*Predicting Chronic Kidney Disease*

*Through*

*Machine Learning Techniques*

Project Report

IST 707: Applied Machine Learning

Professor Dunham

Joshua Olando

Anthony Vo

Samantha Zygmont

***Table of Contents***

Introduction………………………………………………………………………………….....3-4

Literature Review………………………………………………………………….…………...4-5

Data……………………………………………………………………………………..……..6-11

Methods………………………………………………………………………………….…..11-12

Results…………………………………………………………………………….…………13-16

Discussion………………………………………………………………………..…………..15-16

Limitations……………………………………………………………………..…………….16-17

Future Work…………………………………………………………..……………………...17-18

Citations………………………………………………………………………………………….19

***Introduction***

The goal of this project is to develop a machine learning model that can predict whether an individual has Chronic Kidney Disease (CKD). Through using the “Chronic Kidney Disease” dataset from UCLA Irvine, many models were explored and tuned to provide an accurate and efficient tool in diagnosing CKD. Our primary objective is to create a model that can support healthcare decision-making by identifying at-risk individuals, thereby facilitating early intervention and improving patient outcomes.

There are a diverse range of stakeholders who would have a strong interest in this model, each benefiting from its application in different ways. Healthcare providers, such as doctors and nurses, are primary stakeholders as they can use the model to directly enhance diagnostic accuracy and support treatment decisions. Pharmaceutical companies and clinical researchers would also benefit from this model to improve the precision of drug targeting and identify eligible patients for clinical trials. Insurance companies are another potential stakeholder, because they can use the model to evaluate patient risk and ensure drug coverage aligns with the needs of the patients. Patients are also potential stakeholders because they would benefit from early diagnosis, more personalized care, and improving their health outcomes. Finally, government agencies would benefit from implementing this model to inform public health initiatives, allocate resources effectively, and monitor CKD trends.

This model aims to be at an early stage in addressing these stakeholders’ needs. The results of our analysis show that the model demonstrates strong promise, with high accuracy in predicting CKD. With continued data collection, future testing, and refinement, the model has the potential to become a valuable tool in early detection and management of CKD. Ultimately, this work represents a critical first step towards leveraging machine learning to improve healthcare outcomes and address the growing need for early CKD detection.

***Literature Review***

Chronic Kidney Disease is a growing public health issue that affects millions of people worldwide, yet remains widely undiagnosed, particularly in the early stages. The disease is a long-term condition characterized by the gradual loss of kidney function. According to the National Kidney Foundation, “1 in 3 adults in the U.S. are at risk for kidney disease, more than 90,000 people are on the kidney waitlist and 12 die every day, and 1 in 7 U.S. adults have kidney disease and 90% don't know it” (*National Kidney Foundation*). This alarming statistic underscores the urgent need for improved early detection methods. Our group member, Joshua Olando, falls into this statistic in our group of three, and is someone with a form of kidney disease called Polycystic Kidney Disease (PKD). PKD is a congenital disease that can lead to significant kidney damage yet most patients, including Joshua, are often unaware they have it until it is detected by ultrasound or other medical imaging.

Chronic Kidney Disease does not have many symptoms in early stages. The prominent symptoms include nausea, vomiting, loss of appetite, change in urination, and many more, however these symptoms are usually mild in early stages and similar to that of many other diseases, making it very hard to initially detect and diagnose as CKD (Mayo Clinic). However, early detection is the best chance for stopping the progression of this disease. In the late stages of CKD progression, people may need to be put on dialysis due to poor kidney function or ultimately need a kidney replacement if they have complete loss of function (Cleveland Clinic). The challenging nature of early CKD diagnosis in conjunction with the deleterious outcomes of late-stage diagnosis exemplify the importance of creating this model for early detection.

Our project involves several key stakeholders, each with unique needs that this model seeks to address. **Healthcare providers** require efficient diagnostic tools that can support timely decisions and interventions. **Pharmaceutical companies** need a tool for targeting clinical trials or therapeutic interventions, particularly in early-stage kidney disease. **Clinical researchers** could use the model to further understand CKD progression and identify factors that contribute to its onset. **Insurance companies** may find value in the model for risk assessment and to more effectively manage healthcare costs associated with treatment. **Patients** stand to benefit from earlier detection, leading to improved treatment outcomes, better quality of life, and potentially lower healthcare costs. Finally, **government agencies** can leverage the model for public health initiatives, optimizing resource allocation and addressing the growing burden of kidney disease.

