# Project 2 – CKD Prediction Study Proposal

## Business Problem

The Chronic Kidney Disease (CKD) Prediction Study attempts to predict the probability of CKD. Using three different models, the study uses a single dataset to compute the predictions and uses 2 additional datasets as a tertiary source for verifying the predictions. The 2 additional datasets contain only summary data, but the associated studies provide the detail with which to compare the model's results.

## Background

**Definition:** Chronic Kidney Disease is defined by (*Facts About Chronic Kidney Disease*, 2022) as a decrease in the kidney's ability to filter wastes from the body. When kidney function levels decrease, other health problems become prominent. The degree of kidney function reduction is measured by an Estimated Glomerular Filtration Rate test (eGFR). This is a simple blood test that shows the levels of creatinine in the blood. Creatinine is a waste product and is normally excreted through the kidneys. When the kidneys have a reduced filtration rate, the creatinine accumulates in the blood. The eGFR is a computed value based upon the amount of creatinine. eGFR is not the only factor in determining if someone has CKD, other indicators include the amount of protein in the urine or physical damage to the kidneys *(American Kidney Fund, 2022)*.

**Risk Factors: According** to (*kidney disease*, 2022), major risk factors for CKD are diabetes and high blood pressure. Other factors include heart disease and a familial history of kidney failure. Additionally, other studies have shown that African Americans, Hispanics, and American Indians tend to have a greater propensity for CKD. The inference being these groups have higher rates of comorbidity of diabetes and high blood pressure.

## Datasets

### 1. Early Stage of Indians Chronic Kidney Disease (CKD) Dataset

This dataset is from a study conducted by Dr. P. Soundarapandian. M.D., D.M, a Senior Consultant of Nephrology, Apollo Hospitals Tamilnau, India. The population for the study is 400 individuals in India who do or do not have CKD. The dataset is available from the University California Irvine Machine Learning Repository, *(Dua and Graff, 2019)*.

The dataset contains 25 variables and is used by the CKD Prediction project. The variables names and data formats are shown in Table 1 Dataset Variables in Appendix A.

### 2. CURE-CKD Dataset

The CURE-CKD dataset is the focus of a study conducted by *(Tuttle et al., 2019)*. This study's objective is to identify the prevalence of major risk factors for CKD among patients treated in 2 large US health care systems. The study is an observation of the data within the Center for Kidney Disease Research, Education, and Hope (CURE-CKD) Registry. This registry contains data from 2,625,963 adults and children collected from January 2006 – December 2017. Refer to Table 2 CURE-CKD Dataset for details for the variables common to the study dataset.

### 3. Chronic Renal Insufficiency Cohort Study Dataset

The Chronic Renal Insufficiency Cohort (CRIC) Study examines risk factors for progression of chronic renal insufficiency (CRI) and cardiovascular disease (CVD) (*Chronic-Kidney-Disease Research Chronic Renal Insufficiency Cohort Study Kidney Disease*, n.d.). The population group groups include data collected on adults aged 21 to 74 years with a broad spectrum of renal disease severity. Refer to Table 3 CRIC Dataset for the variable's names.

## Methods

The Prediction study is a binary classification problem and uses 3 different models to predict a single outcome based upon a final set of 13 independent variables. The outcome or dependent variable indicates if an observation has or does not have CKD. The 3 models are a Random Forest Classifier, AdaBoost Classifier, and a XgBoost Classifier. Two feature selection models are used to derive the set of most suitable features. These are the RFE and the RFECV models, with the results from the RFCV being used as the final feature set. Refer to figures 1 and 2 for the feature selection results.

The total number of observations in the input dataset is 400. The dataset is split into a training and test set. An additional dataset serves as a secondary test set referred to as the validation set. The validation dataset is a randomized sample from a kidney disease test dataset (Atul, n.d.). This validation set has 60 observations.

The AdaBoost and XgBoost models are built using an initial set of parameters. Due to the limited number of observations in the training and test sets, a Cross-Validation method is run against each model to confirm the accuracy of each model. Additionally, a GridSearch cross validation method was run against each model to fine tune the hyper-parameters. The validation set is used to compare the prediction results from this set to those with the test sets.

As a tertiary source of validation, the predictions for each model are verified against the findings in the CURE-CKD and the Chronic Renal Insufficiency Cohort (CRIC) Studies documents. This task involves comparing the validation set's predicted variables with findings in the authoritative sources.

## Ethical Considerations

Although the source dataset does not include demographic factors, the other supporting datasets do contain demographic segmentation. Diabetes is known to be more prevalent in Hispanics and Blacks than the White population. Demographic segmentation may cause unwarranted biases. Another factor is the age demographic. With any study involving age must be cautious of injecting natural biases and making unfounded conclusions. The third factor is the disease itself. Having certain diseases automatically categorizes people and attaches associations. This study does not identify individuals in the studies but if it did, the individual's privacy and health information is at risk.

