Screening for Chronic Kidney Disease

Introduction

Symptoms: Loss of appetite, Hypertension, Fatigue

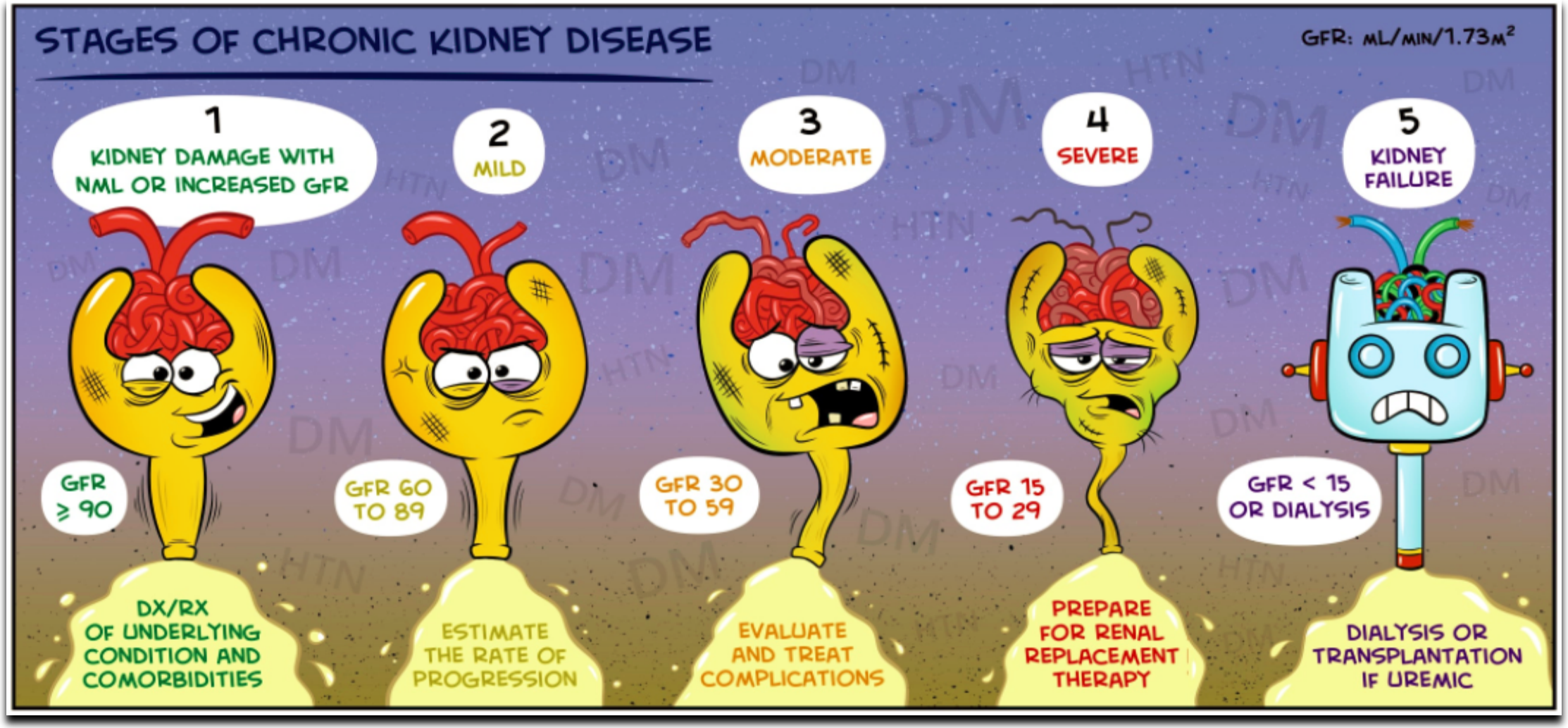
Fact: More than 200,000 US cases per year