The need to address these stakeholders’ concerns lies in the fact that CKD is often diagnosed too late, resulting in more costly and less effective treatments. Early detection through accurate predictive models can help mitigate these challenges. To achieve this goal, we utilized machine learning techniques, which has shown significant promise in healthcare for predictive analytics in other diseases such as Alzheimer’s and COVD-19 (Habehh). Given the rapid advancements in machine learning and its ability to process large datasets with high accuracy, it is well-suited to handle the complexity and variability of medical data, such as the variables associated with CKD.

***Data***

The “Chronic Kidney Disease” dataset used in this analysis was sourced from the UCI Machine Learning Repository and donated by the UCLA Institute for Computational Medicine. The dataset was donated on July 2, 2015, and although the dataset is from 2015, the health information it contains remains highly relevant. The dataset consists of both numeric and categorical variables, capturing various aspects of kidney health and related medical conditions. This dataset is considered reliable and suitable for our project due to the comprehensive nature of its features and UCLA and UCI Machine Learning Repository being credible sources. The data is clean, well-documented, and contains predictor variables and a target variable, which indicates whether the individual has CKD.

The dataset contains **24 columns** and **400 rows**, with a relatively balanced distribution of individuals diagnosed with CKD and those without. This data only has 400 rows, making it a relatively small data set for machine learning. The small sample size was taken into consideration with our analysis. The target variable in the data set called “class” is a binary variable indicating whether the individual has CKD, “ckd”, or does not have CKD, “notckd.” There was some missing data for each of the predictor variables, but there was no missing data for the target variable. While there was a relatively small number of rows, there was a sizeable number of columns. The columns included a mix of 14 **numeric variables** and 10 categorical variables. Table 1 provides an overview of the key numeric variables in the dataset, including their abbreviations, average values, and ranges. One important numeric variable is age, with an average of 51 years and range from 2 to 90 years, suggesting that CKD can affect people of all ages. Another key numeric feature is blood pressure, with an average of 76 mm/Hg and range from 50 to 180 mm/Hg. Blood pressure is an important health indicator in this data set considering that hypertension is often associated with kidney disease (*National Kidney Foundation)*. Another notable numeric feature is serum creatinine, with an average of 3.07 mg/dl and a very wide range from 0.4 to 76 mg/dl. Elevated creatinine levels are indicative of kidney dysfunction, so this variable has the potential for being a strong indicator of CKD (*National Kidney Foundation)*. Many of these numeric variables are therefore important for assessing kidney function.

Table 1: Average Values and Range of Numeric Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Numeric Variable** | **Abbreviation** | **Average** | **Range** |
| Age (years) | age | 51 | 2 to 90 |
| Blood Pressure (mm/Hg) | bp | 76 | 50 to 180 |
| Specific Gravity | sg | 1.02 | 1.005 to 1.025 |
| Albumin | al | 1.02 | 0 to 5 |
| Sugar | su | 0.45 | 0 to 5 |
| Blood Glucose Random (mgs/dl) | bgr | 148 | 22 to 490 |
| Blood Urea (mgs/dl) | bu | 57.4 | 1.5 to 391 |
| Serum Creatinine (mgs/dl) | sc | 3.07 | 0.4 to 76 |
| Sodium (mEq/L) | sod | 137.5 | 4.5 to 163 |
| Potassium (mEq/L) | pot | 4.62 | 2.5 to 47 |
| Hemoglobin (gms) | hemo | 12.5 | 3.1 to 17.8 |
| Packed Cell Volume | pcv | 38.88 | 9 to 54 |
| White Blood Cell Count (cells/cmm) | wbcc | 8406.12 | 2200 to 26400 |
| Red Blood Cell Count (millions/cmm) | rbcc | 4.71 | 2.1 to 8 |

The categorical variables are summarized in Table 2. The categorical variables include mostly binary responses, such as hypertension, diabetes, and anemia, and a nominal response for red blood cells. These categorical variables also have valuable health insights that are important in diagnosing kidney disease.