## Challenges/Issues:

The original intent of the study was to conduct a study of the progression of CKD disease. The major difficulty was procuring datasets with adequate content and time interval data with which to conduct Time-to-Event analysis. Although the Chronic Renal Insufficiency Cohort and the CURE-CKD studies do provide for the time interval outcomes, they are summary results and do not include individual observations. The individual observations are restricted information and are only released with prior authorization. Although these studies do not provide for time intervals and only contain summary results, they are valuable validation sources.

**Analysis:**

## Data Analysis

Conducting an Exploratory Data Analysis (EDA) on the dataset shows a strong correlation between Hemoglobin counts and Packed Cell Volume counts. This makes sense as the two clinical factors are biologically related. The correlation between Anemia and Hemoglobin counts has a strong correlation for the same reason. Refer to Figure 3 – Correlation Heatmap for details. The scatterplots in Figure 4 Albumin to Serum Creatinine Levels illustrate the relationships between the Albumin to Serum Creatine levels by CKD status. Again, there is a biological relationship between the two factors. As the Serum Creatinine increases so does the Albumin levels. Figure 5 shows the relationship between the Hemoglobin levels to Serum Creatinine level by CKD Status. As the Hemoglobin counts increase the Serum Creatinine levels decrease. The CRIC study indicates that people diagnosed with CKD have elevated levels of creatinine, experience anemia, and have low hemoglobin counts. Figure 6 – Class to Best Features shows the correlations of the independent variables to the dependent variable, CKD Status or Class. As shown by figure 6, Albumin has the highest correlation with CKD with Hypertension and Diabetes being the 2nd and 3rd highest in the set of 13 best features. As confirmed by the CRIC and CURE-CKD Studies, the two most prevalent risk factors for CKD are Hypertension and Diabetes with Albumin being a strong indicator of CKD. Albumin is a type of blood plasma protein that leaks into the urine in people with CKD.

## Results Analysis

Using the test and training sets, the 3 models produced accuracy scores close to each other. Also, the ROC and PRC show similar results across the models. The overall accuracy score is 97%. Reviewing the test set Confusion Matrix for each model, the results show 3 False Negatives a piece the Random Forest and the AdaBoost models. Refer to figures 7 and 8 for the Confusion Matrices for the Test and Validation set Confusion Matrices. The XgBoost model has a single False Negative. The Confusion Matrices results for the validation set are different from the test set results. The AdaBoost model has the largest disparity between the test set and the validation set. As shown by the Confusion Matrix, the AdaBoost model shows 25 False Negatives. As stated previously, the AdaBoost Confusion Matrix for the test set shows 3 False Negatives. Additionally, the AdaBoost accuracy rates between the two sets are vastly different. The accuracy rate with the test set is 95%. Whereas the accuracy rate using the validation set is 56.33%. This is a difference of 38.7 percentage points. Yet the same validation set is used for the Random Forest and XgBoost models. Using the validation set, Random Forest has an accuracy rate of 85% and an accuracy score of 96% with the test set. The XgBoost model has an accuracy score of 93.33% with the validation set and a score of 98% with the test set. Table 4 – Accuracy Scores lists the scores for each model.

One explanation for the difference in the AdaBoost model's test and validation predictions may be attributable to Overfitting. The AdaBoost algorithm is prone to Overfitting especially with a noisy dataset. The AdaBoost model was run twice and each time the accuracy rate using the test dataset stayed the same. Using the validation set the accuracy score decreased from 56% to 50%. Additionally, the other two models were run twice more with only a reduction of 1 or 2 percentage points using the validation set.

## Results Validation

To assess the validity of the model results, they are compared with two existing published studies, CURE-CKD and the CRIC Study (*Tuttle et al., 2019)* (*Chronic-Kidney-Disease Research Chronic Renal Insufficiency Cohort Study Kidney Disease*, n.d.). The CURE-CKD study reveals that two-thirds of their adult population were eventually diagnosed with diabetes, hypertension, or prediabetes. During the 3 phases of the CRIC Study, the results show a significant percentage of participants with Diabetes and Hypertension. According to (*kidney disease*, 2022), major risk factors for CKD are diabetes and high blood pressure.

The analysis of the data for the CKD-Prediction project also shows a strong correlation between hypertension (htn), diabetes (dm) and the presence of CKD. Also, the CURE-CKD study reveals that the Urine Albumin-Creatinine Ratio (uACR) levels are elevated among those with CKD. An examination of the relationship between of the Albumin (al) to Serum Creatinine (sc) is confirmed by the CURE-CKD and CRIC studies. Although the Feature Selection does not include the Serum-Creatinine(sc)  variable, the models can accurately predict CKD status without this value.

## Conclusion

Out of the 3 models, XgBoost is the most correct. It maintains accuracy across both the test and validation sets. The Random Forest Classifier is a possible contender, but the XgBoost model has much better accuracy rates. The AdaBoost model is prone to overfitting. Even with the application of some overfitting mitigations, the AdaBoost lags the XgBoost model in terms of performance. XgBoost model is much faster in execution and consumes less computer memory than the AdaBoost model.