Severity: Incurable, Kidney transplant

Diagnosis: Long term care required

Risk factor: Cancer, CVD

Disease Progression

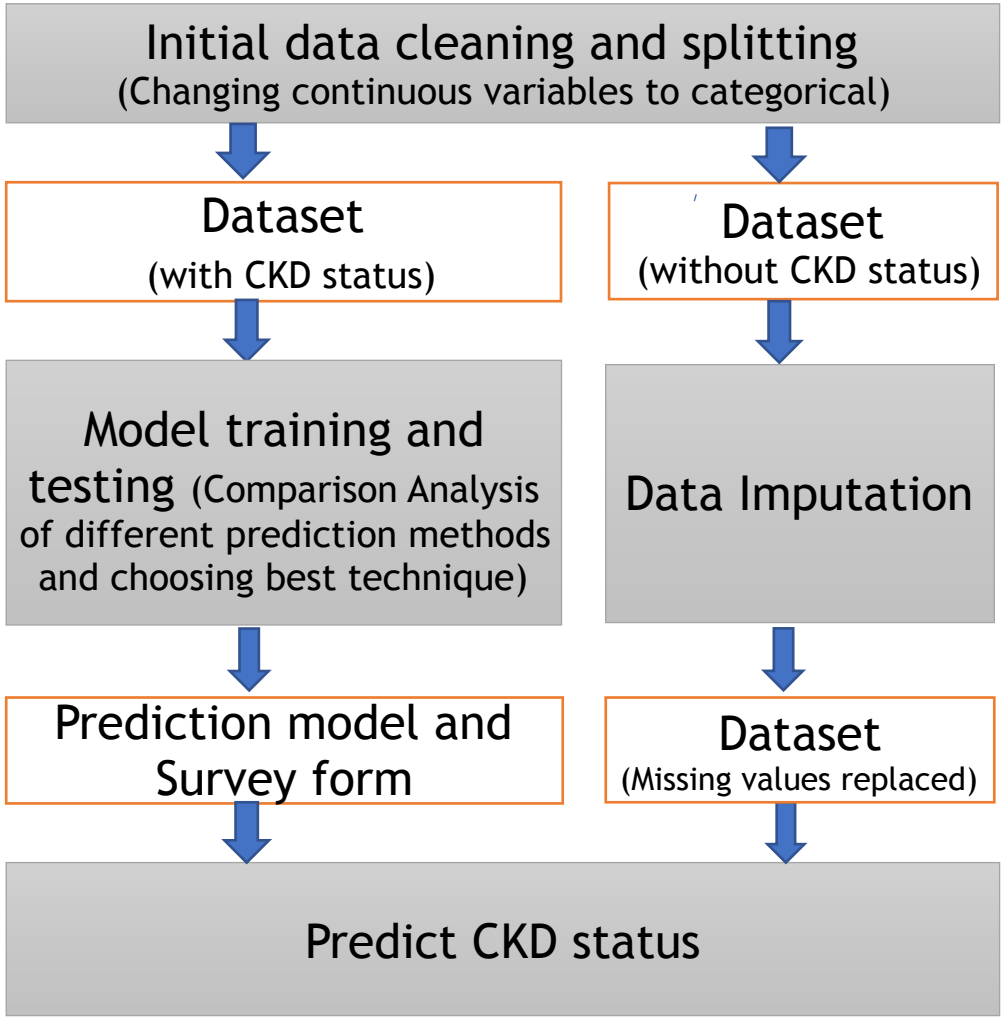


1 in 7

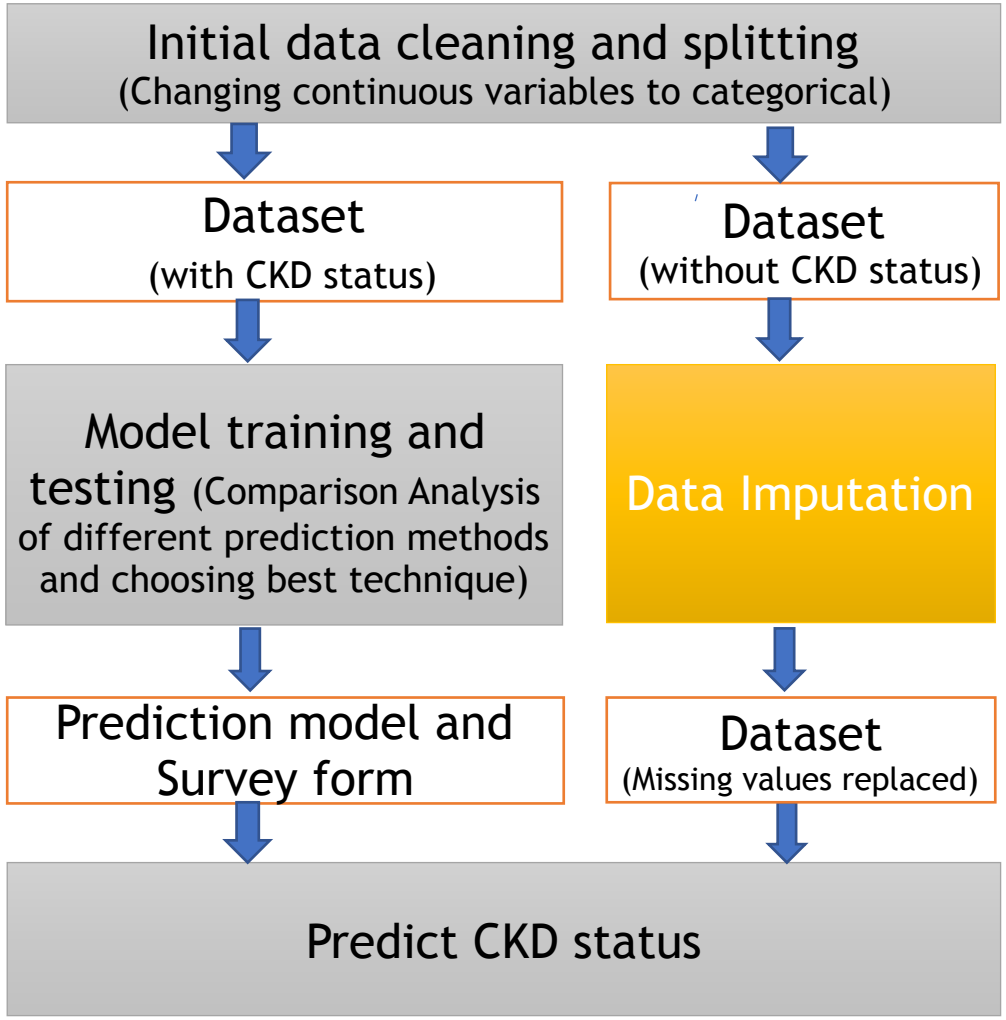
American adults

have CKD

Process Overview



Process Overview



Data Imputation

Case Deletion

- Sample size reduced to great extent
- Used when there is no structure or pattern to the missing data

