**Risk Factor Prediction of Chronic Kidney Disease**

[Risk Factor Prediction of Chronic Kidney Disease - UCI Machine Learning Repository](https://www.archive.ics.uci.edu/dataset/857/risk+factor+prediction+of+chronic+kidney+disease)

Data is [200 rows x 27 columns]

Although it’s a small data but I think the data is of great importance as it can predict Chronic Kidney Disease, therefore, picked my interest. I have been trying to do something great which makes difference, therefore, I was confused and have changed my mind in selecting project. After reviewing data

I noticed that “class” & “affected” are same. I decided to remove “class’’ as “class” was string and “affected” is an integer and hence no data conversion needed

**bp**: Blood Pressure

**sg**: Specific Gravity (of urine)

**al**: Albumin

**rbc**: Red Blood Cells

**su**: Sugar

**pc**: Pus Cell

**pcc**: Pus Cell Clumps

**ba**: Bacteria

**bgr**: Blood Glucose Random

**bu**: Blood Urea

**sc**: Serum Creatinine

**sod**: Sodium

**pot**: Potassium

**hemo**: Hemoglobin

**pcv**: Packed Cell Volume

**wbcc**: White Blood Cell Count

**rbcc**: Red Blood Cell Count

**htn**: Hypertension

**dm**: Diabetes Mellitus

**cad**: Coronary Artery Disease

**appet**: Appetite

**stage**: Stages of CKD

* Chronic Kidney Disease is classified into **5 stages** based on kidney function, often assessed by metrics like the glomerular filtration rate (GFR). For example:
  + Stage 1: Normal kidney function but with signs of kidney damage.
  + Stage 2: Mildly reduced kidney function.
  + Stage 3: Moderately reduced kidney function.
  + Stage 4: Severely reduced kidney function.
  + Stage 5: Kidney failure (end-stage renal disease).
* **GFR (Glomerular Filtration Rate)** is a critical measurement in evaluating kidney function. It estimates how much blood the kidneys filter each minute, helping to determine the health and efficiency of the kidneys.

**Stages of CKD (Based on GFR)**:

* + **Stage 1**: GFR ≥ 90 with kidney damage markers.
  + **Stage 2**: GFR 60-89 (mildly reduced function).
  + **Stage 3**: GFR 30-59 (moderately reduced function).
  + **Stage 4**: GFR 15-29 (severely reduced function).
  + **Stage 5**: GFR < 15 (kidney failure).

**pe**: Pedal Edema

**ane**: Anemia

**grf**: Glomerular Filtration Rate

**age**: Age

-

A screenshot of a computer

Description automatically generated

A blue squares with numbers and a graph

Description automatically generated with medium confidence

**Confusion Matrix:**

* + True Positives (TP): 34 cases were correctly predicted as "Affected."
  + True Negatives (TN): 25 cases were correctly predicted as "Not Affected."
  + False Positives (FP): 0 cases were incorrectly predicted as "Affected" when they were "Not Affected."
  + False Negatives (FN): 1 case was incorrectly predicted as "Not Affected" when it was actually "Affected."
  + The confusion matrix indicates excellent model performance with only one misclassification.

**Classification Report:**

**Not Affected Class:**

* + Precision: 0.96 – Out of all predicted "Not Affected" cases, 96% were correctly classified.
  + Recall: 1.00 – All actual "Not Affected" cases were identified correctly.
  + F1-Score: 0.98 – A balanced metric combining precision and recall.

**Affected Class:**

* + Precision: 1.00 – Out of all predicted "Affected" cases, 100% were correctly classified.
  + Recall: 0.97 – 97% of the actual "Affected" cases were identified correctly.
  + F1-Score: 0.99 – Excellent classification performance for the "Affected" class.
  + Overall Metrics:
  + Accuracy: 0.98 – 98% of all predictions were correct.
  + Macro Average:
  + F1-Score: 0.98 – Average performance across both classes.
  + Weighted Average:
  + F1-Score: 0.98 – Weighted by the number of instances in each class.

**Next step:**

Adjust Class Weights

Analyze Features

Create or derive new features

,…etc.

* Explore Ensemble technique to make it better for generalization, as my fear is that it may not work with unseen data as data is limited and overfitting is a major concern.
* Second 1:1 Capstone Consultation Session on 9th Dec
* Implement Cross-Industry Standard Process for Data Mining (CRISP-DM) framework
* Prepare slides with data for technical audience as well as for non-technical audience
* Any further suggestions to add data into