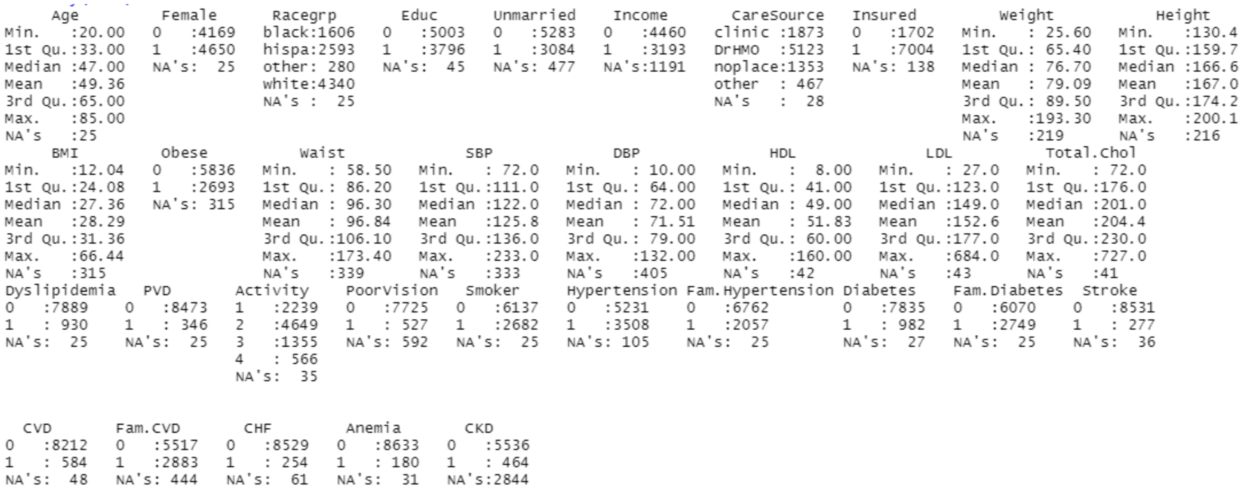
## Introduction

While testing for CKD, or chronic kidney disease, is low-cost, it is an underdiagnosed disease that could benefit from a more effective means of screening to sooner identify those with high-risk indicators for early detection and treatment. The goal of this case study can be divided into two parts: (1) Develop a model that can most effectively predict for CKD where the outcome is not known, (2) Develop a screening tool that maximizes the effectiveness of the model while minimizing its complexity for those surveyed to encourage screening participation. The model is built off a CDC nationwide survey consisting of respondent demographics, self-reported health history, and physical examination and testing from 8819 adult participants.  The first 6000 survey participants have a known outcome of CKD and serve as input for training a predictive model for the remaining 2819 participants where the CKD outcome is unknown.

