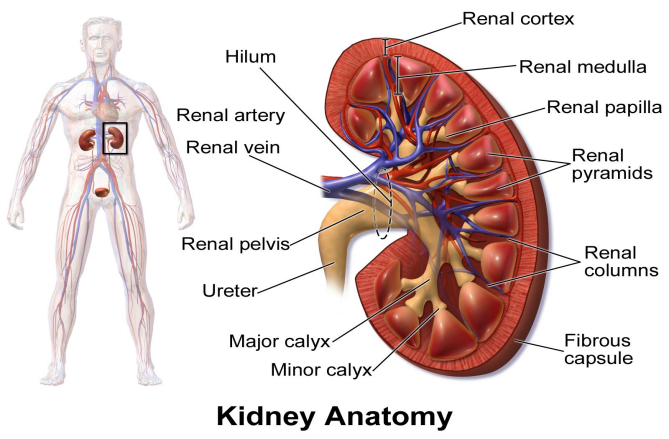


Kidney Disease Prediction

using Classification Models

By: Nick Catanglan

Agenda:

1. Problem Overview
2. Data Wrangling
3. Exploratory Data Analysis
4. Model Selections and Results
5. Future Work & Recommendations

1. PROBLEM OVERVIEW

Chronic Kidney Disease (CKD), is the gradual loss of kidney function leading to kidney failure.

- Kidney filter waste and excess fluid from the blood, which are then excreted as a urine.
- CKD is the 9th leading cause of death in the U.S. in 2016

Medicare spent in 2016 alone:

- **$79 billion** for people with Chronic Kidney Disease.

How does CKD Model Algorithm prediction help?

Machine Learning automatically read and predict thousands of laboratory test and medical records of patients therefore will:

- Save valuable time and money for the government, healthcare industry and patients.
- Early detection, the sooner can get treatment.

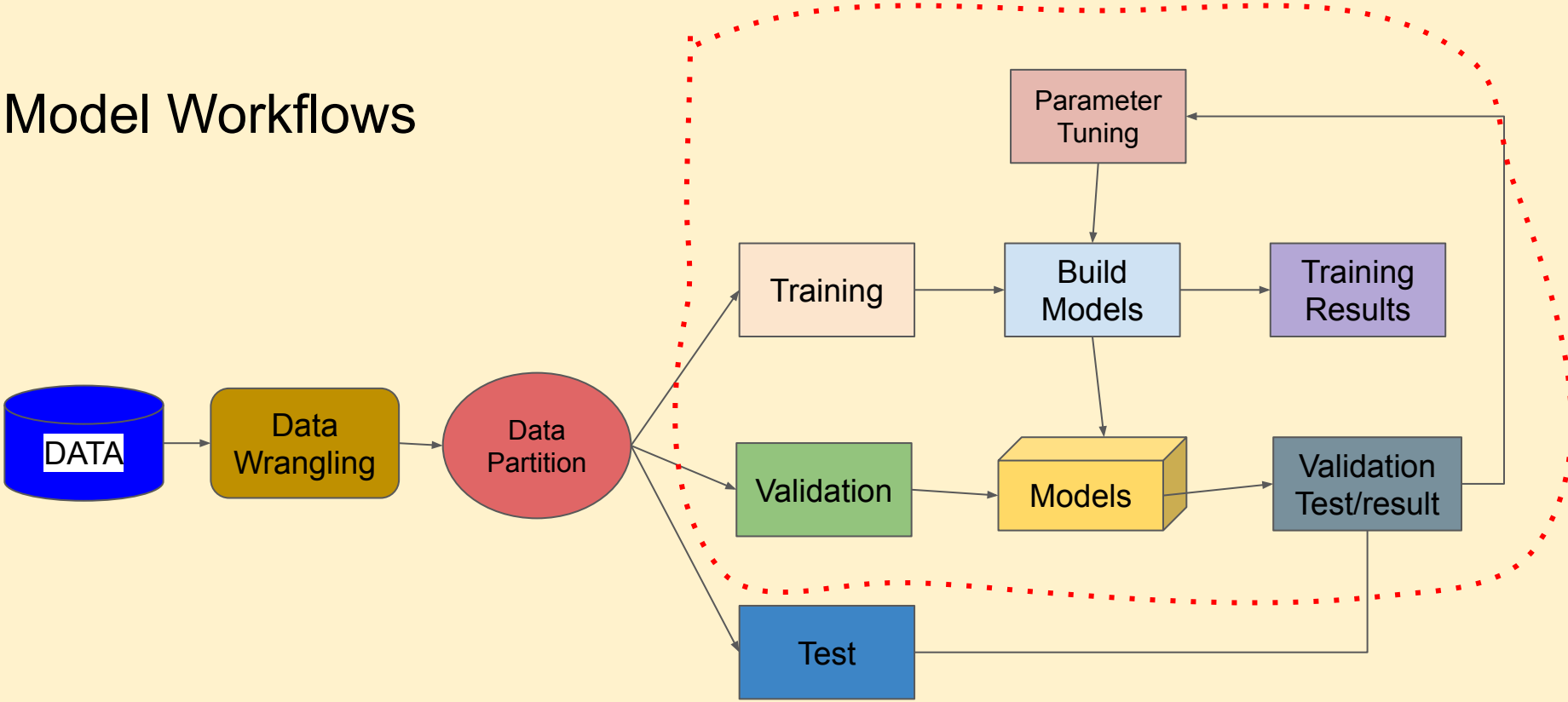
About the dataset :

