Sub-Sahara African Academic Research Publications
September, 2021, Editions

JPAS
Journal of Pure and Applied Science
Vol..21, No.9, ISSN 2318-9315

PREDICTION OF CHRONIC KIDNEY DISEASE WITH EXTRA TREE CLASSIFIER FEATURE SELECTION TECHNIQUE

ILIYAS IBRAHIM ILIYAS[1]; ADATI ELKANAH CHARI[2]; ANDRA UMORU[3]; & ADAMU ABDULLAHI GABRA[4]
[1,3]Department of Mathematical Sciences, University of Maiduguri, Borno State. [2]Department of General Studies, federal College of Education, Yola. [4]Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, Skudia,

ABSTRACT
The Kidney is an organ in the body that removes unwanted Substances from the blood circulating in our body. When it stops functioning, the blood circulates with its waste, causing the disease to the body. Data mining is gradually becoming more prevalent nowadays in healthcare and fraud, Abuse detection, etc. The Kidney is essential for the filtering and purification process of human blood. Death is imminent and inevitable within few days without at least one functioning Kidney. Ignoring kidney malfunction can cause kidney disease leading to death. Prediction of kidney disease can be made efficiently with the feature selection technique. It can be achieved with the assistance of the Attribute selection method. This is achieved by removing the features/attributes of less importance and selecting the critical attribute present in the dataset. Ignoring to do so can affect the model's performance, consumed time, and be expensive. This study aims to predict Chronic Kidney Disease (CKD) with Deep Neural Network (DNN), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM). We apply the use of Extra Tree-Based Classifier to identify rankings of the attributes in our dataset and only choose 12 attributes and apply the machine learning techniques to predict CKD and the accuracy of the selected methods. Our study DNN has surpassed RF, NB, and SVM in terms of accuracy with 99%.

Keywords: *Deep Neural Network,* ***Extra Trees Classifier,*** *Chronic Kidney Disease, Machine learning*

Introduction

A kidney is an organ in the body that removes waste from the body that circulates in our body. When it stops functioning, the blood circulates with its waste, causing the disease to the body. Chronic Kidney Disease (CKD) is a disorder against proper function regarding kidneys. Our kidneys balance the salt and minerals such as calcium, phosphorus, sodium, and potassium that circulate in our blood, filter wastes from the blood, and remove them through urination. This filtering process includes excess fluids from our body Chukwuonye et al., 2018. CKD was ranked in the list of diseases that cause global deaths in the 1990s; by 2010, it had fallen to 28th in the list of global death (Chukwuonye et al., 2018). Levey et al. (2007) made it clear that the level by which it rises was noted to be second only to HIV & AIDs. According to Luyckx & Stanifer (2018), CKD increased globally from 19 million in 1990 to 33 million in 2013, and in 2010, 2.62 million individuals got dialysis around the world.

Data mining is gradually becoming more prevalent nowadays in healthcare and fraud, abuse detection, etc. Koh and Tan (2011) stated that classification is a more useful data mining function to handle items in a collection to class or target categories. For obtaining essential information from medical databases. With the use of machine learning and statistical analysis cleverly. Machine learning techniques comprising of various statistical analyses and datasets assist in extracting hidden patterns and identify relationships from vast among of data (Padmanaban & Parthiban 2016).

A Deep Neural Network is a set of machine learning algorithms with multiple layers that correspond to various concept levels each level. It comprises an input, output, and several hidden layers. It can be used for image processing, object detection, Natural Language Processing, etc. (Chahal & Gulia, 2019).

The kidney can be predicted from the record of clinical information. The clinical information can be maintained either manually or systematically. The systematic maintained data can be simpler and powerful than manually maintained data. The clinical data are stored in the database which contains patient information such as age, blood pressure etc., which are used to predict the disease (Arasu & Thirumalaiselvi, 2017). The Kidney is essential for the filtering and purification process of our blood. Death is imminent and inevitable within few days without at least one functioning Kidney. Ignoring kidney malfunction can cause kidney disease leading to death. CKD and its symptoms are frequently mild and gradual and often go unnoticed for years only to be recognized lately. Purusothaman & Krishnakumari (2015) indicated that kidney failure falls among several illnesses such as heart disease, blindness, etc., resulting in chronic Diabetes. Dialysis and transplant are the only methods to keep the kidneys function artificially. It is a painful and expensive process; early

discovery and characterization of CKD are considered to be an essential component within the administration and control of CKD; use of efficient data mining techniques to reveal and extract hidden information from clinical and laboratory patient data is imminent, which can be supportive to help doctors in maximizing precision for recognizable proof of infection seriousness arrange (Rady & Anwar, 2019). The prediction of kidney disease can be made efficiently with feature selection. It can be achieved with the assistance of the Attribute selection method. It is done by removing the features/attributes of less importance and selecting the critical attribute present in the dataset. Ignoring to do so can affect the model's performance, consumed time, and be expensive.

