Department of Data Science

DTSC 3010: Project Report on Machine Learning

Chronic Kidney Disease Prediction

Author:
Chioma Ifezie
Student ID: 900333450

Supervisor:
Dr. Nawa Raj Pokhrel

April 2024

Contents

1 Background

The aim of this project is to diagnostically predict the presence of Chronic Kidney Disease (CKD) in patients. The goal is to identify individuals with a risk of having CKD by analyzing their medical information using machine learning techniques. I chose this topic because about 35 million people in the US are estimated to have CKD and about 90% of these people do not know that they have it. The lack of early detection and affordable treatment options worsens the situation, leading to severe cases. By employing machine learning techniques to predict CKD from medical data, this project aims to contribute to early diagnosis and timely intervention, potentially saving lives and reducing the healthcare costs associated with the late treatment of kidney disease.

The input to this algorithm consist of a series of medical parameters such as Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count, which are numerical in nature and Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, which are categorical in nature (meaning they have binary values (present or not present)). We then used Logistic Regression to make the prediction. Logistic Regression is particularly suited for binary classification tasks like this, where the outcome is to predict whether or not a patient has CKD.

2 Related Work

The task of diagnosing Chronic Kidney Disease (CKD) using machine learning has been explored extensively in recent years, reflecting various approaches to handling medical data and improving diagnostic accuracy.
Gunarathne W.H.S.D et al. [1] conducted comparative analyses across multiple models and determined that the Multiclass Decision Forest algorithm outperforms others, achieving approximately 99% accuracy on a streamlined dataset with 14 attributes. S.Dilli Arasu and Dr. R. Thirumalaiselvi [2] addressed missing values in a Chronic Kidney Disease dataset. They noted that missing values can diminish both the accuracy and the predictive performance of models. To tackle this issue, they conducted a recalculation of CKD stages, which allowed them to identify and substitute unknown values with recalculated ones. Sahil Sharma, Vinod Sharma, and Atul Sharma [3] evaluated 12 different classification algorithms on a dataset containing 400 records and 24 attributes. They compared their computed outcomes with the actual results to measure the prediction accuracy, employing assessment metrics such as accuracy, sensitivity, precision, and specificity. They discovered that the decision tree method achieved an accuracy of 98.6%, a sensitivity of 0.9720, and perfect scores in both precision and specificity. Charleonnan et al. [4] assessed four machine learning techniques—K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers—to predict CKD using a dataset collected from Apollo Hospitals in India. They compared the performance

of these models to identify the most effective classifier for chronic kidney disease prediction. Their findings indicate that the SVM classifier achieved the highest accuracy at 98.3% and also exhibited the greatest sensitivity after training on the dataset. Boukenze, B. et al. [5] analyzed several algorithms—ANNs, SVM, KNN, and NB—to predict Chronic Kidney Disease, comparing them using the WEKA tool based on prediction accuracy. Their study found that ANN and SVM were the top-performing algorithms, both achieving an accuracy of 62.5%.

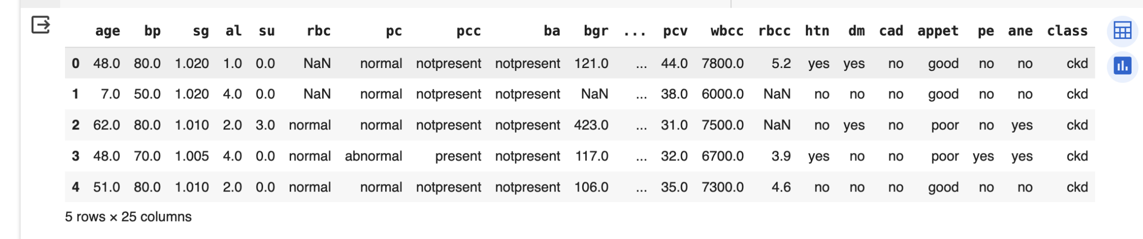
Support Vector Machines: Studies have applied SVM and neural networks to predict CKD. These models are known for their ability to handle complex patterns and large datasets effectively. SVM had good performance with complex patterns; high accuracy, but it can be very computationally expensive. It is more complex than Logistic Regression, potentially offering higher accuracy but at the cost of interpretability and computational inefficiency. Another approach includes the use of decision trees and ensemble methods like Random Forests, which are praised for their interpretability and ease of use in handling both numerical and categorical data. This has high interpretability; effective with incomplete data, but it can suffer from overfitting in some cases. It offers a balance similar to Logistic Regression but with potentially greater complexity and less stability in predictions.