## Data Overview



```
      Age           Female      Racegrp        Educ       Unmarried     Income      CareSource     Insured        Weight            Height
 Min.   :20.00   0    :4169   black:1606   0    :5003   0    :5283   0    :4460   clinic :1873   0    :1702   Min.   : 25.60   Min.   :130.4
 1st Qu.:33.00   1    :4650   hispa:2593   1    :3796   1    :3084   1    :3193   DrHMO  :5123   1    :7004   1st Qu.: 65.40   1st Qu.:159.7
 Median :47.00   NA's:  25    other: 280   NA's:  45    NA's: 477    NA's:1191    noplace:1353   NA's: 138    Median : 76.70   Median :166.6
 Mean   :49.36                white:4340                                          other  : 467                Mean   : 79.09   Mean   :167.0
 3rd Qu.:65.00                NA's :  25                                          NA's   :  28                3rd Qu.: 89.50   3rd Qu.:174.2
 Max.   :85.00                                                                                                Max.   :193.30   Max.   :200.1
 NA's   :25                                                                                                   NA's   :219      NA's   :216
      BMI           Obese          Waist             SBP              DBP              HDL              LDL          Total.Chol
 Min.   :12.04   0    :5836   Min.   : 58.50   Min.   : 72.0   Min.   : 10.00   Min.   :  8.00   Min.   : 27.0   Min.   : 72.0
 1st Qu.:24.08   1    :2693   1st Qu.: 86.20   1st Qu.:111.0   1st Qu.: 64.00   1st Qu.: 41.00   1st Qu.:123.0   1st Qu.:176.0
 Median :27.36   NA's: 315    Median : 96.30   Median :122.0   Median : 72.00   Median : 49.00   Median :149.0   Median :201.0
 Mean   :28.29                Mean   : 96.84   Mean   :125.8   Mean   : 71.51   Mean   : 51.83   Mean   :152.6   Mean   :204.4
 3rd Qu.:31.36                3rd Qu.:106.10   3rd Qu.:136.0   3rd Qu.: 79.00   3rd Qu.: 60.00   3rd Qu.:177.0   3rd Qu.:230.0
 Max.   :66.44                Max.   :173.40   Max.   :233.0   Max.   :132.00   Max.   :160.00   Max.   :684.0   Max.   :727.0
 NA's   :315                  NA's   :339      NA's   :333     NA's   :405      NA's   :42       NA's   :43      NA's   :41
 Dyslipidemia    PVD        Activity    PoorVision    Smoker      Hypertension Fam.Hypertension Diabetes     Fam.Diabetes   Stroke
 0    :7889   0    :8473   1    :2239   0    :7725   0    :6137   0    :5231    0    :6762       0    :7835   0    :6070   0    :8531
 1    : 930   1    : 346   2    :4649   1    : 527   1    :2682   1    :3508    1    :2057       1    : 982   1    :2749   1    : 277
 NA's:  25    NA's:  25    3    :1355   NA's: 592    NA's:  25    NA's: 105     NA's:  25        NA's:  27    NA's:  25    NA's:  36
                           4    : 566
                           NA's:  35


   CVD          Fam.CVD        CHF         Anemia        CKD
 0    :8212   0    :5517   0    :8529   0    :8633   0    :5536
 1    : 584   1    :2883   1    : 254   1    : 180   1    : 464
 NA's:  48    NA's: 444    NA's:  61    NA's:  31    NA's:2844
```

Figure 1: Descriptive Statistics for the Case Study Dataset

Six main risk factors for CKD were identified in the Case Study source material: Diabetes, hypertension, cardiovascular disease, family history of kidney disease, age, and race (specifically First Nations and Pacific Islanders). All but family history of kidney disease was collected in the CKD dataset. Additionally, the dataset does not accurately capture those races considered at-risk categorizing only by “White”, “Black”, “Hispanic” or “Other”. The correlation table in Figure 2 shows a positive correlation between CKD and the four predictors comparable to those outlined in the Case Study, suggesting the dataset is an accurate sample representation of these established risk factors.

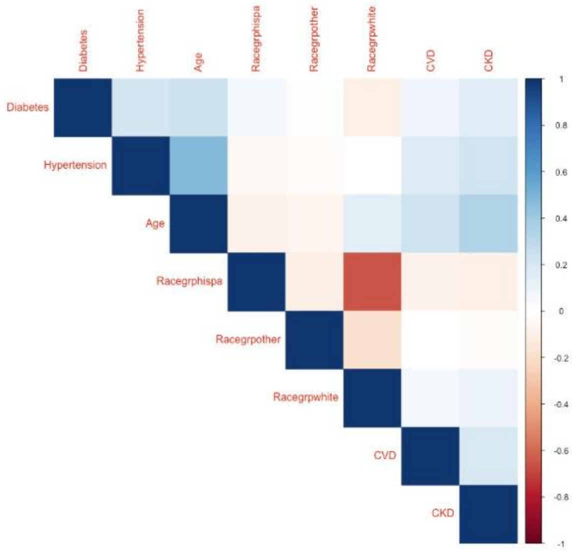


Figure 2: Correlation between CKD and the main risk factors

Missing Data

About 31.5% of the survey participants had at least one missing value among the 32 predictor variables. The distribution of missingness is approximately evenly distributed between the two subsets of respondents with known CKD values and those without. This suggests the missingness between the subsets should not have a significant affect in predictive modeling.

