

Analyzing Various Machine Learning Algorithms for Prediction of Chronic Kidney Disease

**Mr. K. Gopi, M. Bhargavi, A. Tejaswini, S. Ramya Sri,
S. Sai Sowmya, Sd. Taufeeq**

Department of Information Technology, Tirumala Engineering College

ABSTRACT

Chronic Kidney Disease is a serious lifelong condition that induced by either kidney pathology or reduced kidney functions. We examine the ability of several machine learning methods for early prediction of Chronic Kidney Disease. Predictive analytic is used to examine the relationship between data parameters as well as with the target class attribute. It enables us to introduce the optimal subset of parameters to feed machine learning to build a set of predictive models. Using confusion matrix we received the accuracy of all four methodologies such as K-Nearest Neighbour, Random Forest, Decision Tree and Light Gradient Boosted Machine. We receive the accuracy of KNN is 94 percentage and Random Forest, Decision Tree, Light Gradient Boosted Machine accuracy are 98 percentage. Compared to Light Gradient Boosted Machine, Decision Tree and Random Forest we receive less accuracy in K-Nearest Neighbour.

INTRODUCTION

Chronic kidney disease (CKD) is a significant public health problem for worldwide, especially for a low and medium-income countries. CKD means that the kidney does not work and cannot correctly filter the blood. About 10 percent of the population for a worldwide suffering from (CKD), and millions of die each year because of they couldn't get the affordable treatment, with the number increasing in the elderly. In 2010, a study was conducted by International Society of Numerology (ISN) on global burden disease, they reported that CKD has been raised an important cause of the mortality worldwide with the number of deaths increasing by 82.3 percent in the last two decades.

Keywords: Chronic Kidney Disease , Random Forest , Decision Tree, Light Gradient Boosted Machine, K- Nearest Neighbour, Machine Learning, Prediction

1. LITERATURE SURVEY

Jaymin Patel, et al. [4] is represented by the prediction of chronic kidney disease is very important and nowadays it is the leading cause of death. The performance of Decision tree method was found to be 99.25 percent accurate compared to naive Bayes method. Classification algorithm on chronic kidney disease data set the performance is obtained as 99.33 percent Specificity and 99.20 percent Sensitivity. They are also further working on enhancing the performance of prediction system accuracy in neural network and deep learning algorithm.

T Shaikhina, et al.[5] to observe classification algorithms to analyses and predict CKD. They have compared the performance of five classifiers in the prognosis of CKD. The experimental results of our proposed method have demonstrated that RF and XGB have produced superior prediction performance in terms of classification accuracy for our considered data set. For the future, we are going to working for enhancing the performance of prediction system accuracy by ensemble different classifier algorithms.

2. PROPOSED SYSTEM

It is better to search for an alternative approach with promising results and we are going to do that. We going to compare four very commonly known machine learning algorithms and are going to state the best among them using a case study on Chronic Kidney disease Prediction. As stated, the four machine learning algorithms are: Random Forest, Decision Tree, LGBM, and K- Near Neighbour (KNN).



3. TRAINING AND TESTING THE DATASET

The *training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model.* Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.

The training data varies depending on whether we are using Supervised Learning or Unsupervised Learning Algorithms.

For **Unsupervised learning**, the training data contains unlabeled data points, i.e., inputs are not tagged with the corresponding outputs. Models are required to find the patterns from the given training datasets in order to make predictions.

On the other hand, for supervised learning, the training data contains labels in order to train the model and make predictions.

The type of training data that we provide to the model is highly responsible for the model's accuracy and ability. It means that the better the quality of the training data, the better will be the performance of the model. Training data is approximately more than or equal to 60% of the total data for an ML project.