- Data set from UCI repository
- 400 patients with medical records, laboratory test taken in two months period.
- 26 columns and 400 rows (11 numeric, 14 nominals)

Challenges:

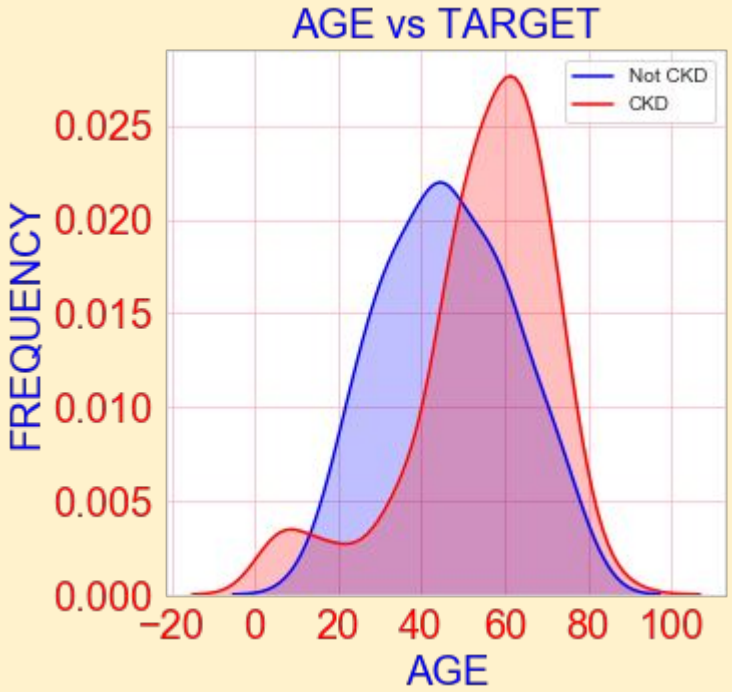
- Missing Values
- Less data points
- Typographical Errors
- Outliers