3 Data-set and Features

The dataset used for this project is the Chronic Kidney Disease (CKD) dataset, which is publicly available from the UCI Machine Learning Repository. It can be accessed through the following URL: UCI Machine Learning Repository: CKD Dataset The CKD dataset comprises 400 instances, each representing a patient's medical record. The dataset is divided into Training and Test sets. The dataset includes 24 features, which are a mix of numerical and categorical types. These features provide a comprehensive view of each patient's medical condition and are deemed essential for effectively predicting CKD. The features include:

Numerical: Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, and Red Blood Cell Count.

Categorical (present or not present): Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, and Anemia.



	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44.0	7800.0	5.2	yes	yes	no	good	no	no	ckd
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38.0	6000.0	NaN	no	no	no	good	no	no	ckd
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	31.0	7500.0	NaN	no	yes	no	poor	no	yes	ckd
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	32.0	6700.0	3.9	yes	no	no	poor	yes	yes	ckd
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	35.0	7300.0	4.6	no	no	no	good	no	no	ckd

5 rows × 25 columns

Figure 1: The first five rows of the dataset

4 Exploratory Data Analysis

In the pre-processing stage of the dataset for predicting chronic kidney disease (CKD), a comprehensive examination of the data was conducted to identify and address missing values, which are particularly prevalent in medical datasets. The dataset was scrutinized for any ‘NaN‘ entries. To maintain the integrity of the dataset without discarding valuable data, missing values in categorical variables were imputed with the mode of their respective columns, reflecting the most common category within each feature. For numerical variables, missing entries were replaced with the mean of their columns, providing a statistically balanced approach to handle data absence. Additionally, to streamline the modeling process and enhance algorithmic performance, categorical binary data initially represented as ’Yes’ and ’No’ were transformed into a more computationally efficient binary format, with ’Yes’ mapped to 1 and ’No’ mapped to 0.
In the analysis, I employed data visualizations to enhance my understanding of the chronic kidney disease (CKD) dataset. Histograms for each feature helped us assess the distribution and identify outliers, ensuring robust model training. The heatmap revealed correlations between variables, informing us about potential predictors of CKD and guiding our feature selection. The bar chart focused on the relationship between hemoglobin levels and CKD prevalence, clearly illustrating how variations in hemoglobin are associated with the occurrence of the disease, which supports its potential as a predictive biomarker in our model.