It is clear from Figure 3 the missing data is missing not at random, (MNAR). Some variables contain more missing values than others. 10 variables contain no missing values. The four variables with the most missing data are Income, self-reported vision quality (“PoorVision”), marital status (“Unmarried”), and family history of cardiovascular disease (“Fam.CVD”) at 13.2%, 6.4%, 5.1%, and 4.8% respectively. It is likely survey participants did not wish to share their income or marital status and may not have had sufficient knowledge of their vision quality or family history. Additionally, missing values of some variables are highly correlated with missing values in other variables. For example, participants without a blood sample drawn did not have values entered for their cholesterol (LDL, HDL, and total cholesterol). Participants without a complete examination often had height, weight, BMI, obesity, and waist information missing or systolic and diastolic blood pressure missing.

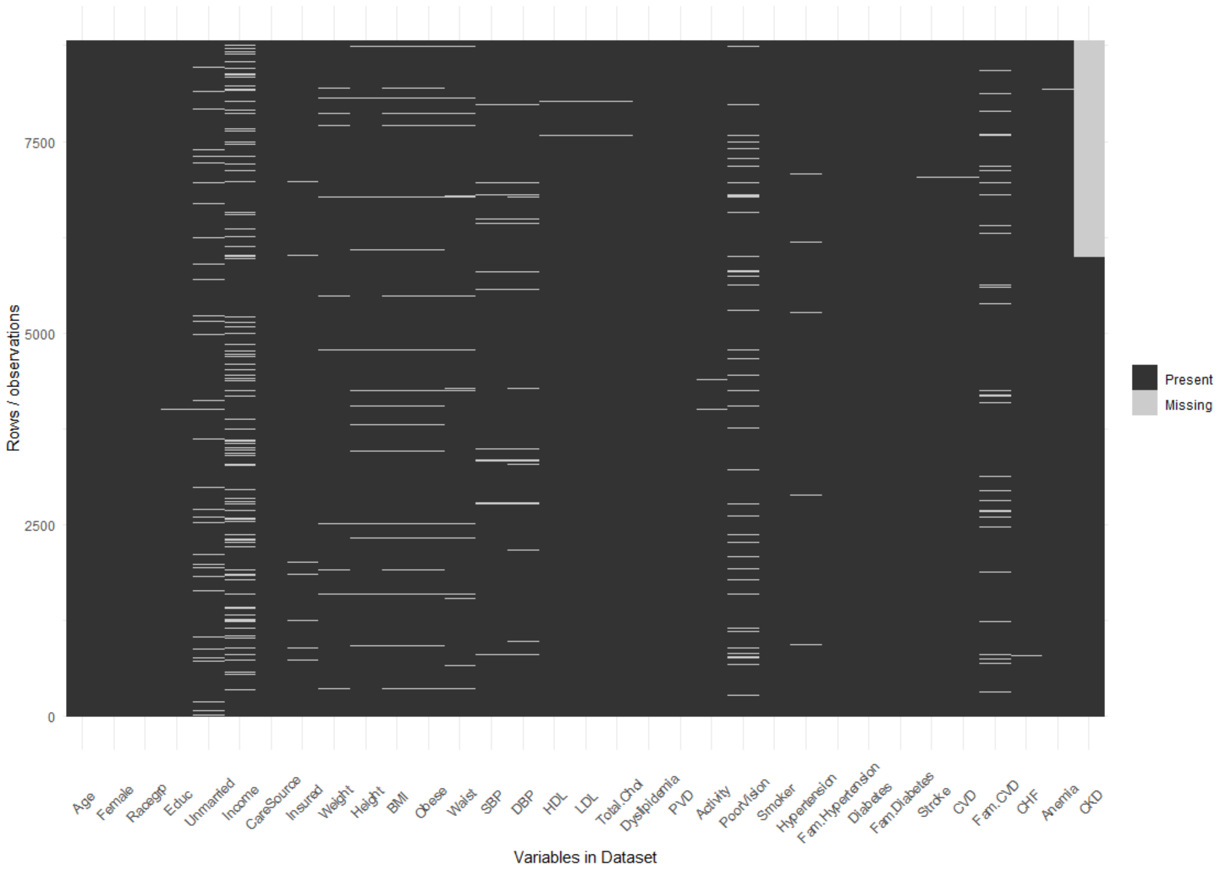


Figure 3: Visualization of missing data

The missing data not being at random suggests any imputation methods would impart some degree of bias on the dataset. For example, the income variable asks participants if their household is below (0) or above (1) the median income. Figure 4 clearly shows below the median income is overrepresented before imputation and may be a result of the missing data. If so, imputation did not effectively correct for this disparity.

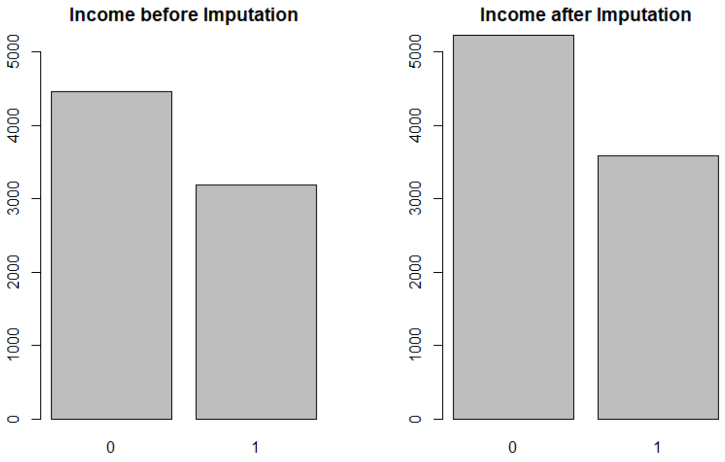


Figure 4: Income before and after imputation