**Questions:**

1. Why did the AdaBoost model perform much worse with the second test set?

This is probably due to the overfitting problem. Overfitting is when a model does not generalize well enough from the training set to the test set. The model performs adequately on the training set but produces poor results on the subsequent training sets. It doesn't perform well on unseen data.

2. What problems may arise with false negative predictions?

In the case of CKD predictions or for any medical predictions, a False Negative may have detrimental effects. Giving someone a false diagnosis has a serious impact on their lives.

3. Is there a relationship between Serum Creatinine and Albumin levels?

Yes, as the Serum Creatinine levels increase so does the Albumin levels. Albumin is a blood plasma protein that leaks into the urine with someone diagnosed with CKD. Albumin levels is a strong indicator of CKD according to the CRIC study. Additionally, as kidney function decreases, the levels of Serum Creatinine increase. Both clinical factors are closely related to each other and both indicator the presence of CKD.

4. Why does Pedal Edema have a high weighted value?

Pedal Edema is a result of reduced kidney function. As kidney function decreases, they cannot rid the body of excess fluids. This fluid then settles in the lower extremities. Pedal Edema not by itself is an indicator of CKD. But it is a contributing factor to the identification of CKD.

5. Why does Appetite have a high weighted value?

One of the symptoms of approaching kidney failure is loss of appetite. As with Pedal Edema, appetite is not an indicator of CKD, but it is in combination with other contributing factors.

6. Which variables are the most prevalent with a positive CKD status?

As said in both the CURE-CKD and CRIC studies, the two highest contributors to CKD are Diabetes and Hypertension. A positive CKD status is confirmed using a blood test called the Extended Glomeruli Filtration Rate (eGFR). This test measures the amount of creatinine and other factors present in the blood.

7. Will the predictions change with the addition of Serum Creatinine into the features set?

The Serum Creatinine levels should be a contributing factor. But looking at the correlation results with a dataset with the Serum Creatinine, the results show neutral correlation with CKD status. This contrasts with National Kidney Foundation reports (*Facts About Chronic Kidney Disease*, 2022), that Creatinine is the substance that is the basis for the eGFR. It is the eGFR measurements that show the presence of CKD.

## 8. Will the predictions change with the deletion of the Blood Pressure values since the Hypertension values are in the feature set?

Deleting the Blood pressure values does not influence the predictions. Blood pressure values have a neutral correlation with the CKD status

## 9. What are the top 5 factors that show the presence of CKD?

Based upon the results of the RFECV feature selection process, the top five features are Albumin, Diabetes, Hypertension, Blood Glucose Random, and Appetite.

## 10. Is it possible to determine the percentage that Diabetes contributes to CKD?

Computing the weights for each variable using the RFE feature selection method, Diabetes has the 3rd highest weighted value. This contrasts with the findings from the CURE-CKD and CRIC studies, Diabetes is the leading cause of CKD.

## References

1. Facts *About Chronic Kidney Disease*. (2022, October 10). National Kidney Foundation. https://www.kidney.org/atoz/content/about-chronic-kidney-disease

2. American Kidney Fund. (2022, November 30). Blood test: eGFR (estimated glomerular filtration rate). https://www.kidneyfund.org/all-about-kidneys/tests/blood-test-egfr

3. Tuttle, K. R., Alicic, R. Z., Duru, O. K., Jones, C. R., Daratha, K. B., Nicholas, S. B., McPherson, S. M., Neumiller, J. J., Bell, D. S., Mangione, C. M., & Norris, K. C. (2019). Clinical Characteristics of and Risk Factors for Chronic Kidney Disease Among Adults and Children. *JAMA Network Open*, *2*(12), e1918169. https://doi.org/10.1001/jamanetworkopen.2019.18169

4. *Chronic-Kidney-Disease - Research Chronic Renal Insufficiency Cohort Study Kidney Disease*. (n.d.). http://www.cristudy.org/Chronic-Kidney-Disease/Chronic-Renal-Insufficiency-Cohort-Study/CRIC-DataView

5. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

6. Atul, A. (n.d.). *Kidney Disease.csv*. https://raw.githubusercontent.com/AP-Atul/Chronic-Kidney-Disease/master/dataset/train.csv

7. *Kidney Disease*. (2022, December 17). National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/kidney-disease