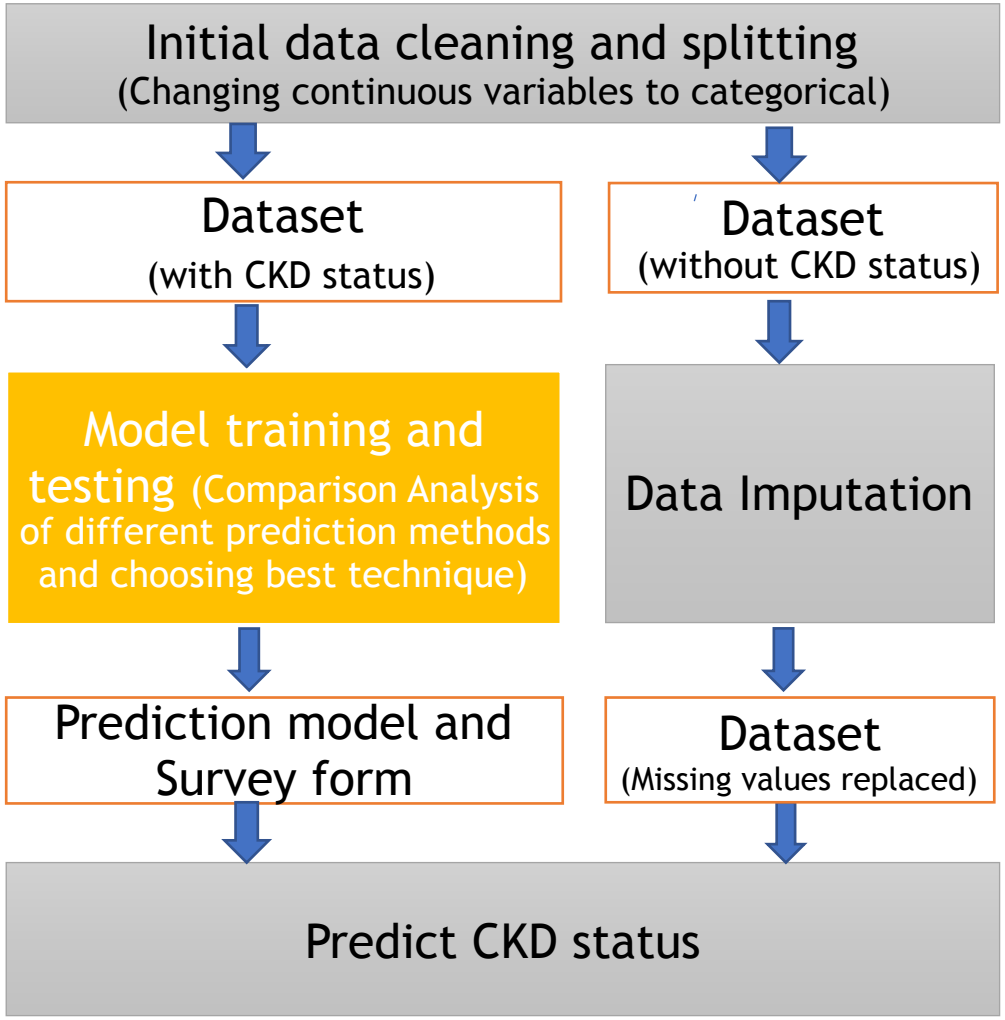
KNN Imputation (Median)

- Predicts both categorical and continuous variables
- It takes in consideration the correlation structure of the data.

Mean Imputation

- Prediction varies depending on outliers
- Single value imputation deflates the variance