Since one of the goals is to predict the outcome of CKD for those participants where it is unknown, all variables used for predictive modeling must be imputed for that “CKD unknown” subset to get probability values. However, imputing methods for one subset should be equally applied to the other for balance. Where bias cannot be avoided, the same bias should be applied to both subsets. For this reason, it was chosen that all variables in the entire dataset would be imputed up front. Imputation was performed using the mice package which applied logistic regression to predict binary variables and predictive mean matching for continuous variables and non-binary categorical variables. This multiple imputation method unfortunately assumes the data is missing at random.

## Part 1: Developing the Best Predictive Model

Logistic regression was used exclusively to develop the best predictive model. Two feature selection algorithms, Boruta and Recursive Feature Elimination (RFE), were then independently applied to the data, and variables deemed important were used as inputs to logistic regression. A logistic regression with no feature selection served as a baseline against each feature selection algorithm. Feature selection using bagging and random forest were also considered but not implemented (see Appendix). The 6000 mice-imputed observations with a known CKD outcome were used in building the model and evaluating performance.

Boruta feature selection compares importance between a randomized version of the dataset and one trained by random forest classification. Variables found to be significant are logged as a hit, and the cycle repeats for a set number of iterations. Boruta applies a wrapper method of selection, meaning it can remove variables that consistently fail to result in a hit. The most important variables have the highest number of hits after all iterations are complete. Boruta feature selection reduced the dimensionality to 20 variables with Age, SBP, CHF, CVD, and Hypertension rounding out the top five (Figure 5).