## Appendix A - Tables and Figures

### Table 1 – UCI Kidney Disease Dataset Variables

| Variable Name | Abbreviation | Data Type |
|---|---|---|
| Age | age | int |
| Albumin | al | int |
| Anemia | ane | binary |
| Appetite | appet | binary |
| Bacteria | bc | binary |
| Blood Glucose Random | bgr | int |
| Blood Pressure | bp | int |
| Blood Urea | bu | float |
| Class | CKD Status | binary |
| Coronary Artery Disease | cad | binary |
| Diabetes Mellitus | dm | int |
| Hemoglobin | hemo | float |
| Hypertension | htn | binary |
| Packed Cell Volume | pcv | float |
| Pedal Edema | pe | binary |
| Potassium | pot | int |
| Pus Cells | pc | binary |
| Pus Cell Clumps | pcc | float |
| Red Blood Cells | rbc | binary |
| Red Blood Cell Count | rbcc | float |
| Serum Creatinine | sc | float |
| Sodium | sod | int |
| Specific Gravity | sg | float |
| Sugar | su | binary |
| White Blood Cell Count | wbcc | float |

### Table 2 – CURE-CKD Dataset Variables

| **Demographics** | **Medical Conditions** | **Clinical Factors** | **Clinical Factors cont...** |
|---|---|---|---|
| **Age (yr)** | Hypertension | Systolic BP (mmHg) | |
| **Sex** | Diabetes | Diastolic BP (mmHg) | |
| **race/ethnicity** | | UACR-urine albumin-to-creatinine ratio | |
| **White** | | UPCR-urine protein-to-creatinine ratio | |
| **Black** | | eGFR category (ml/min per 1.73 m2) | <30 |
| | | | 40 to <50 |
| | | | 50 to <60 |
| | | | >60 |
| **Hispanic** | | | |

### Table 3 – CRIC Dataset Variables

| **Demographics** | **Medical conditions** | **Clinical Factors** |
|---|---|---|
| Age (yr) | Hypertension | Blood Pressure |
| | Diabetes | Roche Adjusted Creatinine |
| | Heart Disease | Urine Protein |
| | | Hemoglobin |
| | | Glucose |