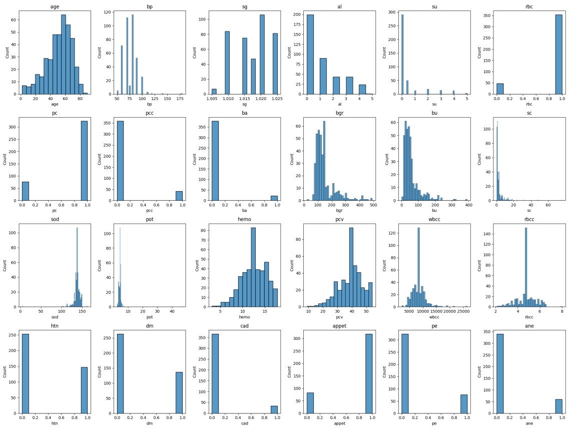


Figure 2: Histogram

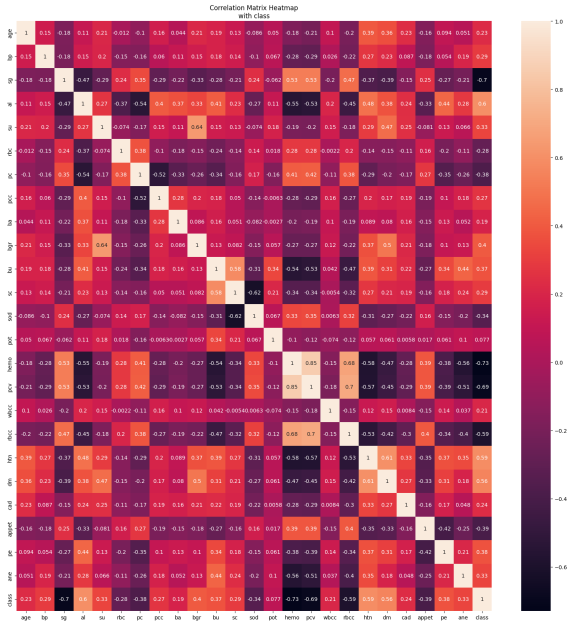


Figure 3: Heat Map

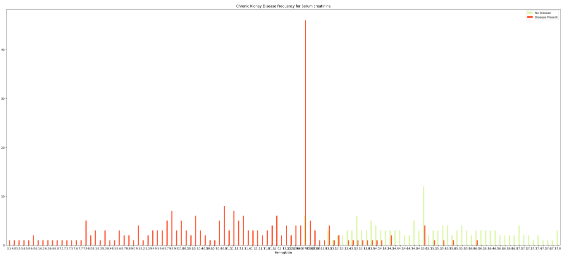


Figure 4: Bar Chart

5 Methods

In predicting chronic kidney disease (CKD), I utilized several machine learning algorithms, each chosen for their specific strengths in handling classification tasks:

Logistic Regression: This model is good for binary classification problems such as CKD prediction. It estimates the probability of the binary outcome (presence or absence of disease) using the logistic function. This function ensures that the model's predictions are confined between 0 and 1, representing probabilities that are converted into classification decisions based on a defined threshold.

$$P(\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{1}$$

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm used for classification by identifying the k closest training examples in the feature space. The prediction for a new instance is determined by a majority vote among its k nearest neighbors, so it is important to select an appropriate k value and distance metric to optimize performance.

Decision Tree: This algorithm segments the dataset into branches to form a tree structure based on decision rules derived from the features. It is particularly useful for visualizing decision-making processes but may require strategies to prevent overfitting, such as pruning.

Random Forest: Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and predicting the outcome based on the majority vote of these trees for classification tasks. This method increases accuracy and robustness by integrating the results of numerous models, each built from a randomly selected subset of training data and features.

6 Result Discussion

In this analysis, several key performance metrics were utilized to evaluate the efficacy of the machine learning models developed for the classification task. These metrics include Accuracy, Precision, Area Under the Curve (AUC), and a Confusion Matrix. Each metric offers unique insights into the performance characteristics of the models:

- **Accuracy** measures the overall correctness of the model and is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total number of samples}}$$

- **Precision** evaluates the model's accuracy in predicting positive labels and is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{False Positives (FP)}}$$

- **Area Under the ROC Curve (AUC)** provides an aggregate measure of performance across all classification thresholds, quantifying the model's ability to discriminate between the positive and negative classes.

- **Confusion Matrix** is a table used to describe the performance of a classification model on a set of test data for which the true values are known. It provides a breakdown of the predictions into four outcomes:

  - **True Positives (TP)**: Correct positive predictions.
  - **True Negatives (TN)**: Correct negative predictions.
  - **False Positives (FP)**: Incorrect predictions where negative instances are predicted as positive.
  - **False Negatives (FN)**: Incorrect predictions where positive instances are predicted as negative.