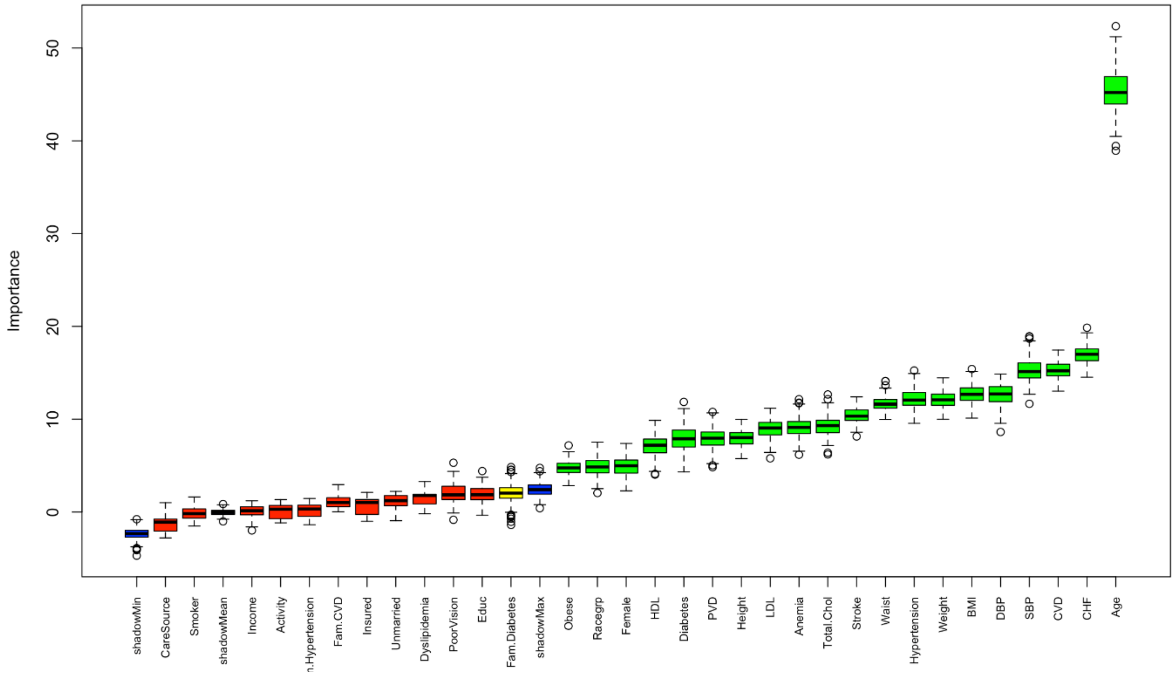


Figure 5: Importance of variables using Boruta Feature Selection

Recursive feature elimination (RFE) applies a machine learning model on a subset of the data to estimate feature importance. It then uses a type of backward selection (also a wrapper method) and cross validation to gradually eliminate unimportant variables and re-rank predictors with each new iteration. Taking the highest accuracy RFE reduced the dimensionality to 25 variables with Age, CHF, CVD, SBP, and Hypertension in the top five in importance.

The model is scored by rewarding $1300 for each true positive and deducting $100 for each false positive. The logistic regression assigns a probability of CKD to each observation and the threshold for determining which ones have CKD and which ones don't is entirely based on maximizing true positives and minimizing false positives to maximize return. The baseline model with no feature selection had the most true positives (see Table 1) but also had a significant number of false positives bringing the monetary return down to $409,200 with an optimal threshold at 6.7%. Boruta performed slightly better returning $410,600 at a 7.8% threshold and RFE slightly worse with $409,000 at 7.4%. Logistic regression using Boruta feature selection was chosen for predicting the 2819 observations with an unknown outcome for CKD.