Machine Learning is a growing field concerned with studying enormous and several variable data and grown from pattern recognition and computational learning theory in Artificial Intelligence (AI), having computational methods, algorithms, and techniques for analysis and prediction. From Medical Science's viewpoint, Machine Learning techniques have shown success in predicting and diagnosing numerous critical diseases. Furthermore, human professionals and experts are limited to finding hidden patterns from data. Hence, the alternative is to use computational methods to investigate the raw data and exciting information for the decision-maker (Kumar, 2016). Feature selection has been a functioning exploration territory in design acknowledgment, insights, and data mining networks. The fundamental thought of feature selection is to pick a subset of info factors by wiping out features with almost no prescient data. Feature selection can fundamentally enhance the conceivability of the subsequent classifier models and frequently manufacture a model that sums up better-concealed focuses (Alind, 2020).

You can get the feature rankings of each feature/attribute of your dataset by using the model's feature importance property. Feature selection gives you a score or rankings for each feature/attribute of your data. The higher the rankings, the more relevant or important is the feature/attribute towards your output variable. The feature importance used in this work is an inbuilt class that is originated from Tree-Based Classifiers. We used Feature importance, an inbuilt class with a Tree-Based Classifier, to extract the top features from our dataset and determine the attributes with the highest influence of CKD prediction using the dataset we used.

In this study, Deep Neural Network (DNN), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) model is used to predict CKD at the University of California Irvine (UCI) dataset with Tree-Based Classifier feature selection technique. Performance evaluation of the model was computed by computing the Sensitivity, Specificity, Recall, Precision, F1 Score, Receiver Operating Characteristic (ROC) Score, and Cohens Kappa.

Literature Review

Arasu & Thirumalaiselvi (2017) used Weighted Average Ensemble Learning Imputation (WAELI) to perform feature selection and predicted CKD with the selected features on the UCI dataset, the algorithm used by the authors in the prediction is: SVM and ANN with an accuracy of 73% for both algorithms while after feature selection, an accuracy of 78% for both ANN and SVM was achieved. The dataset used by the authors is an imbalanced dataset which occurs when there are too little data of a particular class, these can affect the accuracy of the models.

Arafat, Fatema & Islam (2018) Studied an automated detection of CKD with clinical data using RF and NB based on a comparative study on the UCI dataset. They computed the weight of each attribute used in the dataset. Their result shows that RF has higher accuracy of 98%, followed by LR and NB with 96% for each. They were able to compute the weight of each attribute but did not validate the use of RF and NB on a reduced feature to verify if they can achieve better accuracy.

Misir, Mitra & Samanta (2017) Predicted CKD and NCKD with reasonable accuracy using a lesser number of features on a balanced dataset gotten from the UCI repository dataset, and they performed feature extraction and reduction using CFS, with WEKA as a tool. Their work was able to achieve good accuracy with the use of two classifiers, namely: Correlation-based feature subset selection and Levenberg–Marquardt on 8 attributes. They ignored the issue of missing values; instead, they deleted the instances with missing values and only used 158 cases.

Alshebly & Ahmed (2019) applied different machine learning algorithm, which is ANN and LR, to a problem in the domain of medical diagnosis and analyzed their efficiency of the prediction on the University of California Irvine (UCI) dataset with 153 cases and 11 attributes of CKD patients, the observed performance of the ANNs classifier is better than LR mode with the accuracy of 84.44%, the sensitivity of 84.21, specificity of 84.61% and Area Under the Curve (AUC) of 84.41% and found that the most critical factors that have a clear impact on CKD patients are creatinine and urea, they ignored cases with missing values and only used 153 points instead of addressing it and also the dataset is an imbalanced data.