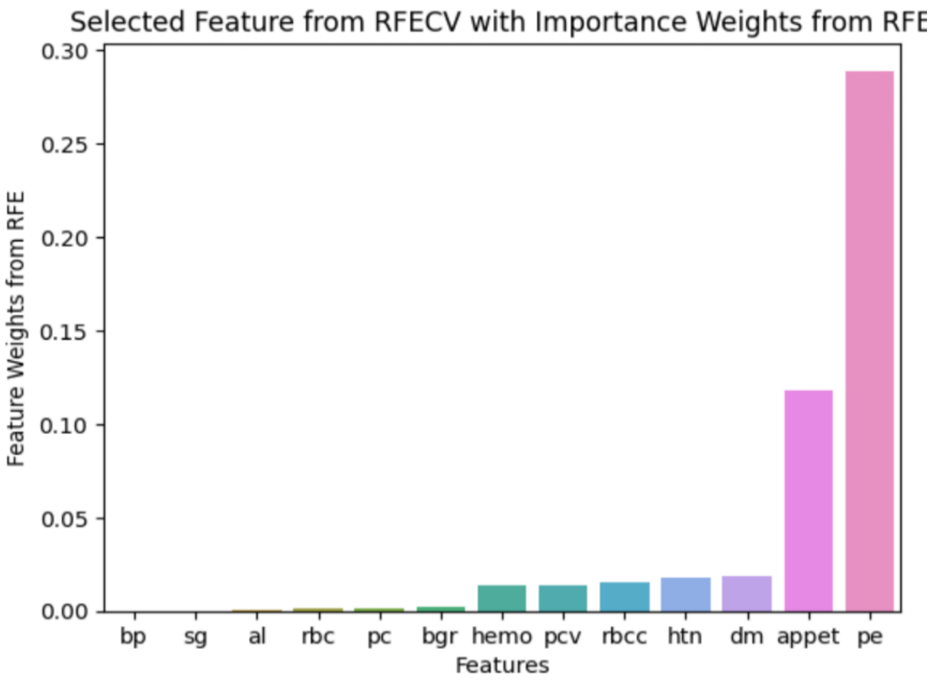


Figure 1 Feature Selection Weights

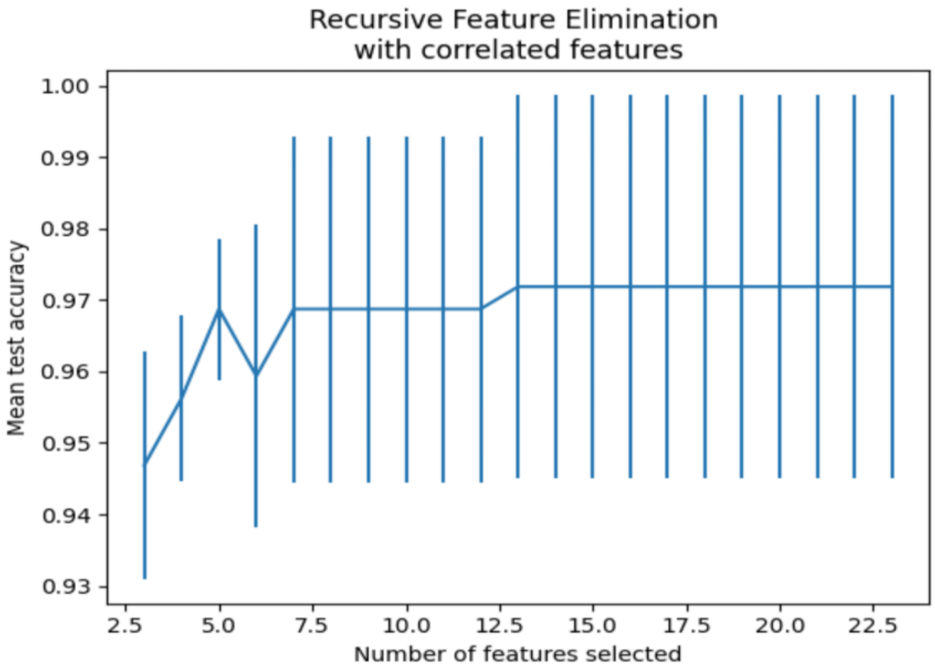


Figure 2 RFE Results

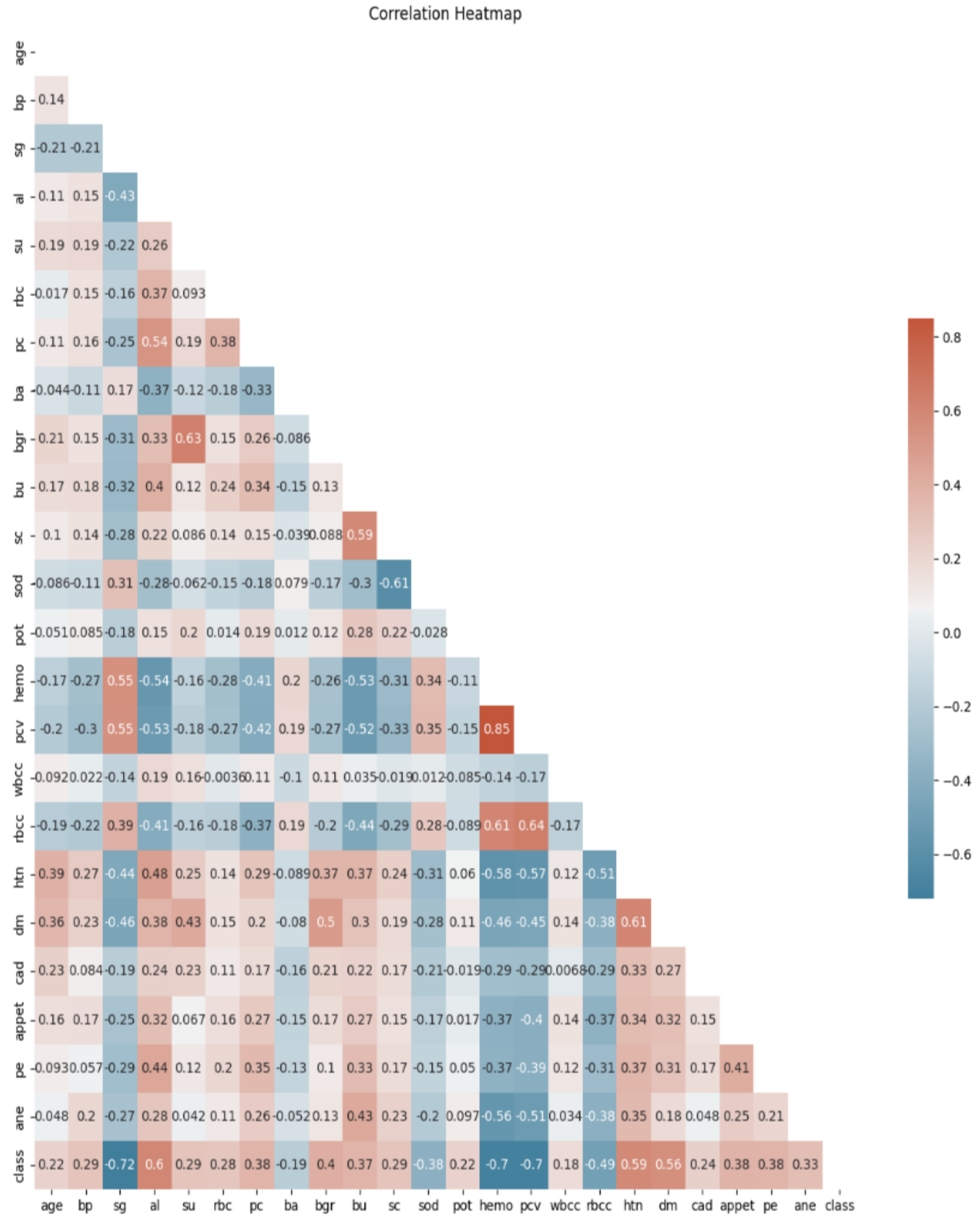