| | No Feature Selection | Boruta | Recursive Feature Elimination |
|---|---|---|---|
| AIC | 2233.6 | 2221.6 | 2223.4 |
| Null Deviance - Deviance | 1108.9 | 1088.8 | 1101.1 |
| Null °Free - Residual °Free | 37 | 21 | 28 |
| Log Likelihood | 2157.6 | 2177.6 | 2165.4 |
| Area Under the Curve | 90.27% | 90.07% | 90.21% |

| | | | |
|---|---|---|---|
| True Positive | 411 | 405 | 406 |
| False Positive | 1251 | 1159 | 1188 |
| True Negative | 4285 | 4377 | 4348 |
| False Negative | 53 | 59 | 58 |
| Accuracy | 78.3% | 79.7% | 79.2% |
| Precision | 24.7% | 25.9% | 25.5% |
| Recall | 88.6% | 87.3% | 87.5% |
| True Positive Rate | 88.6% | 87.3% | 87.5% |
| False Positive Rate | 22.6% | 20.9% | 21.5% |
| F-Measure | 38.7% | 39.9% | 39.5% |

Table 1: Logistic regression performance based on feature selection methods

## Part 2: Developing an Efficient but Easy-to-Use Screening Tool

When building a survey for future patients to help determine if they are at risk for CKD, we need to strike a balance between the simplicity of our survey and the accuracy of our results. By reducing the number and complexity of questions in our survey, patients are more likely to answer everything, which will lead to having much less NA values, hence, making us less reliant on imputation. However, by having fewer questions, the accuracy of our model may decrease. Our survey will not include questions that generally relate to statistics that patients need special equipment or assistance to calculate, such as details regarding their blood pressure or cholesterol. The main objective of this survey is to enable patients to answer the questions without any assistance from a medical professional and determine their need for getting tested for CKD.

The first model that we used to find the best solution for the survey consisted of the following 14 variables: Waist, history of Hypertension in the family, both history of Diabetes in the family and Diabetes, Cognitive Heart Failure (CHF), Body Mass Index (BMI), both history of CVD in the family and CVD, Gender, Peripheral Vascular Disease (PVD), Race Group, Weight, Anemia, and Age. These variables were determined based on the results from a backwards-selection logistic regression analysis, which identified 19 variables as significant. We analyzed those variables and removed any that we determined were unrelated to CKD, such as education, and any that would likely not be answered in a survey due to lack of knowledge, such as questions related to blood pressure statistics.

The model was run using two different thresholds values. The reason behind this was to see our values with two different goals. The first goal was to maximize the payout and the second was to find a balance between a high payout and a low number of false negative identifications. Falsely identifying a patient as not at risk for CKD would be detrimental to the patient's health and is a $1300 opportunity cost.

| Goal | Payout | AUC | TP | FP | FN | Threshold |
|---|---|---|---|---|---|---|
| High Payout, Low FN | $395,300 | 89.56% | 402 | 1273 | 62 | 0.071 |
| Highest Payout | $399,100 | 89.56% | 382 | 975 | 82 | 0.1 |

Table 2: Payout Calculations for Backwards-Selection Survey Model

As shown in the table above, the second model had the highest payout value however the False Negative (FN) value increased by 20 points. This means that our high-payout model misdiagnosed the patients who had CKD.

Even though the payouts for both models are high, it is still not considered an ideal model because it was produced using the backward selection method. Backwards Selection is biased as the variables chosen are solely dependent on their level of significance. The model starts by choosing all variables and then starts eliminating the variables with p-values higher than 0.5 in each iteration, leaving you with the model consisting of only the significant variables. The model also does not do a great job with categorical variables. When we initially ran the model with backwards selection, it considered the entire RaceGrp variable as significant even when only one of three categories had a p-value less than 0.5. The fourth category was included in the intercept of the model. Moreover, the model in the end also included 'Education' as a variable. This would not be the ideal model as the level of education a participant has is associated with them having CKD.

With this in mind, we designed our second model for the survey using a mix of variables deemed significant in a decision tree analysis and variables that the original case study found were significant.  This led us to create a model that determined whether a participant should be tested for CKD based on the Age, Weight, Waist, Height, Diabetes, CVD, and Female variables.  As before, certain variables that the decision tree found to be significant, such as blood pressure statistics, were removed from our model so that participants who do not have this information readily available can still use our survey.  We included the Female variable as it is a simple binary question that provided a boost to both our payout and Area Under the Curve.  Research also suggests that one's gender determines which factors contribute to CKD, which makes gender a worthwhile addition to our model (Chang).