Process Overview

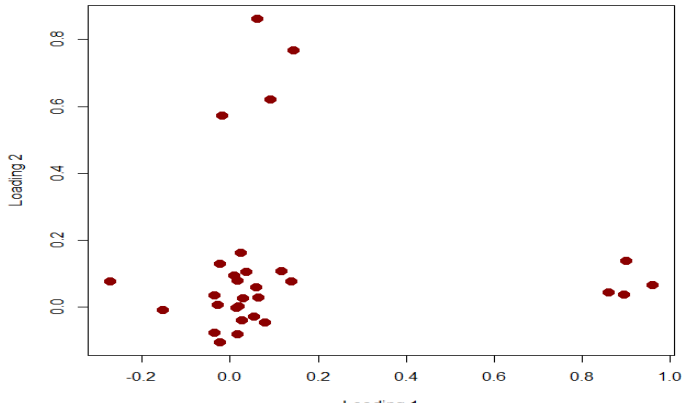


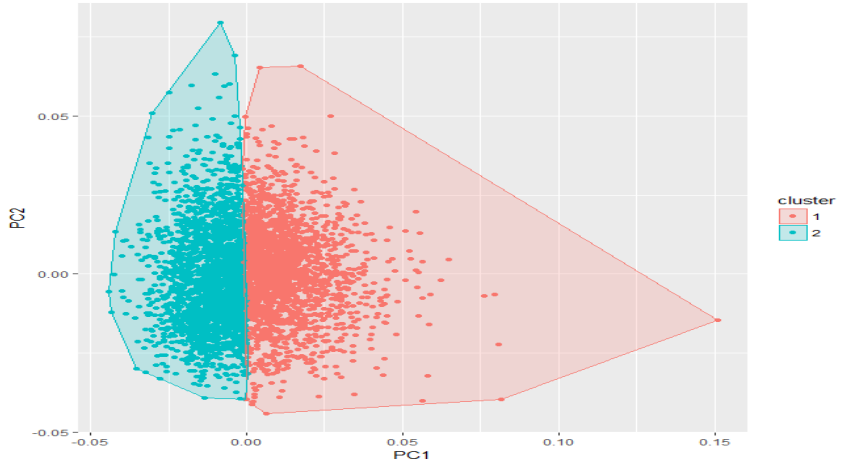
Model Training and testing

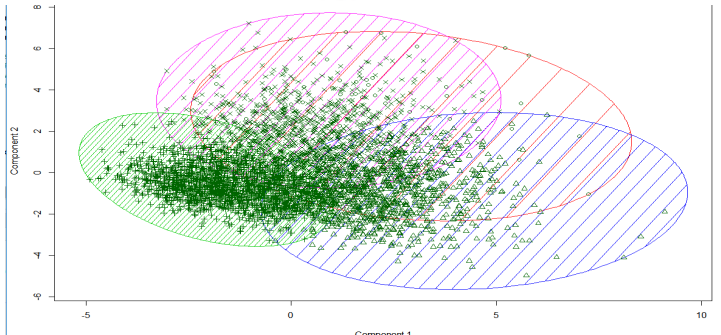
Factor Analysis
- Helps in finding structure and pattern in data
- Interpretation of the factor is subjective
- Choosing number of factors is difficult

Principal Components Analysis
- Easy data visualization
- Dominant variables in data
- Information loss

Clustering
- Choosing number of clusters is difficult
- Not able to perform hierarchical clustering on large data
- Non-separable clusters due to many outliers







We decide to use logistic regression methodology

- It determines an outcome which can be success or failure based on a set of predictor variables
- The relationship between the dependent and independent variable is not linear rather follows an S-shaped (or sigmoidal) curve.
- When true probabilities are extreme ,the linear model can predict values which are greater than 1 or less than 0 –Achilles heel

Transform the continuous numeric variables to categorical variables

Age

Continuous Numeric	Categorical
0-10	Level-One
11-20	Level-Two
21-30	Level-Three
31-40	Level-Four
41-50	Level-Five
51-60	Level-Six
61-70	Level-Seven
71-80	Level-Eight
81-90	Level-Nine
>90	Level-Ten

Weight

Continuous Numeric	Categorical
0-20	Level-One
21-40	Level-Two
41-60	Level-Three
61-80	Level-Four
81-100	Level-Five
101-120	Level-Six
121-140	Level-Seven
141-160	Level-Eight
161-180	Level-Nine
>180	Level-Ten

Height

Continuous Numeric	Categorical
0-110	Level-One
111-120	Level-Two
121-130	Level-Three
131-140	Level-Four
141-150	Level-Five
151-160	Level-Six
161-170	Level-Seven
171-180	Level-Eight
181-190	Level-Nine
>190	Level-Ten

BMI

Continuous Numeric	Categorical
<18.5	Underweight
18.5-24.9	Healthy
25.0-29.9	Overweight
>29.9	Obese

...

LDL

Continuous Numeric	Categorical
0-50	Level-One
51-100	Level-Two
101-125	Level-Three
125-150	Level-Four
151-175	Level-Five
176-200	Level-Six
201-300	Level-Seven
301-500	Level-Eight
>500	Level-Nine

