CHRONIC KIDNEY DISEASE ANALYSIS

Department Of Computer Engineering
University Of Peradeniya

Group 02
Shamra Marzook (E/16/232)
Subhash Rathnayake (E/16/320)
Vindula Rathnayake (E/16/319)

ABSTRACT

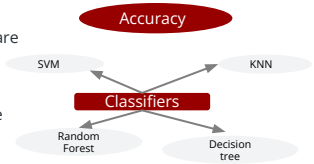
Chronic kidney disease (CKD) is a condition where the kidneys get damaged,making one unhealthy. Early detection can often prevent the disease from getting worse and being fatal.

In this project machine learning techniques like KNN, SVM, random forest and decision tree classifiers have been explored. The objective is to compare the performance of these classifiers on the basis of accuracy for the prediction of CKD in patients and from the experimental results it is observed that the accuracy and performance of random forest classifier is the best.

INTRODUCTION

Machine Learning is a rising field concerned with the study of huge and multiple variable data. In Medical Science perspective, machine learning and data mining techniques together have proved success in prediction and diagnosis of various critical diseases.

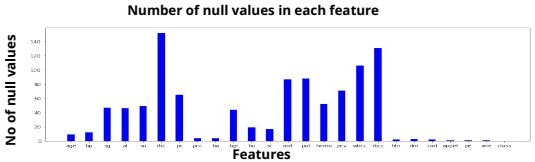
Various classification approaches and machine learning algorithms are applied for prediction of chronic diseases. Here we are concerned about CKD, which is an abnormal function of kidney or a progressive failure of renal function over a period of months or years.



Often, chronic kidney disease is diagnosed as a result of screening of people known to be at risk of kidney problems, such as those with high blood pressure or diabetes and those with a blood relative with CKD.

DATA SET

Source : UCI - Machine Learning Repository
No of Instances : 400
No of Attributes : 25 (11 numeric and 14 nominal)



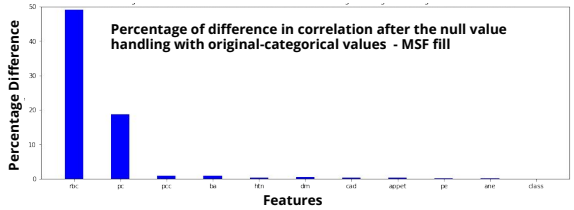
METHODOLOGY

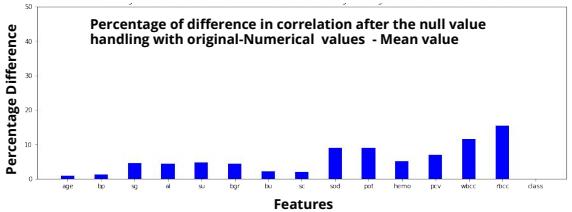
Data Preprocessing
Categorical data in the form of "Yes", "No" , "Normal" , ect . Therefore replace those values with 1 & 0 .

Approaches for null value handling

- Numerical type Data - Mean Value
- Categorical type Data - Most frequent value

Then analyze the change of correlation of each feature with class label. And decided to drop the feature call rbc which had 48% change in correlation and 0.28 amount of correlation change.



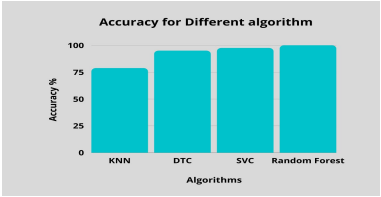


Divide 20% of the data set as Test set & remainder as train set

Model Evaluation
Use 4 classifier algorithms for training model and evaluate them using accuracy & the confusion matrix. Algorithms use to train models,

1. K-Nearest Neighbors Classifier
2. Support Vector Classifier
3. Random Forest Classifier
4. Decision Tree Classifier

RESULT



Random forest classifier algorithm gives the best accuracy which is 100 %

CONCLUSION

Conclusion

- Best alternative to state whether the patient has CKD positive or negative.
- Further improvement is required.

Application

- Fast processing and immediate results with high accuracy.
- Minimizing human effort and cost efficient databases

The benefits of this model are

- Straightforward results
- Accurate performance Calculations.

Future scope

- Database should be expanded on which the system will be tested much better.With new dataset we can train this model to predict the chronic kidney disease stage as well.

.

REFERENCES

https://www.datarobot.com/wiki/fitting/
https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
https://www.datacamp.com/community/tutorials/random-forests-classifier-python
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://machinelearningmastery.com/machine-learning-in-python-step-by-step/