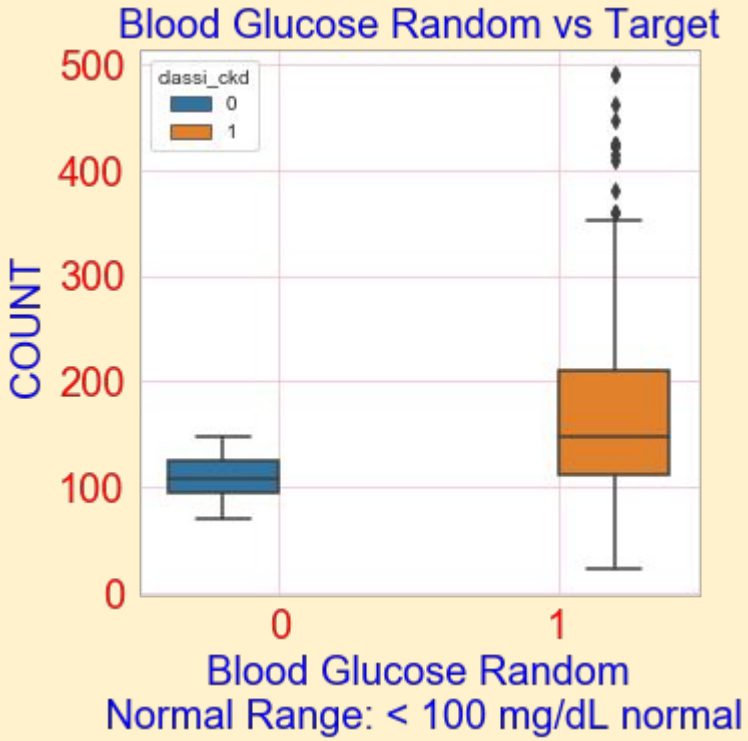
Model Workflows

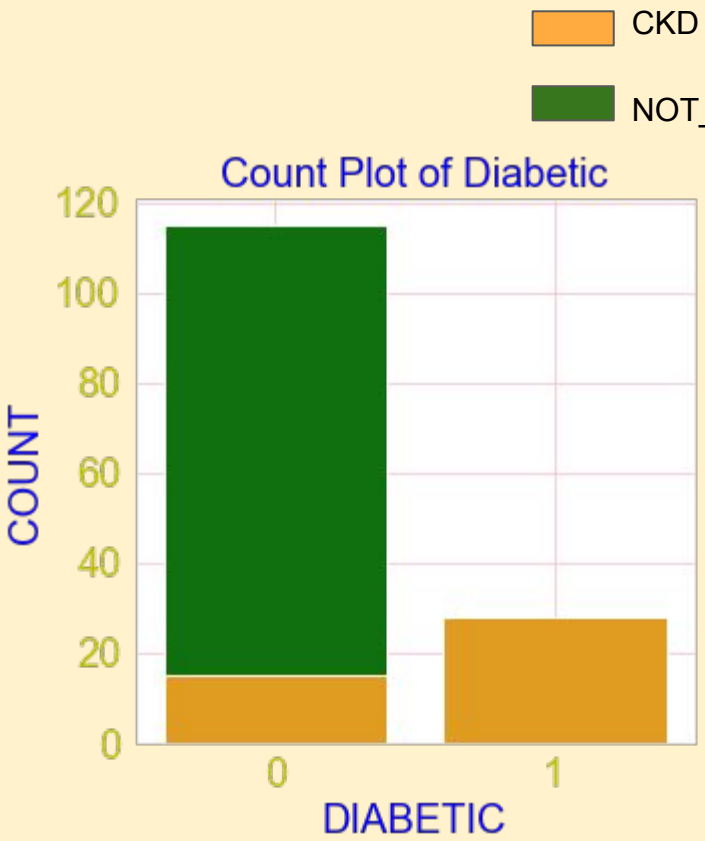


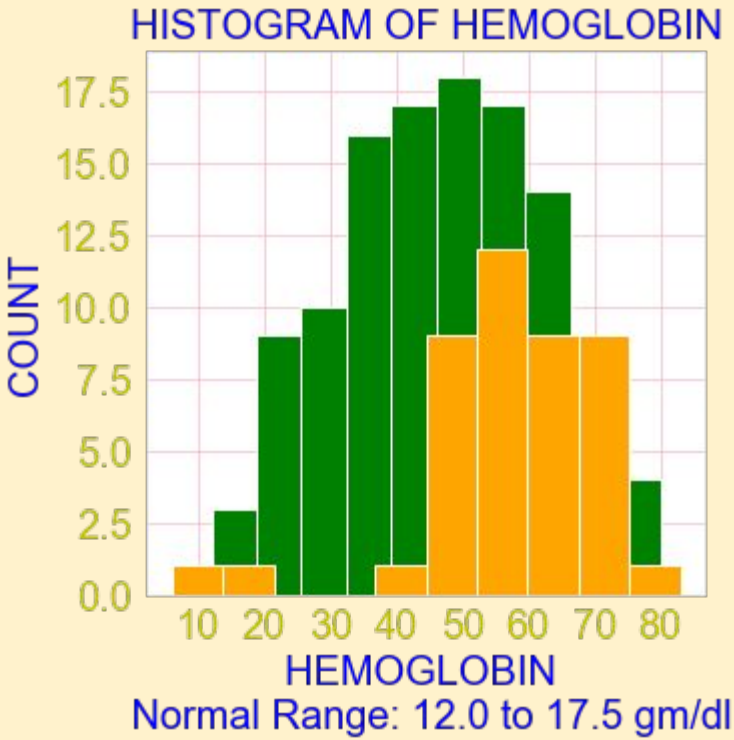
Model Training & Tuning

3. EXPLORATORY ANALYSIS






CKD = 1
NOT_CKD = 0
Count Plot of Diabetic
120
100
80
60
40
20
0