Figure out logistic regression model and then a simple screening tool

Logistic Regression Model

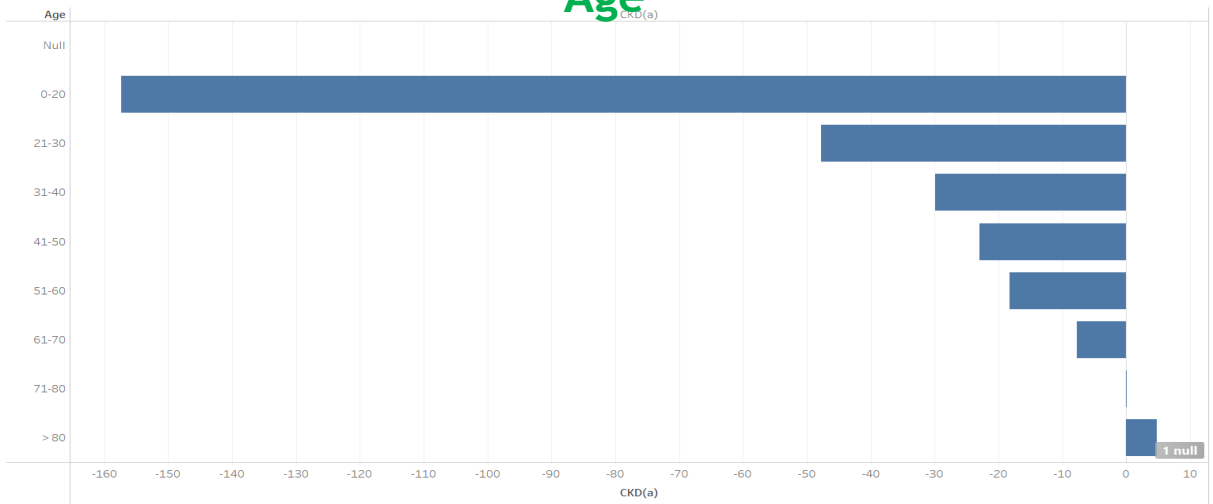
```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.19821    0.56755  -5.635 1.75e-08 ***
Female           0.30798    0.16452   1.872 0.061213 .
Racegrphispa    -0.94819    0.26962  -3.517 0.000437 ***
Racegrpother    -0.20882    0.56752  -0.368 0.712905
Racegrpwhite     0.12621    0.21018   0.601 0.548172
Unmarried        0.27953    0.15939   1.754 0.079483 .
PVD              0.38130    0.22885   1.666 0.095680 .
Hypertension     0.72034    0.17718   4.066 4.79e-05 ***
Diabetes         0.60920    0.17389   3.503 0.000459 ***
CVD              0.82417    0.19122   4.310 1.63e-05 ***
Anemia           1.07386    0.50079   2.144 0.032006 *
Age_LevelFive   -2.29269    0.31920  -7.183 6.84e-13 ***
Age_LevelFour   -2.99336    0.44207  -6.771 1.28e-11 ***
Age_LevelNine    0.47569    0.20472   2.324 0.020145 *
Age_LevelSeven  -0.76862    0.19084  -4.028 5.63e-05 ***
Age_LevelSix    -1.82197    0.26676  -6.830 8.50e-12 ***
Age_LevelThree  -4.77085    1.01727  -4.690 2.73e-06 ***
Age_LevelTwo   -15.72912  427.68290  -0.037 0.970662
HDL_LevelFive    0.67153    0.51523   1.303 0.192452
HDL_LevelFour    1.16447    0.50846   2.290 0.022009 *
HDL_LevelNine    0.61372    0.80241   0.765 0.444367
HDL_LevelOne   -13.23142 1546.48434  -0.009 0.993174
HDL_LevelSeven   0.51387    0.57224   0.898 0.369190
HDL_LevelSix     0.49190    0.53473   0.920 0.357622
HDL_LevelTen    -0.06845    0.92728  -0.074 0.941153
HDL_LevelThree   1.16326    0.51871   2.243 0.024923 *
HDL_LevelTwo     1.25516    0.60925   2.060 0.039382 *
```