Zeynu & Patil (2018) used KNN, DT, ANN, NB, and SVM to diagnose CKD. They build two important models, feature selection method and ensemble subset evaluator with a best-first search engine. They used info gain attributes evaluator with ranker search engine and wrapper subset evaluator with the best first search engine. Their result shows that KNN using wrapper subset evaluator with best first search engine feature selection method has 99%, J48 with info

gain attribute evaluator with a ranker search engine has 98.75%, ANN with wrapper subset evaluator with a best-first search engine has 99.5% accuracy, NB with wrapper subset search engine has 99%, SVM with info gain attribute evaluator with ranker has 98.25% accuracy, the second model-building method ensemble model was examined by five heterogeneous classifiers based on a voting algorithm. The effectiveness of their proposed ensemble model was examined by comparison of the base classifier. The experiment result showed that the proposed ensemble achieved 99% accuracy. Their study was executed on WEKA and on the UCI dataset.

Polat et al., (2017) used the Support Vector Machine classification algorithm to diagnose Chronic Kidney Disease. During the diagnosis, two essential types of feature selection methods, namely, wrapper and filter approaches, were chosen to reduce the Chronic Kidney Disease dataset dimension. The classifier subset evaluator with greedy stepwise search engine and wrapper subset evaluator with the Best First search engine was used in the wrapper approach. In the filter approach, correlation feature selection subset evaluator with greedy stepwise search engine and filtered subset evaluator with the Best First search engine was used. The dataset used by the authors was from the UCI repository. Their results showed that, SVM without feature selection: 97.75%, SVM with ClassifierSubsetEval with Greedy stepwise: 98%, SVM with WrapperSubsetEval with Best First: 98.25%, SVM with CfsSubsetEval with Greedy stepwise: 98.25%, SVM with FilterSubsetEval with Best First: 98.5%.

Salekin & Stankovic (2016) considered 24 predictive parameters and created a machine classifier to detect CKD, they were able to achieve an accuracy of 0.993 according to an f1 score with 0.1084 roots mean square error. They performed feature selection to determine the most relevant attributes for detecting CKD and rank them according to their predictability. The dataset used by the authors was the UCI dataset, KNN, Random forest, and ANN was the classifier used. They performed feature selection using two methods: wrapper method and LASSO regularization, they found out that RF with reduced attributes set of 12 attributes can detect CKD with the highest accuracy of 0.998 using the f1 measure and with 0.107 roots mean square, error. Which is a 57% RMSE reduction compared to the state-of-the-art solutions. Through their evaluation, they find hemoglobin, which is an indicator of anemia, diabetes mellitus, specific gravity, hypertension, and previously explored serum creatinine, and albumin which are highly predictive attributes for CKD. Also, through cost analysis considering all 24 attributes, they identified a cost-effective, highly accurate detection classifier using only 5 attributes: specific gravity, albumin, diabetes mellitus, hypertension, and hemoglobin.

This study will compare DNN with other techniques, perform missing value handing with multiple imputations, and apply feature selection to determine features with the highest prediction contribution.