Figure 3 – Correlation Heatmap

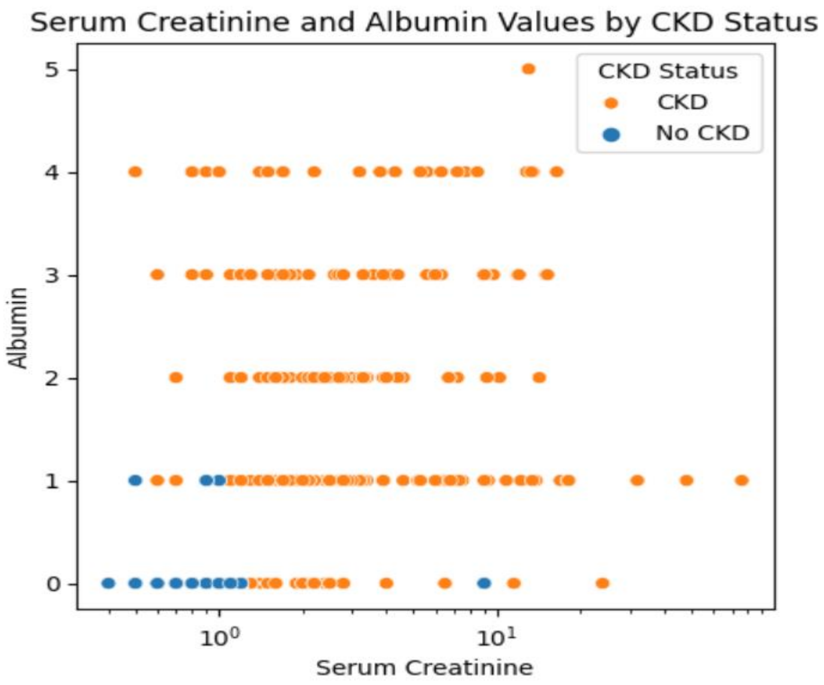


Figure 4 – Hemoglobin vs Red Blood Cells Scatterplot