Table 2: Categorical Variables Abbreviations and Descriptions

|  |  |  |
| --- | --- | --- |
| **Categorical Variables** | **Abbreviation** | **Description of Response Types** |
| Red Blood Cells | rbc | nominal (1.005, 1.010, 1.015, 1.020, or 1.025) |
| Pus Cell | pc | normal or abnormal (binary) |
| Pus Cell Clumps | pcc | present or not present (binary) |
| Bacteria | ba | present or not present (binary) |
| Hypertension | htn | yes or no (binary) |
| Diabetes Mellitus | dm | yes or no (binary) |
| Coronary Artery Disease | cad | yes or no (binary) |
| Appetite | appet | good or poor (binary) |
| Pedal Edema | pe | yes or no (binary) |
| Anemia | ane | yes or no (binary) |

The numeric variables were further analyzed based on correlations. A correlation heatmap, shown in Figure 1, was used to visualize relationships between the numeric variables in the dataset. The heatmap revealed several strong positive correlations between certain features, such as packed cell volume, hemoglobin, red blood cell count, and specific gravity. Based on this information, we can infer that these variables are closely related and likely provide similar insights into an individual's health. For instance, higher values in PCV, hemoglobin, and RBC count are often associated with better blood oxygen-carrying capacity, which may correlate with kidney function as well (*National Kidney Foundation*).