  The confusion matrix is particularly useful for visualizing the performance of a classifier beyond mere accuracy, as it illustrates the types of errors made by the model.

The results of these metrics are summarized in the following table and were visually supported by ROC curves and plots of model parameters (not shown here).

Model	**TN**	**FP**	**FN**	**TP**	**Accuracy (%)**	**AUC**	**Precision**
Logistic Regression	38	0	1	61	99.0	0.991935	1.0
KNN	38	0	1	61	99.0	0.991935	1.0
Decision Tree	38	0	0	62	100.0	1.00000	1.0
Random Forest	38	0	0	62	100.0	1.00000	1.0

Table 1: Model Performance Metrics where TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive

6.1 Overfitting Considerations

Overfitting is a significant concern in machine learning, where a model learns the details and noise in the training data to an extent that it negatively impacts the performance of the model on new data. To mitigate this, I employed techniques such as k-fold cross-validation and adjusted the complexity of the models through hyperparameter tuning. Specifically, Figure 5 demonstrates the relationship between tree depth and accuracy for the Decision Tree model. It is observed that increasing the tree depth beyond 2 does not significantly enhance the accuracy, suggesting that a depth of 2 is optimal for balancing model complexity and generalization capability.

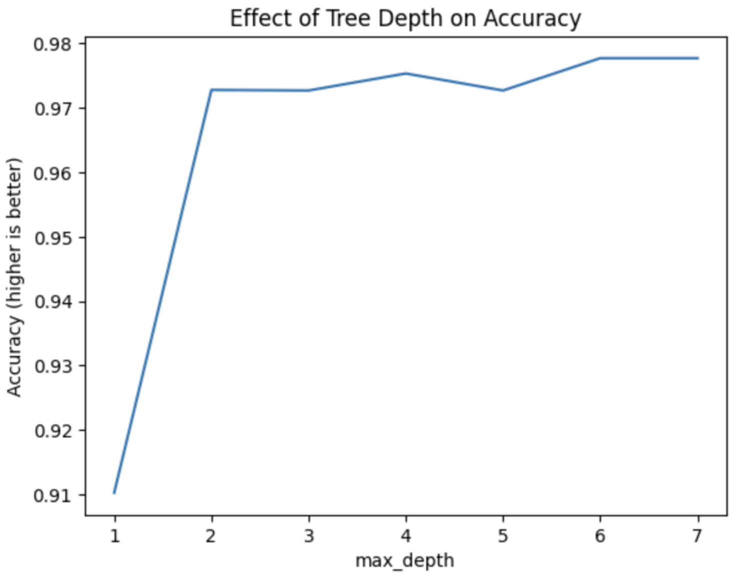
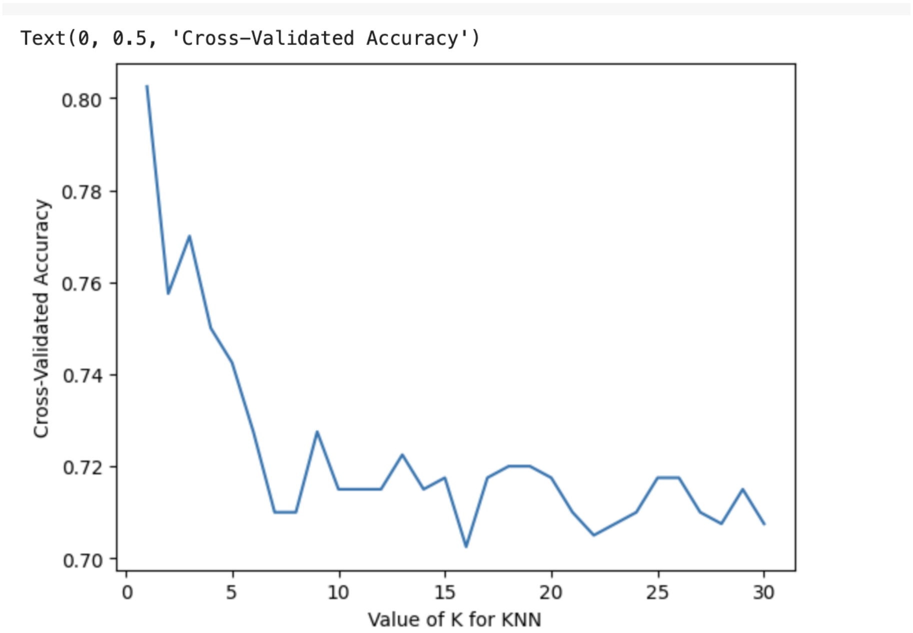