Methodology

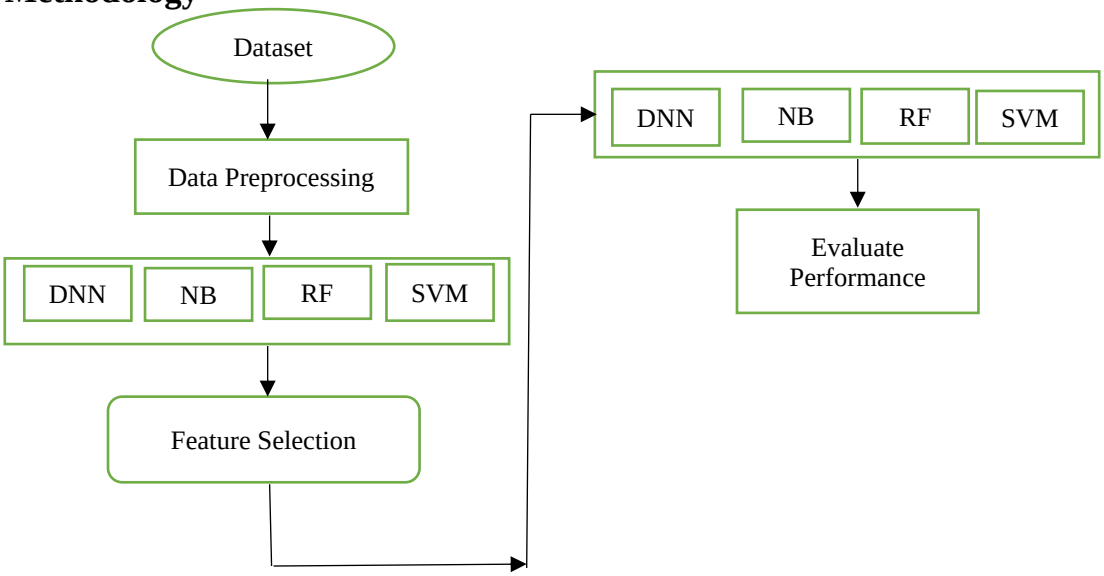


Figure 3.1: Flow of Research Methodology

Dataset

In this study, we used a dataset from UCI machine learning repository named: University of California Irvine (UCI) dataset, which contained 400 instances consisting of an input and target variables, 25 variables and 1 variable for the input and target variables, respectively (14 nominal and 11 numeric), the target variable consist of a binary classification called Chronic Kidney Disease (CKD) and Not Chronic Kidney Disease (NCKD) 150 are not having CKD and 250 are having CKD. We split our dataset into 10 parts, and 10-fold cross-validation was performed during each experiment to evaluate the algorithms. We implemented classification algorithms in the open-source anaconda tool using Python Programming Language.

Table 1. Dataset Attributes

Attribute	Description
Age in years	Number of years
Blood pressure of a patient	Given in mm/Hg
Specificity Gravity	Ranges from 1005 to 10025 (the higher the risk)
Albumin	The range is 0 to 5 (the higher, the better)

Sugar level	5 levels indicating the severity
Red Blood Cells	Is abnormal or normal
Pus Cell	Is normal or not normal (high number lead to the urinary tract)
Pus Cell clumps	Can be present or not present
Bacteria	Can be present or not present
Blood Glucose	It is in mgs/dl
Blood urea	It is in mgs/dl
Serum creatinine	High level is not good
Sodium	It is measured in mEq/L
Potassium	It is measured in mEq/L
Hemoglobin	Less than 15 is kidney failure
Packed Cell Volume	This is numerical
White Blood Cell Count	This is the numerical cell count
Red Blood Cell Count	Should not be higher or less than normal
Hypertension	It is categorical (yes or no)
Diabetes (Mellitus)	It is categorical (yes or no)
Artery disease (Coronary)	It is categorical (yes or no)
Appetite	Is it poor or good (yes or no)
Pedal Edema	It is categorical (yes or no)
Anemia	It is categorical (yes or no)
Class	Given as CKD or NOTCKD

Data Pre-Processing

The UCI Dataset contains 400 instances of CKD patients. To obtain accurate results which are not misleading, sufficient data pre-processing should be done. Many healthcare datasets usually have missing values, noisy and inconsistent data. All these contribute to the quality of data. Low-quality data results in extremely poor machine learning results. The data was cleaned, and we used simple imputations to handle the issue for easy execution by applying MissForest, a machine learning-based imputation technique. It uses a Random Forest algorithm to do the task. It is based on an iterative approach, and at each iteration, the generated predictions are better.

Performance Evaluation

In this work, we used eight (8) performance measures: accuracy, specificity, sensitivity, kappa statistic, precision, recall, ROC Score, and F1 Score. We also performed feature selection to find the importance of our attributes.

Confusion Matrix: confusion matrix indicates the model's statistical suitability and its compatibility with the dataset. It can be defined as a table layout used

explicitly to visualize algorithm performance (Alshebly & Ahmed, 2019). Table 3.3 shows the summary table.

Table 3.3: Confusion Matrix

Classification		**Observation**	
		Negative	**Positive**
Positive	**Negative**	**True Negative (TN)**	**False Positive (FP)**
	Positive	**False Negative (FN)**	**True Positive (TP)**

i. **Accuracy**- It is used to classify the number of correctly predicted data points out of all data points. The number of expected data points was correctly divided by the total number of data points prediction made (Alshebly & Ahmed, 2019). It is expressed in equation 3.1: $$\boldsymbol{Accuracy} = \frac{\boldsymbol{TN+TP}}{\boldsymbol{TP+FP+FN+TN}} \ldots\ldots (\mathbf{3.1})$$

ii. **Precision:** It is defined as the portion of relevant instances among the retrieved instances. It is given as the correlation number between the correctly classified modules to entire classified fault-prone modules (Alshebly & Ahmed, 2019). It is expressed in equation 3.2: $$\boldsymbol{Precsion} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{3.2})$$

iii. **Recall/ Sensitivity:** Recall is a metric that measures the number of correct positive classified data points made out of all the positive data points that are supposed to be made (Alshebly & Ahmed, 2019). It is expressed in equation 3.3: $$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{3.3})$$

