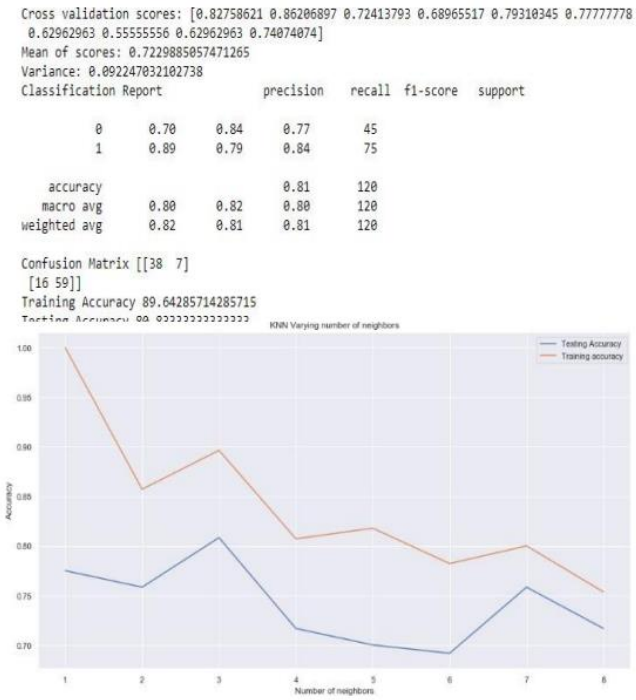

Fig A.3 K-Nearest Neighbors

The KNN Classifier fitted to predict whether the patients have CKD or not has come with fair results. From the KNN model, The first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are fall on range between 0.55-0.82, and there are chances that the model's accuracy will drop down to eighties. The mean of the scores obtained is about 0.72998, variance of about 0.0944. From the graph It can be inferred that in the testing accuracy there is a peak when the k value is 3. So, consider the k value is 3 for proceeding the model. The classification report which displays the precision, recall, F1 and support scores for the model.

- Precision Score for class 0 (negative) is 0.70 and for class 1 (positive) is 0.89, indicating the preciseness of the model which is fair.
- Recall value for class 0 is 0.84 and class 1 is 0.79, which describes the amount up – to which the model can predict the output. The positive class and negative class's recalls are quite fine.

- As the precision and recall values are quite different, the dominance of classes may occur at some places.

From the confusion matrix it is seen that there are 38 true positives, 7 false positives, 16 false negatives and finally 59 true negative values. The model made small errors on all the instances of the confusion matrix. And that is why the accuracy has reduced to 80s when compared to the other models. The training and test accuracy of the model is 89.64% and 80.83% respectively. The KNN model has the accuracy of 80.83%.

IV. Results And Discussions

The classifier used at first for CKD prediction is Logistic regression. The model has worked well and classified CKD with an accuracy of 100%. The model is also not over-fitted as model optimization techniques like K fold cross validation is done for 10 folds. So, the Logistic regression model can effectively classify the patients without even a slightest error. Naive Bayes has also classified perfectly and has the accuracy of 100%. The KNN Classifier used for predicting CKD has fair results. The model has classified the CKD with slight errors with an accuracy of 80.83%. Errors in the medical fields is not acceptable, even though the model tried is best after optimization too. The model is also not under - fitted as model optimization techniques like K fold cross validation is done for 10 folds.

Challenges and Future Directions

Despite the success of machine learning in CKD prediction, challenges persist, such as the need for more extensive and diverse datasets, ethical considerations, and interpretability of complex models. Future research directions should focus on implementing big data-oriented tools and techniques to enhance the speed and effectiveness of predictive models in healthcare.

V. Conclusion

In conclusion, the experiment on predicting chronic kidney disease using three classifiers Logistic Regression, Naive Bayes, and K-Nearest Neighbors (KNN). Both Logistic Regression and Naive Bayes classifiers demonstrate an impressive 100 percent accuracy in prediction task. The exceptional accuracy of Logistic Regression and Naive Bayes models suggests that these classifiers are well-suited for the given dataset and features. Further the robust cross-validation technique is used to confirm the generalization capabilities of these models. It's also essential to consider other performance metrics beyond accuracy, such as precision, recall, and F1 score, to gain a comprehensive understanding of classifier performance. Moreover, the interpretability and ease of implementation of each classifier should be taken into account when choosing the most suitable model for real-world applications. Hence we conclude that among the three classifiers experimented the Logistic Regression and Naive Bayes preforms Better with respective accuracy.

REFERENCES

[1] Gunarathne W.H.S.D, Perera K.D.M, Kahandawaarachchi K.A.D.C.P, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Brain Tumours”, 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[2] S.Ramya, Dr. N.Radha, "Diagnosis of Lung Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[3] S.Dilli Arasu and Dr. R.Thirumalaiselvi, “Review of Diabetes Mellitus based on Data Mining Techniques”,International Journal of Applied Engineering Research ISSN 0973- 4562 Volume 12, Number 23 (2017) pp.13498-13505

[4] L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,”2015.

[5] S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning : a review,” IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[6] Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms: A Survey",International Journal Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8

[7] Akhter T., Islam M.A., Islam S. Artificial neural network based covid-19 suspected area identification. *J Eng Adv.* 2020;1:188–194. [Google Scholar]

[8] Almasoud M., Ward T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *Int J Soft Comput Appl.* 2019;10 [Google Scholar]