Simple Screening Tool

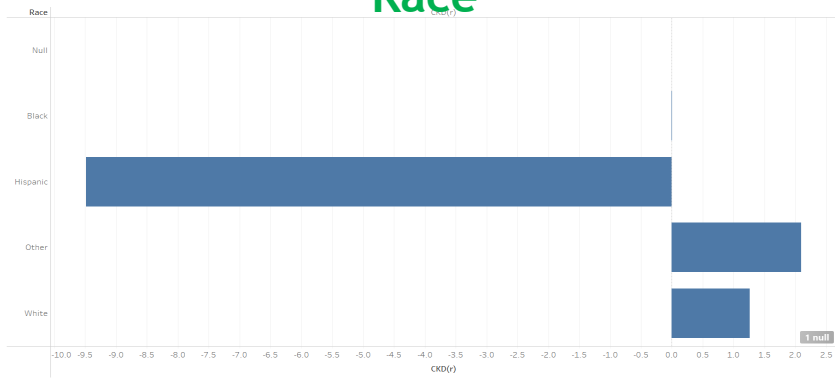
1. What are your gender? (If **Female**, 3.08 points; **Male**, 0 points)
   Female ☐ Male ☐
2. What is your age (years)? (if **0-20**, -157.29 points; **21-30**, -47.71 points; **31-40**, -29.93 points; **41-50**, -22.93 points; **51-60**, -18.22 points; **61-70**, -7.69 points; **71-80**, 0 points; **>80**, 4.76 points)
   0-20 ☐ 21-30 ☐ 31-40 ☐ 41-50 ☐ 51-60 ☐ 61-70 ☐ 71-80 ☐ >80 ☐
3. What is your race? (if **White**, 1.26 points; **Black**, 0 points; **Hispanic**, -9.48 points; **Other**, -2.09 points)
   White ☐ Black ☐ Hispanic ☐ Other ☐
4. Are you unmarried? (If **Yes**, 2.80 points; **No**, 0 points)
   Yes ☐ No ☐
5. Do you have PVD? (If **Yes**, 3.81 points; **No**, 0 points)
   Yes ☐ No ☐
6. Do you have Hypertension? (if **Yes**, 7.20 points; **No**, 0 points)
   Yes ☐ No ☐
7. Do you have Diabetes? (If Yes, 6.09 points; No, 0 points)
   Yes ☐ No ☐
8. Has a doctor ever told you that you had angina pectoris, myocardial infarction, or stroke? (if Yes, 8.24 points; No, 0 points)
   Yes ☐ No ☐
9. Have you received treatment for anemia in past three months or hemoglobin at exam lower than 11g/dL? (If **Yes**, 10.74 points; **No**, 0 points)
   Yes ☐ No ☐
10. What is your HDL level (mg/dL)? (if **0-20**, -132.31 points; **21-30**, 12.55 points; **31-40**, 11.63 points; **41-50**, 11.65 points; **51-60**, 6.72 points; **61-70**, 4.92 points; **71-80**, 5.14 points; **81-90**, 5.64 points; **91-100**, 6.14 points; **>100**, -0.68 points)
    0-20 ☐ 21-30 ☐ 31-40 ☐ 41-50 ☐ 51-60 ☐
    61-70 ☐ 71-80 ☐ 81-90 ☐ 91-100 ☐ >100 ☐