iv. **F1 Score**: This used to determine the mean between precision and recall.It is used to describe the preciseness (number of instances can be correctly predicted by the algorithm) and robustness (whether it avoids missing any important number of instances) of a technique (iliyas, 2020. The equation of f1-score is in equation 3.7: $$\boldsymbol{F1\ Score} = \mathbf{2} \times \frac{\mathbf{1}}{\frac{\mathbf{1}}{\boldsymbol{precision}} + \frac{\mathbf{1}}{\boldsymbol{recall}}} \ldots\ldots\ldots\ldots\ldots (\mathbf{3.4})$$

Naive Bayes (NB)

Naive Bayes is a machine learning algorithm or classifier which uses the **Bayes theorem** with independent assumptions between features. The one-dimensional Naive Bayes classifier computes the ratio of the log probabilities of the features belonging to all the classes. The naive Bayes classifier computes the class value probability assuming each of the attributes independently. This means Naive Bayes does not consider the correlation between attributes. Naive Bayes is a very scalable classifier, but it can create a bias towards one or more attributes,

resulting in accuracy (Arafat et al., 2018). Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x), and P(x|c), The expression of Naïve Bayes is in equation 3.5:

$$\boldsymbol{P(c|x) = \frac{P(c|x)P(c)}{P(x)}} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (\mathbf{3.5})$$

Support Vector Machine (SVM)

The main advantage of the SVM classifier is to discover the improved decision border, which exemplifies the most extraordinary decisiveness (maximum margin) amidst the classes. The standard of SVM begins from resolving the problems of linear separable then expands to treat the non-linear cases. (Obaid et al., (2018). An SVM model is a representation of different classes in a hyperplane in a multidimensional space. The hyperplane will be generated iteratively by SVM so that the error can be minimized. A paradigm of the SVM framework is shown in Figure 3.1. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

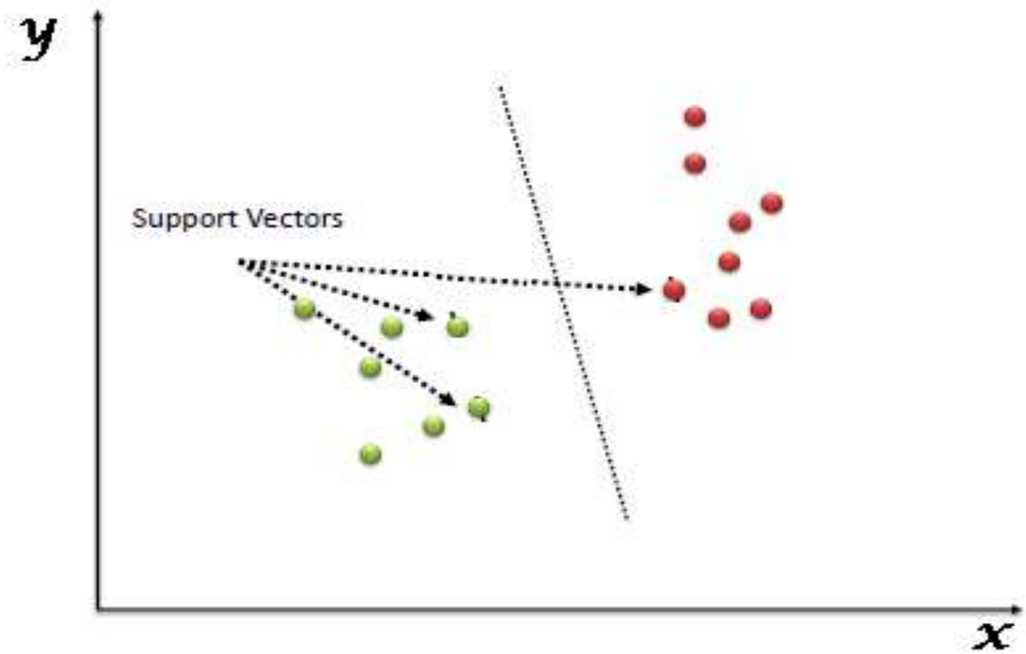


Figure 3.2 Support Vector Machine

Deep Neural Network (DNN)

A Deep Neural Network (DNN) is a deep learning technique that comprises an input layer, several hidden layers, and an output layer. Each layer consists of several units called neurons. These neurons are also referred to as artificial

neurons. A neuron obtains several inputs, performs a weighted summation over its inputs with a bias, then the resulting sum goes activation process with an activation function to yield output. Each neuron contains a vector of weights associated with its input size and a bias is optimized during the training process (Chahal & Gulia, 2019). Figure 3.2 shows the structure of DNN.