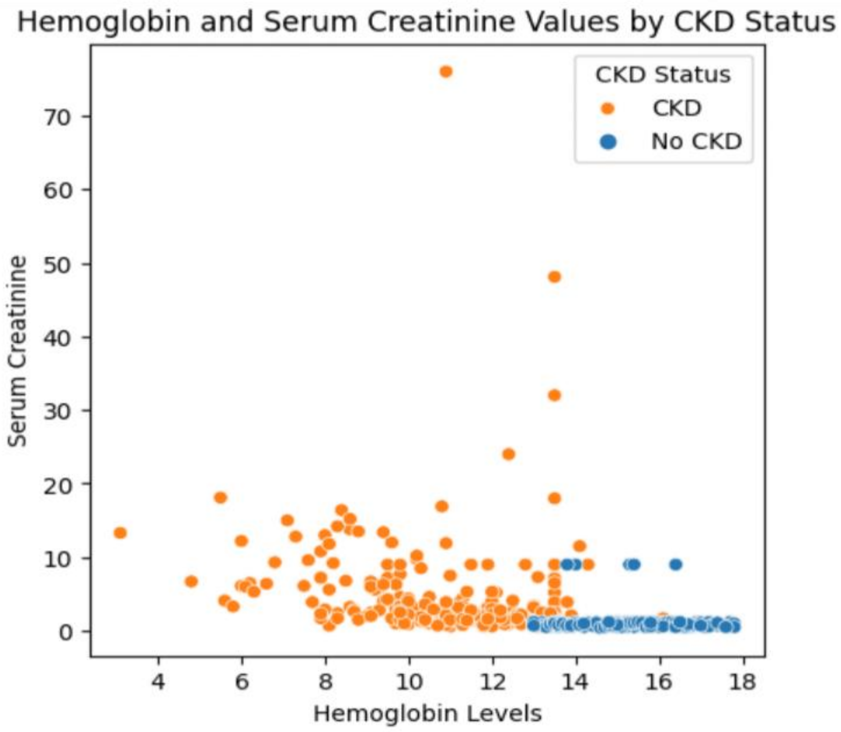


Figure 5 – Hemoglobin vs Serum Creatinine Levels Scatterplot

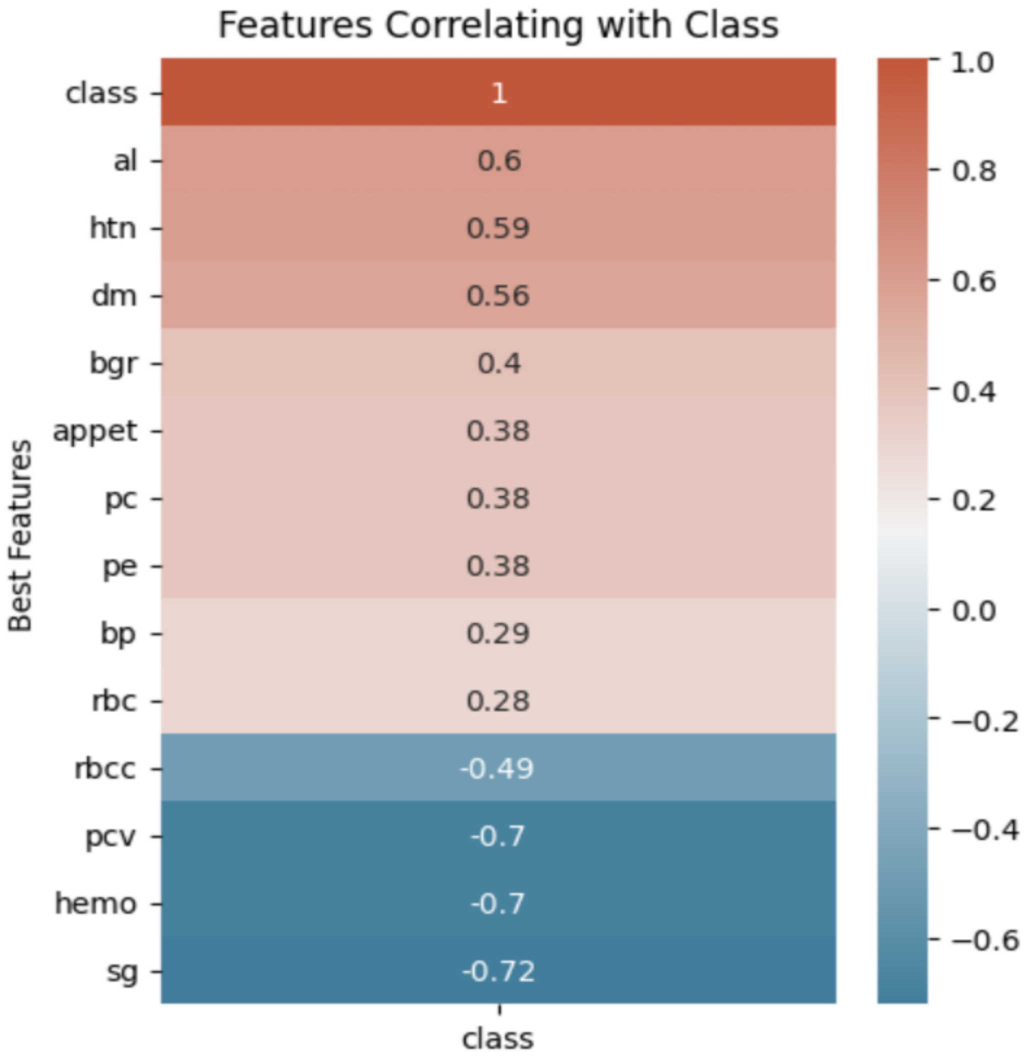
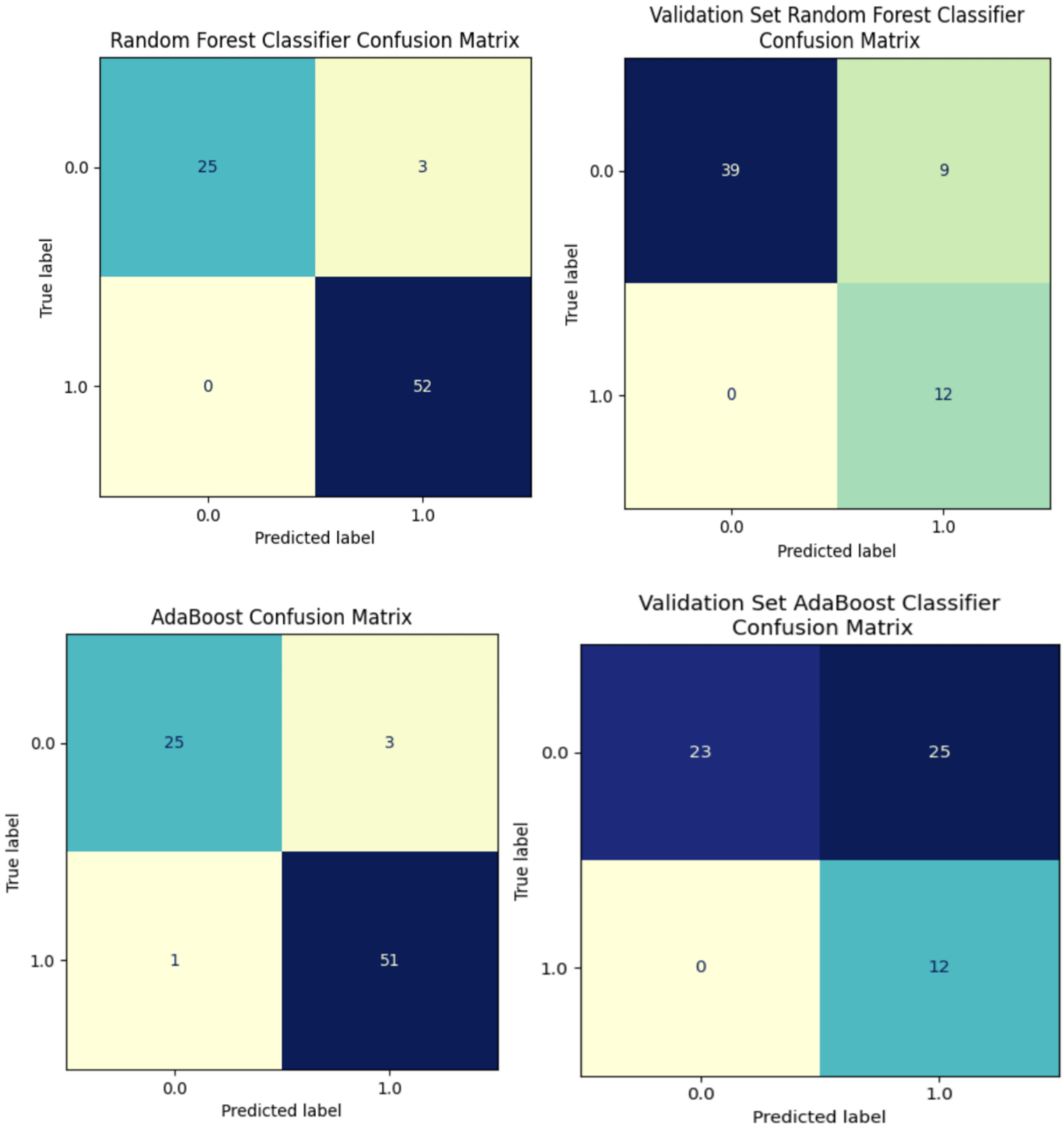