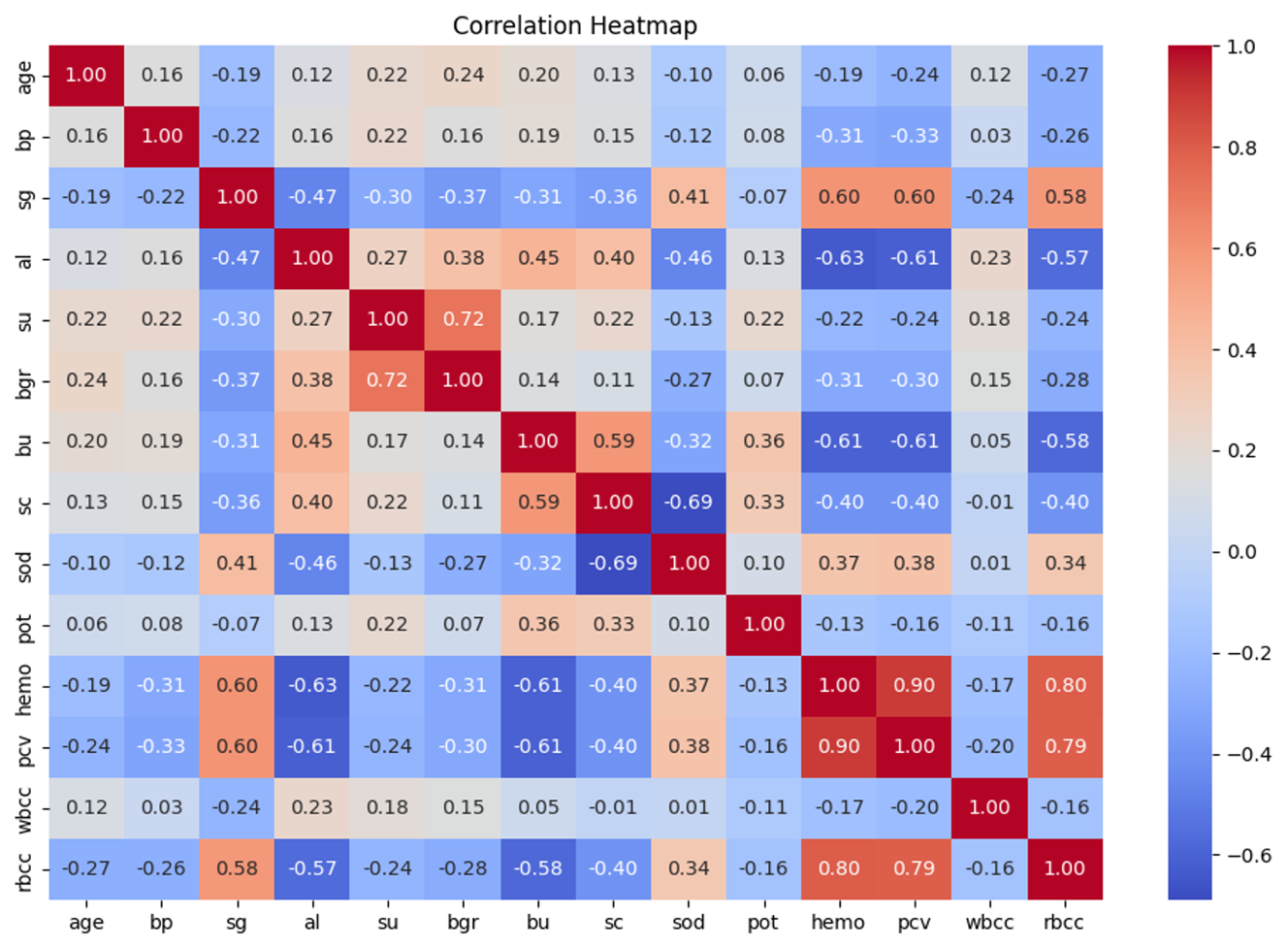


Figure 1: Correlation Heat Map of Numerical Data from the data frame

The age variable was also analyzed thoroughly to understand the distribution of age and whether the individual has CKD. The **age distribution for individuals with CKD in Figure 2** illustrates a high concentration of cases in older age groups, centered around 60 years. However, there are individuals diagnosed with CKD at younger ages as well. This distribution indicates that CKD can affect individuals across various age groups but is more prevalent in older individuals. The age distribution for individuals without CKD is illustrated in **Figure 3**. This plot shows a much more normal distribution centered around age 45. It is important to have a wide range of data for individuals without CKD to successfully train the model to be able to differentiate between the two groups across different age demographics.

A graph of a number of individuals with chronic kidney disease

Description automatically generated

Figure 2: Distribution of Ages in the Data Set for Individuals with Chronic Kidney Disease

A graph of age distribution

Description automatically generated

Figure 3: Distribution of Ages in the Data Set for Individuals without Chronic Kidney Disease

Overall, this data set represents a wide range of health indicators important for predicting CKD. This data is reliable and relatively clean, making it a very adequate data set to be used for machine learning and predicting CKD.

***Methods***

First, we loaded the data set from the UCI Machine Learning Repository into a Pandas data frame. After exploring the data, we confirmed the size of the data frame with our 24 predictor variables, target variable, and 400 rows. Because of the size, cross-validation was used to help split the data into multiple folds and train the model on different subsets. Additionally, we can ensure with cross-validation that the model's performance is not dependent on a particular train-test split, reduce overfitting by evaluating the model on unseen data multiple times, and in our case where data is limited, cross-validation allows for better utilization of the available data such that each data point gets to be in the training set and the validation set, providing a more comprehensive evaluation. However, because we’re using cross-validation, our code needs to be set up and preprocess the data to prevent any data leakage.

Our step for preprocessing the data was to utilize a pipeline to scale, impute, and encode it. It begins by importing necessary libraries and defining lists of numerical and categorical columns in the data set because certain transformations can only be done to one or the other. The numerical\_transform function applied Winsorization to limit extreme values, imputed missing values using KNNImputer, and scaled the data with MinMaxScaler. The categorical\_transform function imputed missing categorical values with the most frequent value and encoded them using LabelEncoder. These transformations were then wrapped in the FunctionTransformer object to ensure they can be applied to the data. The ColumnTransformer was then used to combine these transformations, applying the numerical transformations to numerical features and categorical transformations to categorical features. This setup ensured that both numerical and categorical data were preprocessed appropriately before any modeling. The final preprocessor object can be used in our machine learning pipeline to streamline the preprocessing steps.

The next method of our code focused on building the machine learning models to predict chronic kidney disease. Four distinct algorithms were applied: Logistic Regression, Decision Tree, SGD Classifier, and Random Forest, offering diverse approaches to the prediction task. Each model was trained and evaluated using cross-validation of five folds offering a good mix between bias, variance, and computational cost. The F1 score served as the primary metric for comparing model effectiveness. In addition to printing out the results, we used visual representations of the results to aid in identifying the most promising algorithm for chronic kidney disease prediction.