Using this model, we set out to maximize our payout while keeping our number of false negatives as low as possible.  While the highest payout we identified with this model is a bit lower than the highest payout with the model based on backwards-selection, our number of false negatives is slightly lower.

| Goal | Payout | AUC | TP | FP | FN | Threshold |
|---|---|---|---|---|---|---|
| High Payout, Low FN | $389,300 | 88.87% | 403 | 1346 | 61 | 0.072 |

Table 3: Payout Calculation for Mixed Survey Model

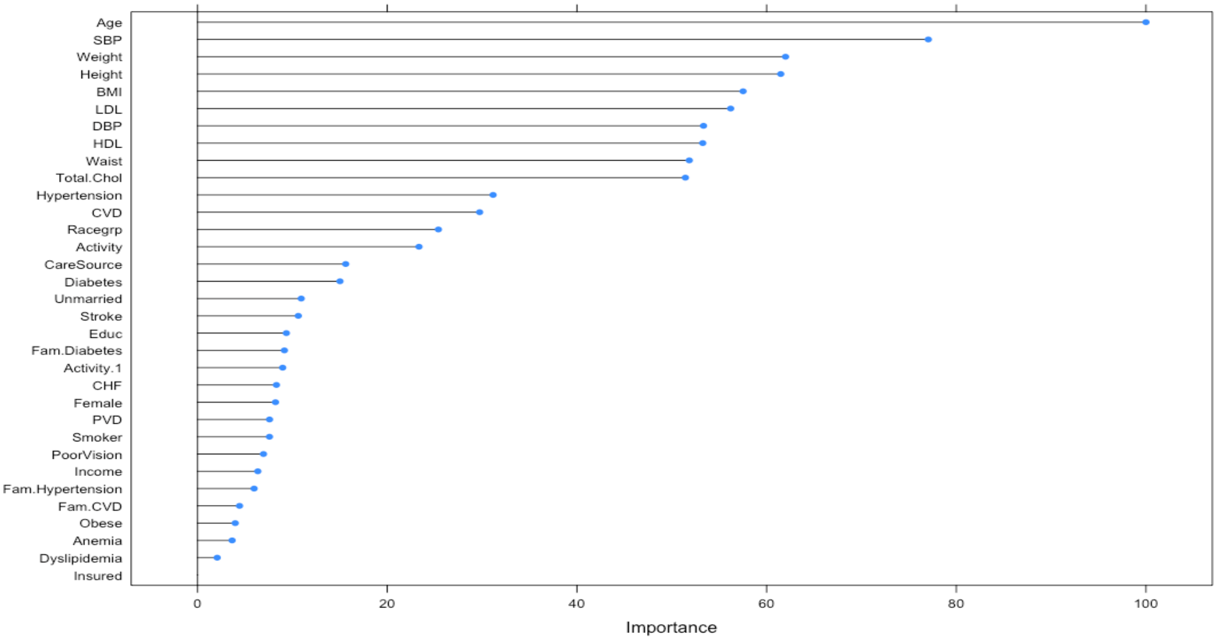
Our survey generates the below coefficients for each variable.  Using this formula (Score = -16.559 + 0.106*Age + …), we can assume that any participant whose total value is greater than **0.072** should be tested for CKD.  The scores assigned to each participant in the study can be observed in the third column of the .csv file included with this report.  The original intercept value was –14.415, but due to a necessary adjustment that required us to convert the variables of Diabetes, CVD, and Female to numeric values, the binary values were changed from 0 = "No" and 1 = "Yes" to 1 = "No" and 2 = "Yes".  By making an adjustment to the intercept, we can maintain our original threshold without seeing any change in our final values.

| Intercept | Age | Weight | Waist | Height | Diabetes | CVD | Female |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -16.559 | 0.106 | 0.013 | -0.009 | 0.028 | 0.655 | 0.802 | 0.687 |