Figure 6 Class Correlations



Random Forest Classifier Confusion Matrix

Validation Set Random Forest Classifier Confusion Matrix

AdaBoost Confusion Matrix

Validation Set AdaBoost Classifier Confusion Matrix

Figure 7 Confusion Matrices – Random Forest and AdaBoost

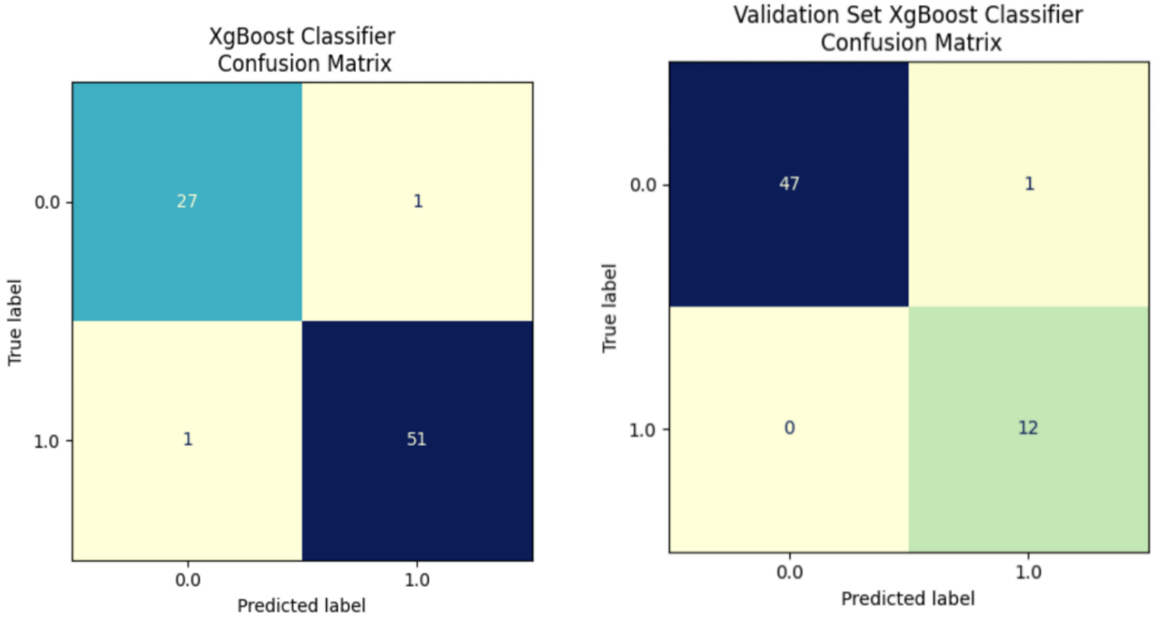


Figure 8 Confusion Matrices – XgBoost

Table 4 Accuracy Scores

| Model | Test Accuracy Score | 1st Validation Accuracy Score | 2nd Validation Accuracy Score |
|---|---|---|---|
| Random Forest | 0.96 | 0.85 | .85 |
| AdaBoost | 0.95 | 0.57 | .5 |
| XgBoost | 0.98 | 0.93 | .92 |