Lastly, the code optimized each model's performance by systematically exploring different hyperparameter combinations. This process involved defining a search space with potential hyperparameter values for each model type. Using GridSearchCV, all possible combinations within this search space were evaluated. Each combination was then tested using 5-fold cross-validation again to ensure reliable assessments. The primary goal was to find the hyperparameter set that maximizes the model's average F1 score across the cross-validation folds. Finally, GridSearchCV identified and reported the best-performing hyperparameters, leading to the selection of the optimal model configuration for each algorithm. Additional code was added to summarize each model by its test accuracy, cross-validation accuracy, and average F1 score. The data was then compared in a bar chart to summarize the average F1 score for each model before and after tuning.

***Results***

After testing multiple models, including Logistic Regression, Decision Tree Classifier, SGD, and Random Forest, we evaluated their performance. The results of the initial model analysis before tuning are illustrated in Figure 4. The figure shows the average F1 scores, calculated based on cross-validation with cv=5, for each of the four models. The figure shows the model with the best F1 score before tuning was SGD with an F1 of 0.979, then Random Tree Classifier with an F1 of 0.973, then Logistic Regression with an F1 of 0.971, and lastly, Decision Tree Classifier with an F1 of 0.948. Therefore, based on this initial analysis before tuning, SGD classifier had the highest average F1 score.

A graph of a number of different colored bars

Description automatically generated with medium confidenceFigure 4: Comparison of the Model’s Performance Prior to Hyperparameter Tuning

After the hyperparameters were tuned, the model performance was again compared. In Figure 5, the model’s average F1 score is illustrated before and after tuning for each model tested. The Logistic Regression model, the SGD model, and the Random Forest model’s average F1 scores improved after tuning. However, the Decision Tree model’s F1 score worsened after tuning, going from 0.948 before tuning to 0.898 after tuning. The highest average F1 score after tuning was that of SGD classifier with an F1 of 0.990, then Logistic Regression with an F1 of 0.987, then Random Forest with an F1 of 0.980, and lastly Decision Tree with an F1 of 0.898. Based on the outputs, the optimal hyperparameters for the SGD model were found to be **alpha = 0.0001, loss = hinge, max\_iter = 1000,** and**penalty = l2,** which led to the SGD classifier F1 score increasing to 0.990. Based on this information, even after tuning, the SGD model remains the best performing model based on the average F1 score.

A graph showing a comparison of model performance

Description automatically generated

Figure 5: Comparison of the Model’s Performance Before and After Hyperparameter Tuning

The accuracy of the models was further summarized in Table 3 for test accuracy, cross-validation accuracy, and the average F1 score. The results showed that the SGD model outperformed the others in both test accuracy and F1 score, with a test accuracy of 98.75%, CV accuracy of 98.00%, and an F1 score of 99.01%. In comparison, Logistic Regression achieved a test accuracy of 97.50%, CV accuracy of 96.75%, and an F1 score of 98.68%. The Decision Tree model demonstrated a similar test accuracy of 97.50% but had a significantly lower F1 score of 89.75%. Random Forest also achieved a test accuracy of 97.50%, CV accuracy of 97.25%, and an F1 score of 97.99%.

Table 3: Summary of Classifiers and their respective Test Accuracy, Cross-Validation Accuracy, and F1 Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Logistic Regression | Decision Tree Classifier | SGD Classifier | Random Tree Classifier |
| Test Accuracy | 97.50% | 97.50% | 98.75% | 97.50% |
| Cross Validation Accuracy (cv=5) | 96.75% | 96.50% | 98.00% | 97.25% |
| F1 Accuracy | 98.68% | 89.75% | 99.01% | 97.99% |