Testing:

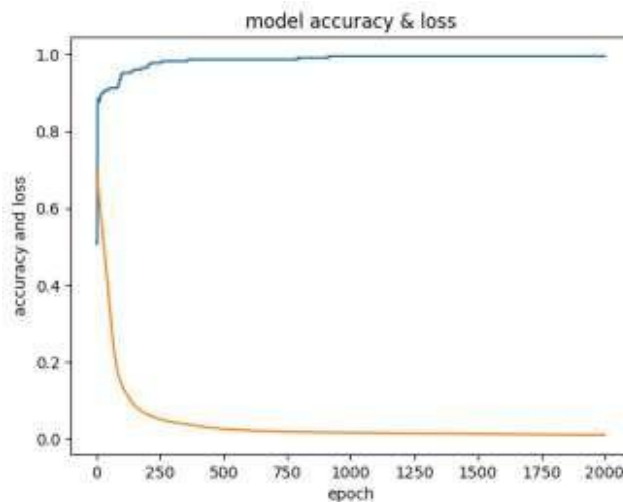
Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. *The test dataset is another subset of original data, which is independent of the training dataset.*

However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

At this stage, we can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset. If the accuracy of the model on training data is greater than that on testing data, then the model is said to have overfitting.

The testing data should:

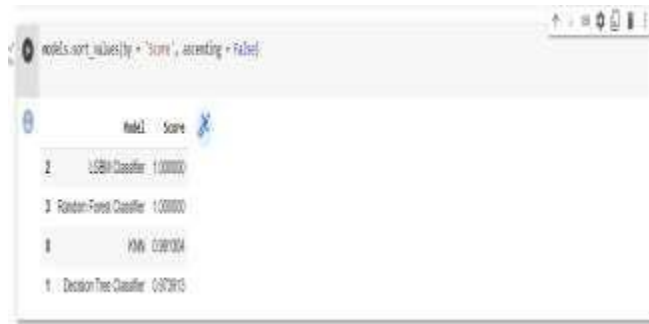
- Represent or part of the original dataset.
- It should be large enough to give meaningful predictions.



RESULTS

As we have mentioned the earlier that it is better to search for an alternative approach with promising results and we are going to do that. So, in this study, we are going to compare four very commonly known machine learning algorithms and we are going to state the best among them using a case study on Chronic Kidney disease Prediction. As stated, the four machine learning algorithms are: Random Forest Decision Tree LGBM K-Nearest Neighbour (KNN) Why we are going with only four algorithms is that we don't want to overload and confuse the learning procedure by including more machine learning algorithms. So, we thought that four would be an optimal choice and the machine would work just fine.

Using confusion matrix we received the accuracy of all four methodologies such as K-Nearest Neighbour, Random Forest, Decision Tree and Light Gradient Boosted Machine. We receive the accuracy of KNN is 94 percentage and Random Forest, Decision Tree, Light Gradient Boosted Machine accuracy are 98 percentage. Compared to Light Gradient Boosted Machine, Decision Tree and Random Forest we receive less accuracy in K-Nearest Neighbour.



Model	Score
2 - USM Classifier	1.00000
3 - Random Forest Classifier	1.00000
4 - SVM	0.99104
1 - Decision Tree Classifier	0.97815

4. CONCLUSION

CKD means that the kidney does not work and cannot correctly filter the blood. About 10 percent of the population for a worldwide suffering from (CKD), and millions of die each year because of they couldn't get the affordable treatment, with the number in- creasing in the elderly. In 2010, a study was conducted by International Society of Numerology (ISN) on global burden disease, they reported that CKD has been raised an important cause of the mortality worldwide with the number of deaths increasing by 82.3percent in the last two decades. Also, the number of patients reaching End-Stage Renal Disease (ESRD) is increasing, which requires kidney transplantation or dialysis to save the patient 's lives. Machine learning models can effectively aid clinicians to achieve this goal due to their fast and accurate recognition performance. By using these ML algorithms, it is easy to identify the kidney disease in early stage. Random Forest, Light Gradient Boosted Machine, these two are best algorithms.

REFERENCES

- [1] J. Aljaaf et al, "Early prediction of chronic kidney disease using machine learning supported by predictive analytic," in 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, .
- [2] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel,et al., "Heart Disease Prediction Using Ma- chine learning and Data Mining Technique", International Journal Of Computer Science Com- munication, Vol. 7, No. 1, pp.129 – 137. (2019)
- [3] T Shaikhina, Torgyn, et al. "Decision tree and random forest models for outcome prediction in anti body incompatible kidney transplantation." Biomedical SignalProcessing and Control (2019).
- [4] C.T. Tran, et al., Multiple Imputation and Ensemble Learning for Classification with Incomplete Data, Springer International Publishing, pp. 401-415. (2020)
- [5] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Transnational Medicine, vol. 17,(1), pp. 119, 2019.