COUNT
0
1
DIABETIC


HISTOGRAM OF HEMOGLOBIN
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
COUNT
10
20
30
40
50
60
70
80
HEMOGLOBIN
Normal Range: 12.0 to 17.5 gm/dl

HEATMAP

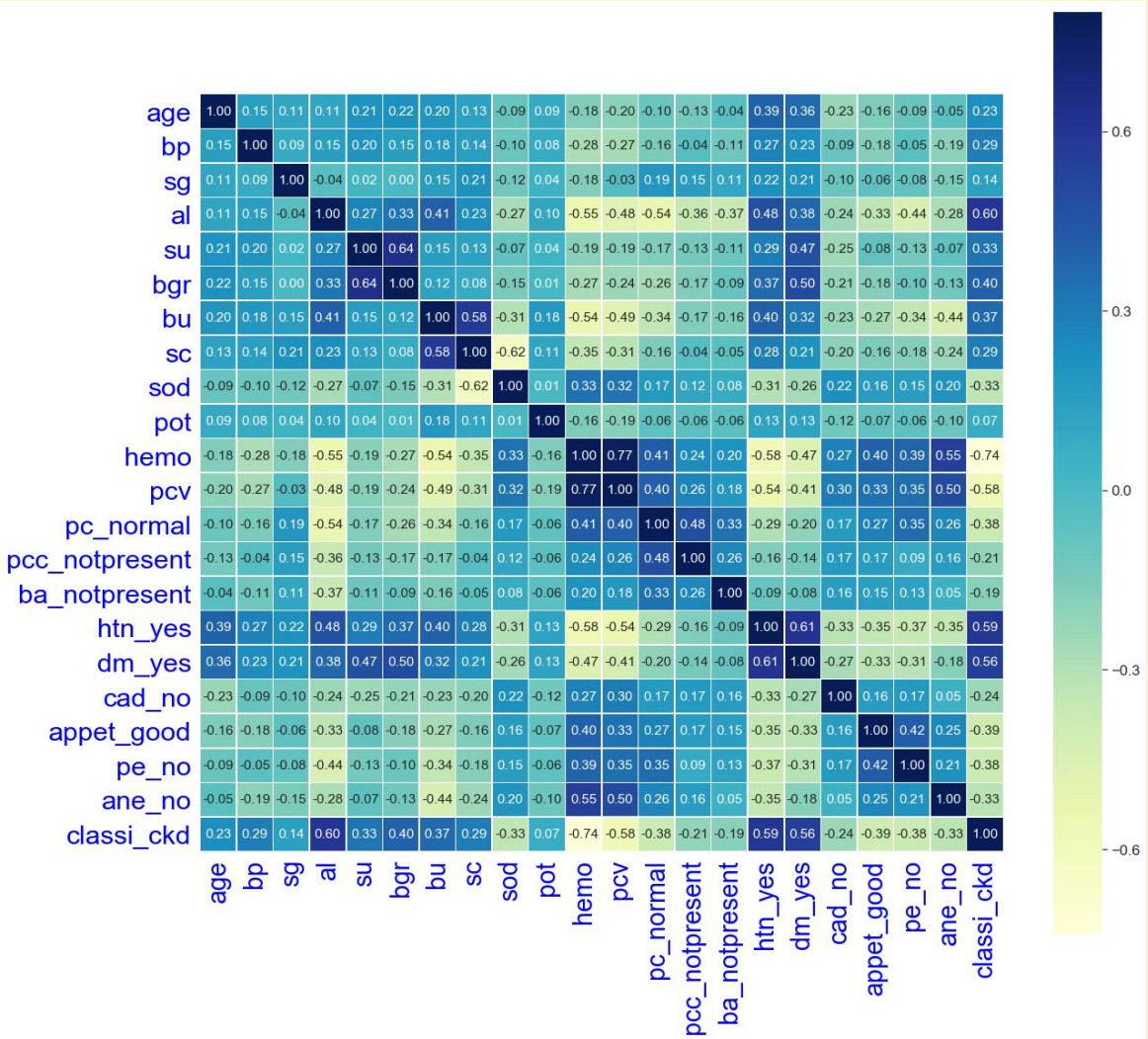
High correlated variables(Features)