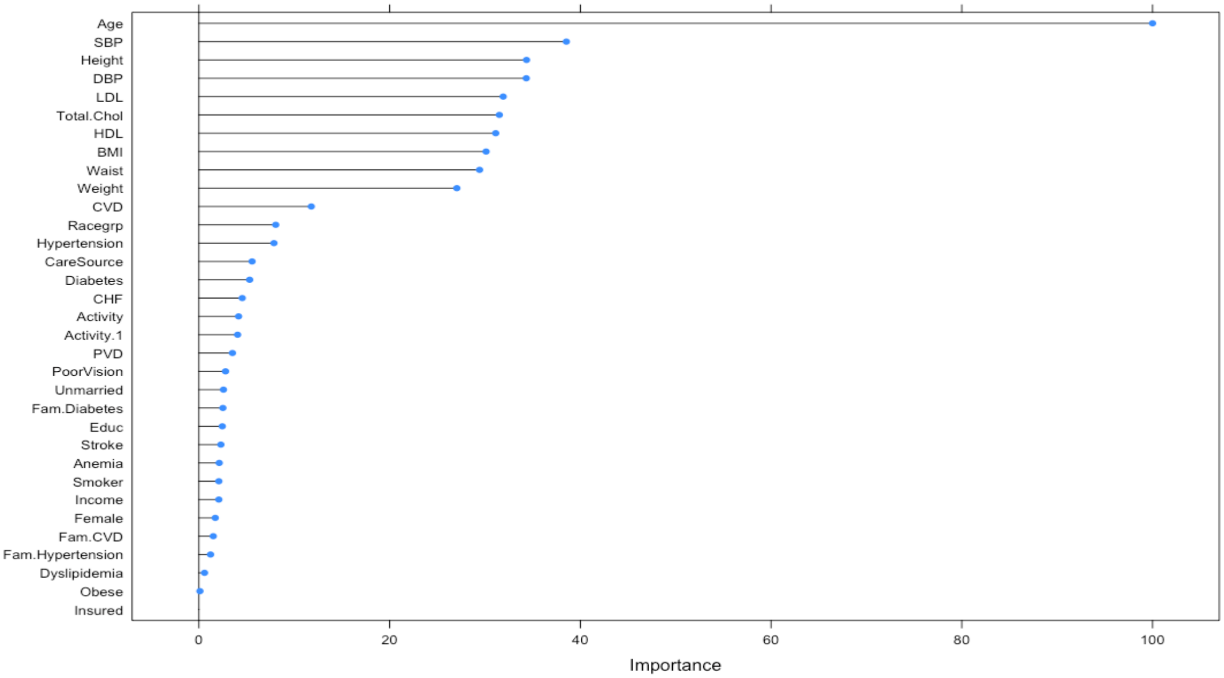
Table 4: Coefficients for Survey Score Equation

## Appendix



| Age | Weight | Waist | Height | Diabetes | CVD | Female |
|---|---|---|---|---|---|---|
| 100 | 61.995 | 51.847 | 61.492 | 15.015 | 29.741 | 8.21513 |

Figure 6: Importance of variables using Bagging model for Feature Selection



| Age | Weight | Waist | Height | Diabetes | CVD | Female |
|---|---|---|---|---|---|---|
| 100 | 27.048 | 29.433 | 34.357 | 5.324 | 11.788 | 1.719 |

Figure 7: Importance of variables using Random Forest for Feature Selection

## References

Chang, Po-Ya. Chien, Li-Nien. Lin, Yuh-Feng. Wu, Mai-Szu. Chiu, Wen-Ta. Chou, Hung-Yi. “Risk factors of gender for renal progression in patients with early chronic kidney disease.” Medicine (Baltimore). July 29, 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5265827/#:~:text=CKD%20progression%20may%20differ%20depending%20on%20sex.,-%5B4%2C5%5D&text=Male%20patients%20show%20a%20substantially,those%20observed%20in%20female%20patients.