Figure 5: Effect of Tree Depth on Accuracy.

6.2 Computational Efficiency

In pursuit of computational efficiency, I analyzed the impact of varying the number of neighbors in the KNN algorithm along with adjusting the maximum depth in decision trees. As demonstrated in Figure 6, the cross-validated accuracy of the KNN model changes with different numbers of neighbors. Specifically, the accuracy remains high at 0.99 for KNN = 5 and KNN = 6, but begins to decrease to 0.96 for

KNN = 10 and remains at 0.96 for KNN = 20. This suggests that increasing the number of neighbors beyond 5 does not significantly enhance accuracy but potentially increases the computational burden. Consequently, I chose KNN = 5 for our model, which provides a good balance between high accuracy and computational efficiency, preventing unnecessary computational expenses while maintaining optimal model performance.



```
Text(0, 0.5, 'Cross-Validated Accuracy')
```

Figure 6: Cross-Validated Accuracy as a Function of the Number of Neighbors for KNN.

6.3 Discussion

The models have all performed exceptionally well, demonstrating high accuracy and precision. However, despite the Random Forest model achieving perfect scores, the Decision Tree model is preferred for this analysis. Given the simplicity and the non-complex nature of our dataset, the Decision Tree sufficiently captures the necessary patterns without the additional complexity and computational cost associated with the Random Forest. This decision not only ensures efficiency but also preserves computational resources, making the Decision Tree model the most suitable choice for this scenario.

7 Conclusion/Future Work

This study evaluated the effectiveness of Logistic Regression, KNN, Decision Tree, and Random Forest models in predicting Chronic Kidney Disease (CKD), revealing high accuracy and precision across all models, with the Decision Tree particularly noted for its balance of simplicity and effectiveness. For future enhancements, I propose exploring additional classifiers such as SVM and deep learning algorithms, incorporating more comprehensive datasets to improve generalization, and developing a user-friendly interface for the application. These steps aim to enhance the models' accuracy and accessibility, facilitating easier adoption in clinical settings for real-time diagnostic support.

References

[1] W.H.S.D Gunarathne, K.D.M Perera, and K.A.D.C.P Kahandawaarachchi. Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd). In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 291–296, 2017. doi: 10.1109/BIBE.2017.00-39.

[2] S Dilli Arasu, R Thirumalaiselvi, et al. Review of chronic kidney disease based on data mining techniques. *International Journal of Applied Engineering Research*, 12(23):13498–13505, 2017.

[3] Sahil Sharma, Vinod Sharma, and Atul Sharma. Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis, 2016.

[4] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, and Nitat Ninchawee. Predictive analytics for chronic kidney disease using machine learning techniques. *2016 Management and Innovation Technology International Conference (MITicon)*, pages MIT–80–MIT–83, 2016. URL https://api.semanticscholar.org/CorpusID:37366631.

[5] Basma Boukenze, Hajar Mousannif, and Abdelkrim Haqiq. Performance of data mining techniques to predict in healthcare case study : Chronic kidney failure disease. *International Journal of Database Management Systems*, 8:1–9, 06 2016. doi: 10.5121/ijdms.2016.8301.