***Discussion***

Based on the results of this analysis, it was evident that all four models, Logistic Regression, Decision Tree, SGD Classifier, and Random Forest, demonstrated promising performance in predicting chronic kidney disease. Initially, all models achieved average F1 scores above 0.9 during cross-validation. Hyperparameter tuning further enhanced the performance of most of the models, leading to even higher F1 scores, with the exception of the Decision Tree Classifier. The SGD model had the most significant improvement after tuning, making it our highest-performing model. The optimal hyperparameters for SGD were an alpha of .0001, perceptron for the loss, max iterations of 1000, and elastic net for the penalty. The Decision Tree Classifier exhibited interesting behavior, in that its average F1 score decreased after hyperparameter tuning. We hypothesized that the F1 score of the Decision Tree classifier may have decreased after hyperparameter tuning due to overcomplicating the model, poor parameter options, or potentially overfitting. However, all the other models remained very competitive with exceptionally high F1 scores.

Overall, we are very happy with the results of this analysis. We were able to achieve excellent F1 scores, extremely high test accuracies, and extremely high cross-validation accuracies. We believe the project addresses stakeholder needs by providing a tool for early and accurate detection of CKD, leading to improved patient outcomes and advancements in medicine.

***Limitations***

There were some limitations with the “Chronic Kidney Disease” data set that could impact the accuracy and generalizability of the model. First, the sample size of only 400 responses is relatively small, which can limit the model's ability to learn from the training data. This constraint may limit the model’s accuracy and reproducibility when tested with other data sets. Additionally, due to the small sample size, issues such as overfitting and data leakage become significant concerns. Therefore, while we did take measures to prevent data leakage, it is possible that some data was lost and could give us difficulty when testing the model. Another concern with our dataset was that there were 24 predictor variables related to health indicators. Given the large number of variables, it is challenging to determine which of these variables are most significant in predicting CKD. There is also a possibility that important health factors were not included in the dataset, which could further limit the model's predictive power. The data set is also nearly 10 years old, considering that it was acquired in 2015. While the health data is still relevant, it is very possible that medical advances or changes in diagnostic criteria could affect the relevance of some variables over time. These limitations highlight the need for careful consideration of the model’s accuracy and additional testing.

***Future Work***

While we had high success with our model during this machine learning project, we recognize that this is just the start for creating a model capable of accurately predicting CKD in diverse populations. The first area we would like to focus on in the future is improving the model by utilizing larger, more diverse datasets from a range of healthcare sources. A larger sample size would help to refine the model and enhance its performance across different demographics. Additionally, we aim to test the model’s accuracy by collecting data from known individuals with and without CKD, to further evaluate its accuracy. Early detection is another critical objective, and we plan to further tune the model to identify at-risk patients even before they show clinical symptoms.

Furthermore, we hope to expand the scope of our work by investigating potential environmental or lifestyle factors that may influence CKD risk. After identifying at-risk patients, we will conduct population studies to understand how factors such as diet, exercise, and socio-economic status may be correlated with CKD progression. Addressing the needs of our stakeholders is also central to the success of this model. In the future, we hope that this model can be tuned with close collaboration with healthcare providers, pharmaceutical companies, clinical researchers, insurance companies, government agencies, and most importantly patients, to ensure the model is meeting their needs. Through these future efforts, we hope to use machine learning and the power of data science to enhance both the clinical utility and societal impact of our predictive model for CKD detection.

***Citations***

1. *National Kidney Foundation*, www.kidney.org/. Accessed 11 Nov. 2024.
2. Mayo Clinic. "Chronic Kidney Disease." Mayo Clinic, Mayo Foundation for Medical Education and Research, 6 Jan. 2023, [www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521. Accessed 11 Nov. 2024](http://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521.%20Accessed%2011%20Nov.%202024).
3. Cleveland Clinic. "Chronic Kidney Disease." Cleveland Clinic, Cleveland Clinic Foundation, www.my.clevelandclinic.org/health/diseases/15096-chronic-kidney-disease. Accessed 11 Nov. 2024.
4. Habehh, Hafsa, and Suril Gohel. “Machine Learning in Healthcare.” *Current genomics* vol. 22,4 (2021): 291-300. doi:10.2174/1389202922666210705124359
5. "Chronic Kidney Disease." UCI Machine Learning Repository, University of California, Irvine, 2014, archive.ics.uci.edu/dataset/336/chronic+kidney+disease.