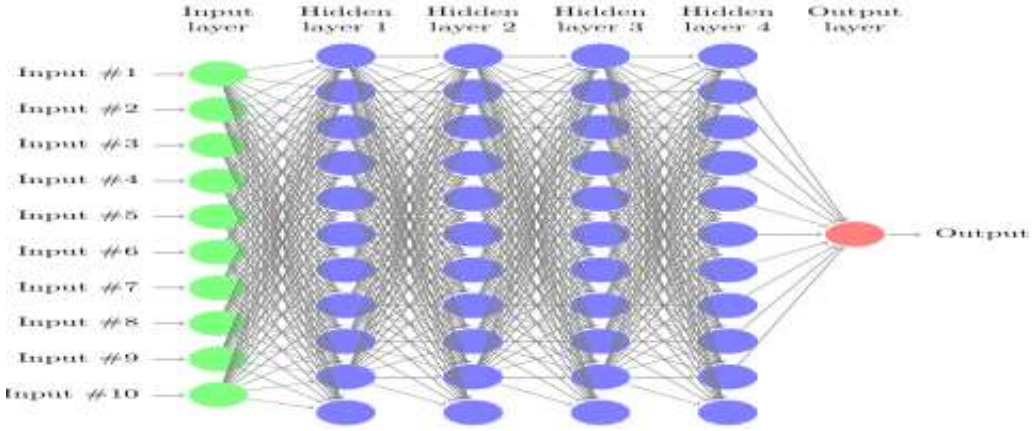


Figure 3.3: DNN Structure

Random Forest

Random forest (RF) is a well-known ensemble learning method applied in different fields, including high-dimensional classification and pattern recognition. When using RF, many single individual decision trees will be created utilizing classification and regression algorithms (Chimwayi et al., 2017). The Random Forest Algorithm comprises different decision trees, each with the same nodes, but using other data leads to different leaves. It merges the decisions of multiple decision trees, which represents the average of all these decision trees, as shown in Figure 3.3.