- Hemoglobin & PCV (0.8)
- Sugar & BGR
- SC & BU

Very low correlation to Target

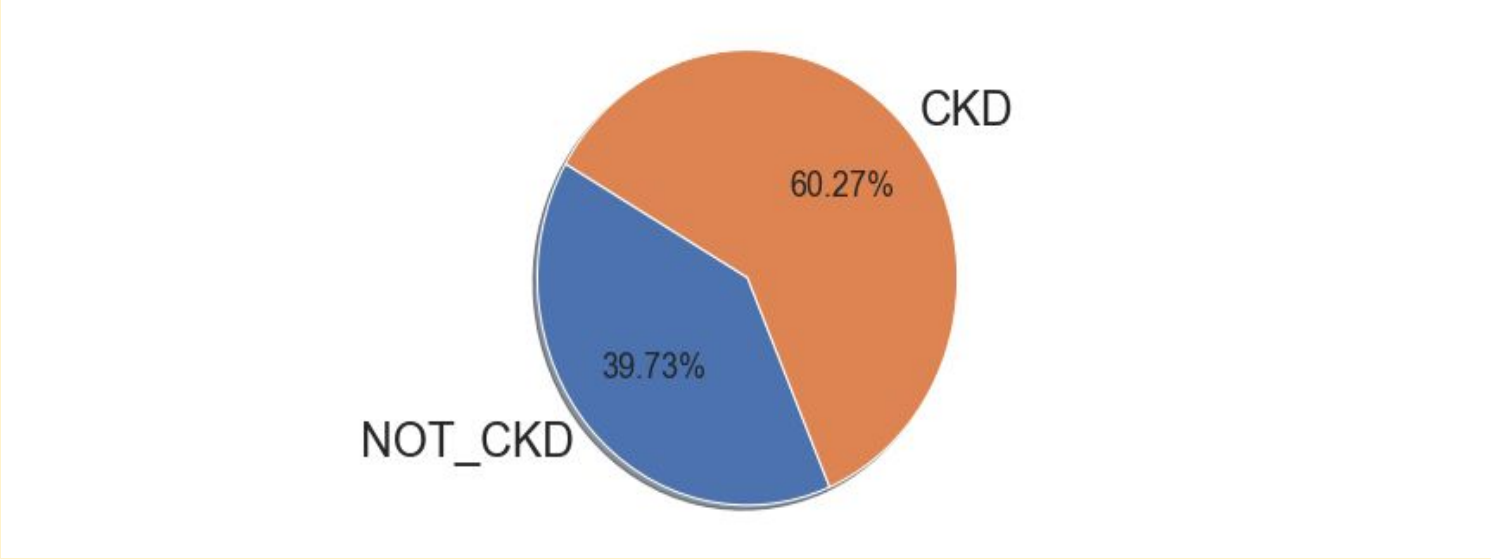
- Potassium

After all cleaning: 368 rows, 16 col.