Here are the significant variables and their corresponding points

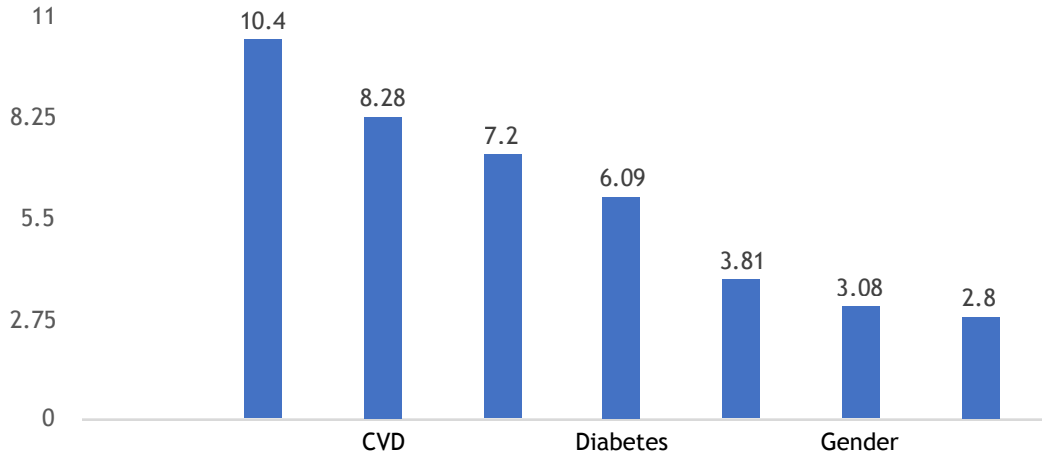
Age



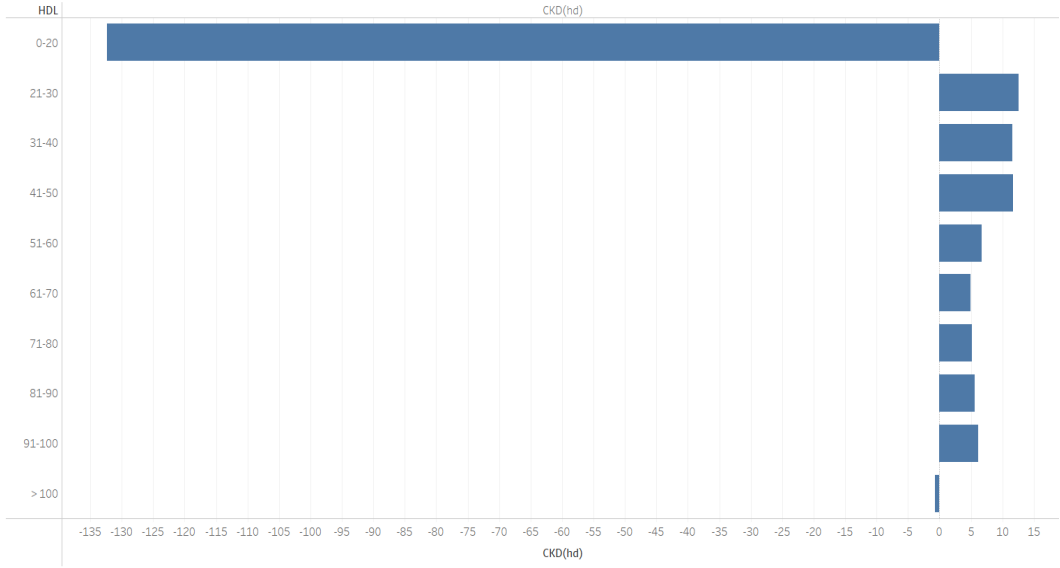
Race



Other variables



HDL



Conclusion

"An 80-year white unmarried female with PVD, Hypertension, CVD, Anemia and a very low HDL (ranging from 21-50) has very high chance of CKD"

Actual example #3971 from ckddata.csv-

Age	Female	Racegrp	Unmarried	HDL	PVD	Hypertension	Diabetes	CVD	Anemia	CKD
85	1	white	1	36	1	1	1	1	1	1

Interpretation of results

Confusion Matrix	True Condition		
Predicted Condition	Total population	Condition Positive	Condition Negative
	Predicted Condition Positive	True Positive 11	True Negative 1284
	Predicted Condition Negative	False Positive 66	False Negative 18

Accuracy = 94%

F-measure = 0.21

Final Cost = $200 * FN + $100 * FP = $10,200

AIC = 936.68

Limitations

- Identifying independent variables
- Limited outcome variables- cannot predict continuous outcomes
- Overfitting the model
- Assumption regarding the relationship between predictor and dependent variables

Recommendations

- To implement random forest algorithm as it can be used for classification as well as regression
- Need to observe BIC values

References

- https://pdfs.semanticscholar.org/4172/f558219b94f850c6567f93fa60dee7e65139.pdf
- https://www.gstatic.com/healthricherkp/pdf/chronic_kidney_disease.pdf
- https://www.kidney.org/news/one-seven-american-adults-estimated-to-have-chronic-kidney-disease
- https://classroom.synonym.com/disadvantages-logistic-regression-8574447.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543767/
- http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

Thank you