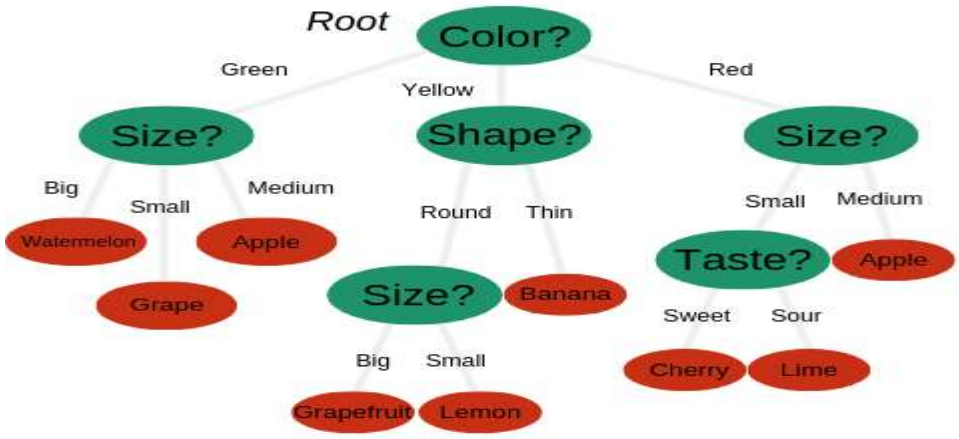


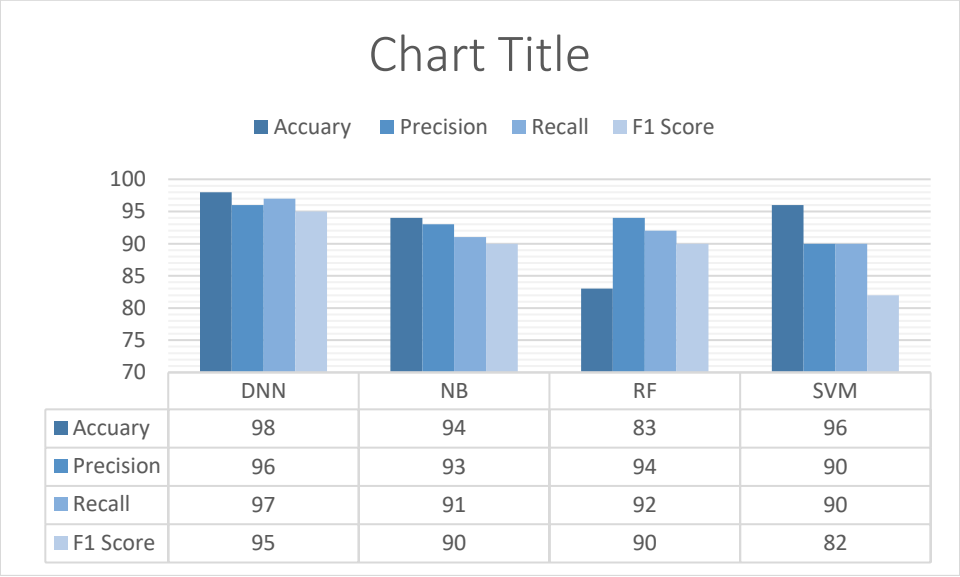
Figure 3.4: Random Forest

Highly Randomized Trees Classifier (Extra Trees Classifier)

Highly Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in constructing the decision trees in the forest. Each Decision Tree in the Extra Trees Forest is built from the original training sample. Then, at each test node, each tree is provided with a random selection of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees (Alind, 2020).

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature, and the user selects the top k features according to his/her choice (Alind, 2020).

Results and Discussion



	DNN	NB	RF	SVM
Accuary	98	94	83	96
Precision	96	93	94	90
Recall	97	91	92	90
F1 Score	95	90	90	82

Figure 4.1: Performance Evaluation

Figure 4.1 indicated the performance metrics of the techniques based on Accuracy F1 Score, Recall and Precision, in terms of accuracy: DDN: 98%, NB:

94%, RF: 83%, and SVM: 96% while in terms of Precision, DDN: 96%, NB: 94%, RF: 94% and SVM: 90%, in terms of Recall, DDN: 97%, NB: 91%, RF: 92% and SVM: 90% and in terms of F1 Score: DDN: 95%, NB: 90%, RF: 90% and SVM: 82%, the comparison has shown that DNN performed better in terms of accuracy than other techniques.

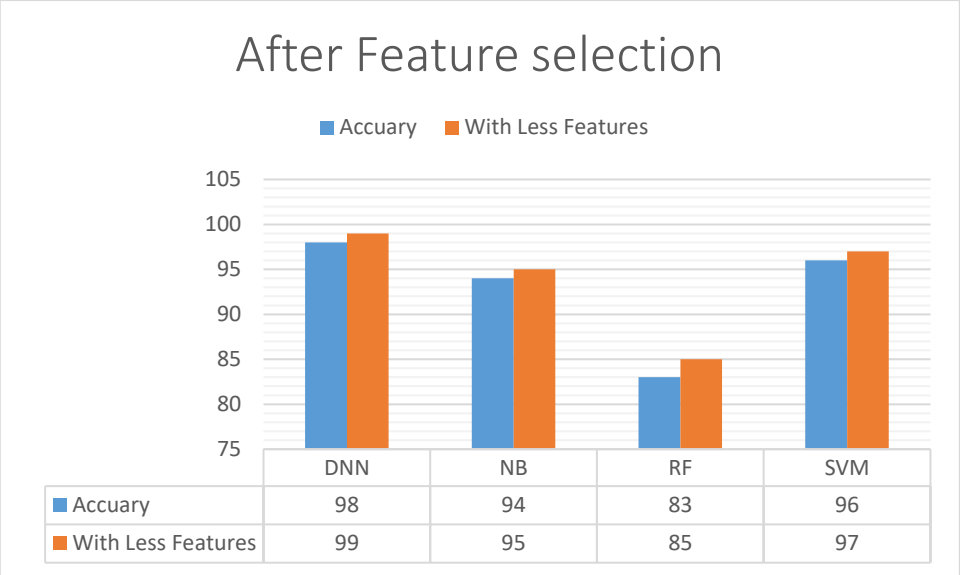


Figure 4.2: Results comparison with feature selection

Figure 4.2 shows the result of machine learning techniques used to predict CKD before and after applying feature selection. After the feature selection, we were able to use only 12 attributes, the accuracy of the methods tends to increase after using lesser attributes, we achieved an accuracy of 99% for DNN, ND: 95%, RF: 85%, SVM: 97%, indicating an improvement in the accuracy.