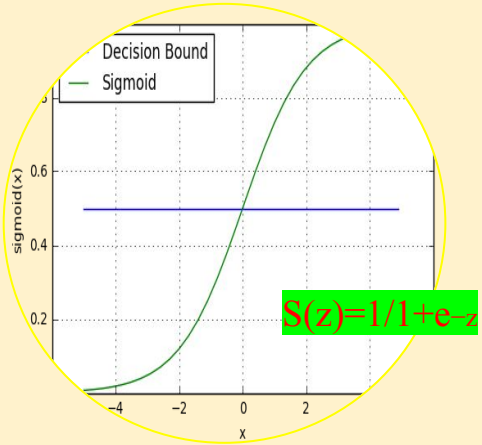
PIE CHART SHOWS THE DIVISION OF OUR TARGET VARIABLE

CKD = 60.27%
NOT_CKD = 39.73%



4. Model Selections

LOGISTIC REGRESSION



- Probability
- Sigmoid function
- 0 to 1 value
- With threshold (0.5)

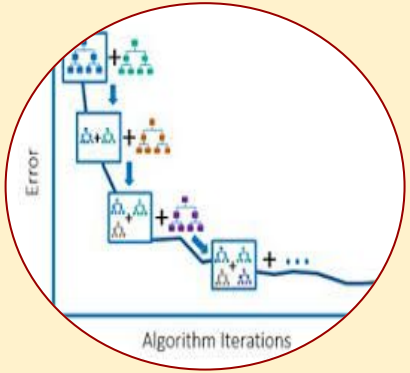
DECISION TREE

- Split data on features
- Repetitive splitting procedure
- Continue repetitive splitting until each node left w/ same class label
- entropy & info gain

RANDOM FOREST

- Ensemble model Bagging(Parallel)
- Compose of many decision trees
- Average performance of trees

GRADIENT BOOSTING



- Ensemble model (Sequential)
- Iterate multiple times.
- Optimizing the loss function(error) of previous learner.

Model Comparison

	Logistic Regression	Decision Tree	Random Forest	KNN	Naive Bayes	SVM	Gradient Boosting
Accuracy	95%	96%	99%	91%	92%	89%	98%
Precision	100%	100%	100%	100%	100%	100%	100%
Recall	96%	94%	98%	89%	87%	89%	97%
Cross_Val	93%	95%	97%	89%	90%	85%	96%

The best performing model is the Random Forest

CONFUSION MATRIX: Random Forest

- Use to measure performance of algorithms

	ACTUAL	ACTUAL
PREDICTED	NOT_CKD	CKD
NOT_CKD	43 => TP	0 => FP T-1 error
CKD	2 =>FN T-2 error	68=>TN

Accuracy = (TP+TN)/(TP+TN+FN+FP) = 98%
98% correct prediction
2% mis-classification error

Precision = TP/(TP+FP) = 100%

Recall = TP/(TP+FN) = 95.6%

SUMMARY:

GOAL:

- Make a model that can predict CKD given some laboratory result and medical history.

RESULTS:

- Model was able to predict with 98% accuracy.
- 2% mis-classification error
- Precision 100%
- Recall 95.6%

RISKS:

- Wrong Diagnosis, model incorrectly classified 2% error as likely as CKD in fact it is NOT CKD.
- Review with group of medical and engineering professions for further study before implementation.

5. FUTURE WORK:

- Collect more data for it has only 400 observation
- Add more important features and laboratory results
  - RBC lab results, Age stratified classifiers, Lifestyle, Work, Married etc.
- Model Improvement: combine multiple classifiers
  - Other Ensemble
- Balance label data:
  - SMOTE
  - Oversampling or other methods.

THANK YOU!!!!!!

Any Question?