**Chronic KIdney Disease dataset**

Link Kaggle: <https://www.kaggle.com/mansoordaku/ckdisease>

**# Context**

First, I am new to ML, and just in case I slip up, apologies in advance!!  
So, I am doing an online ML course and this is an assignment where we are supposed to practice scikit-learn's PCA routine. Since the course has been ARCHIVED - which means the discussion posts are not answered!! - hence my posting of the problem here.

What better way to learn than to get so many experts giving me feedback … right?

**# Content**

The data was taken over a 2-month period in India with 25 features ( eg, red blood cell count, white blood cell count, etc). The target is the 'classification', which is either 'ckd' or 'notckd' - **ckd=chronic kidney disease**. There are 400 rows

The data needs cleaning: in that it has NaNs and the numeric features need to be forced to floats. Basically, we were instructed to get rid of ALL ROWS with Nans, with no threshold - meaning, any row that has even one NaN, gets deleted.

Part 1: We are asked to choose 3 features **(bgr, rc, wc)**, visualize them, then run the PCA with n\_components=2.  
the PCA is to be **run twice**: one with **no scaling** and the second run WITH scaling. And this is where my issue starts … **in that after scaling I can hardly see any difference!**

I will stop here for now till I get feedback and then move to Part 2.

# Acknowledgements

The dataset is available at: [https://archive.ics.uci.edu/ml/datasets/ChronicKidneyDisease](https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease)

# Inspiration

I would like to get an intuitive and a practical understanding of PCA.