Conclusion

In this paper, we were able to apply multiple imputations of our dataset. We then predicted CKD using NB, SVM, RF, and DNN, before and after identifying feature importance of our dataset in the prediction of CKD, to see if we can achieve better accuracy with lesser attributes which can cost less and consumed less time, DNN has shown to have a better accuracy compared to other techniques with 98%. Extra-based Tree Classifier is the technique used in performing feature importance selection, and in the end, we achieved a greater accuracy with just 12 attributes. This indicates that an extra tree classifier can eliminate dominant attributes in prediction, saving a lot of money and saving time since more excellent attributes come with more cost and time.

Reference

[1]. Alind, G. (2020). ML Extra Trees Classifiers for Feature Selection. https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/amp/
[2]. Alshebly, O. Q., & Ahmed, R. M. (2019). Prediction and Factors Affecting of Chronic Kidney Disease Diagnosis using Artificial Neural Networks Model and Logistic Regression Model. *Iraqi Journal of Statistical Sciences*, 16(28), 140-159.
[3]. Arafat, F., Fatema, K., & Islam, S. (2018). Classification of chronic kidney disease (CKD) using data mining techniques (Doctoral dissertation, Daffodil International University), 30(2), 3321-3443
[4]. Arasu, S. D., & Thirumalaiselvi, R. (2017). A Prediction of Chronic Kidney Disease Using Feature based Priority Assigning Algorithm. *International Journal of Applied Engineering Research*, 12 (20), 9500-9505.
[5]. Chahal, A., & Gulia, P. (2019). Machine Learning and Deep Learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(2), 2278-3075.
[6]. Chimwayi, K. B., Haris, N., Caytiles, R. D., & Iyengar, N. C. S. (2017). Risk Level Prediction of Chronic Kidney Disease Using Neuro-Fuzzy and Hierarchical Clustering Algorithm (s), 8(12), 112-119.
[7]. Chukwuonye, I. I., Ogah, O. S., Anyabolu, E. N., Ohagwu, K. A., Nwabuko, O. C., Onwuchekwa, U., & Oviasu, E. (2018). Prevalence of chronic kidney disease in Nigeria: systematic review of population-based studies. *International journal of nephrology and renovascular disease*, 1(3)11, 165.
[8] Iliyas, I, I., Rambo, S. I., Dauda, A. B., & Tasiu, S. (2020). Prediction of Chronic Kidney Disease Using Deep Neural Network. *FUDMA Journal of Sciences*, 4(4), 34-41
[9]. Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 65. https://doi.org/10.4314/ijonas.v5i1.49926
[10]. Kumar, M. (2016). Prediction of chronic kidney disease using random forest machine learning algorithm. *International Journal of Computer Science and Mobile Computing*, 5(2), 24-33.
[11]. Levey, A. S., Atkins, R., Coresh, J., Cohen, E. P., Collins, A. J., Eckardt, K. U., ... & Powe, N. R. (2007). Chronic kidney disease as a global public health problem: approaches and initiatives–a position statement from Kidney Disease Improving Global Outcomes. *Kidney international*, *72*(3), 247-259.
[12]. Luyckx, V. A., Tonelli, M., & Stanifer, J. W. (2018). The global burden of kidney disease and the sustainable development goals. *Bulletin of the World Health Organization*, *96*(6), 410-414.
[13]. Misir, R., Mitra, M., & Samanta, R. K. (2017). A reduced set of features for chronic kidney disease prediction. *Journal of pathology informatics*. 1(6), 333-340.
[14]. Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., & Taha, F. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. *International Journal of Engineering & Technology*, 7(4.36), 160-166.
[15]. Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. *Indian Journal of Science and Technology*, *9*(29), 1-6.
[16]. Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of medical systems*, 41(4), 55.

[17]. Purusothaman, G., & Krishnakumari, P. (2015). A Survey of Data Mining Techniques on Risk Prediction: Heart Disease. *Indian Journal of Science and Technology*, *8*(12), 1-8.
[18]. Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, 10(2), 178-183.
[19]. Salekin, A., & Stankovic, J. (2016, October). Detection of chronic kidney disease and selecting important predictive attributes. In 2016 *IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 262-270). IEEE.
[20]. Zeynu, S., & Patil, S. (2018). Prediction of chronic kidney disease using data mining feature selection and ensemble method. *International Journal of Data Mining in Genomics & Proteomics*, 9(1), 1-9.