ACTIVITY 5: PRINCIPAL COMPONENT ANALYSIS

LUSANDA MDHLALOSE
PGDIP DATA SCIENCE

TABLE OF CONTENTS

INTRODUCTION

- I will be presenting the Principal Component Analysis (PCA). results that I derived from my Jupyter Notebook. But before that I will start by presenting the other important areas such as Dataset , Exploratory Data Analysis and Data Preprocessing which will help us understand the data we are dealing with.

DATASET:

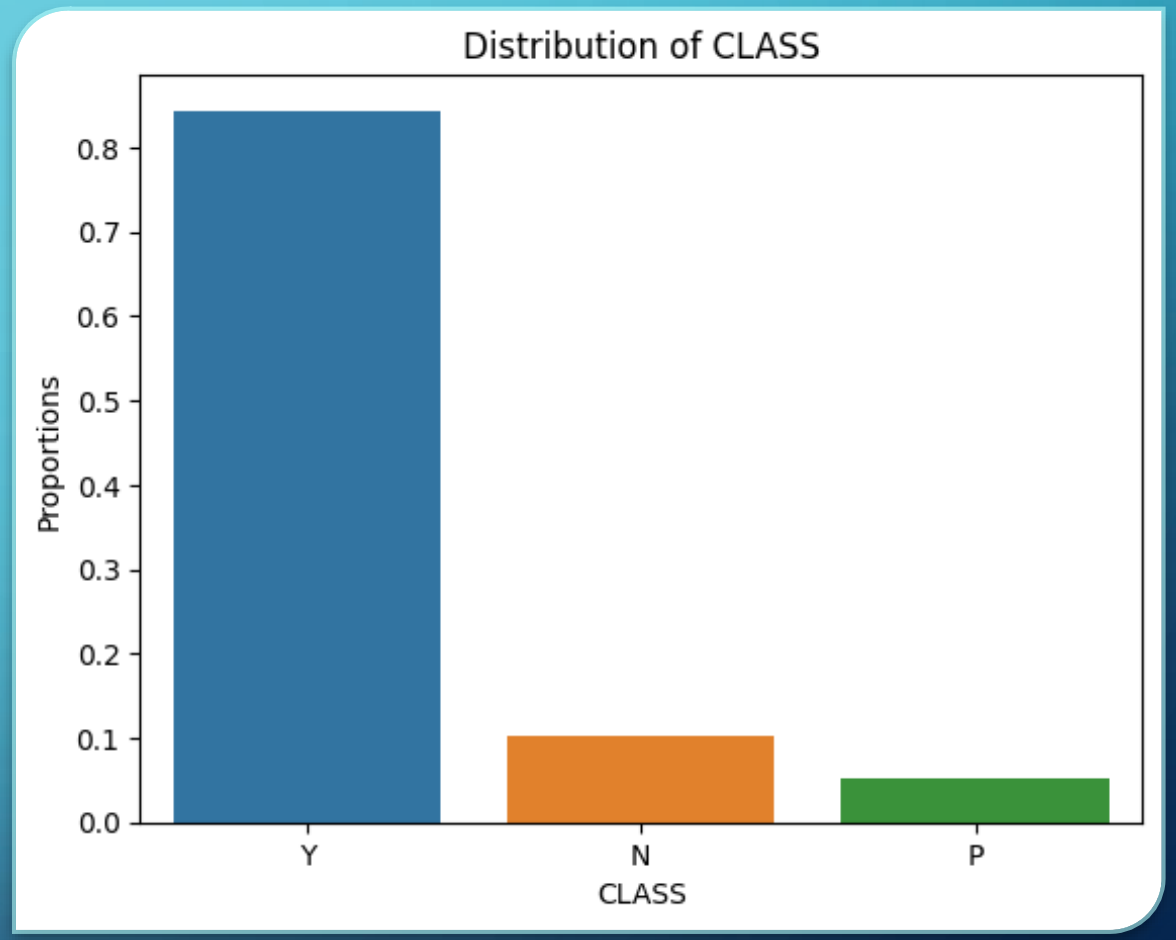
- The dataset used was collected from the laboratory of Medical City Hospital.
- The dataset consists of 1001 Patients with the following recorded attributes; Unique identifier (ID) , No_Pation, Gender, Age, Urea, Creatinine ratio (Cr), Hemoglobin A1c level (HbA1c), Cholesterol (Chol), Triglycerides (TG), HDL Cholesterol, LDL Cholesterol, VLDL Cholesterol, Body Mass Index (BMI) and a Class.
- Under Class, there are three classifications, which are Diabetic(Y), Predict-Diabetic(P) and Not Diabetic (N).

PROCEDURE:

1. I uploaded the dataset and Python libraries
2. Performed a Data Overview to see the basic info and stats of the dataset.
3. Performed an Exploratory Data Analysis before the PCA, to understand the distribution of the data and feature to feature relationship.
4. Performed a Data Preprocessing to prepare the data for PCA analysis.
5. Performed the PCA Algorithm.
6. Provided a visualization of the reduced dimensional space performed by the PCA algorithm.

EXPLORATORY DATA ANALYSIS

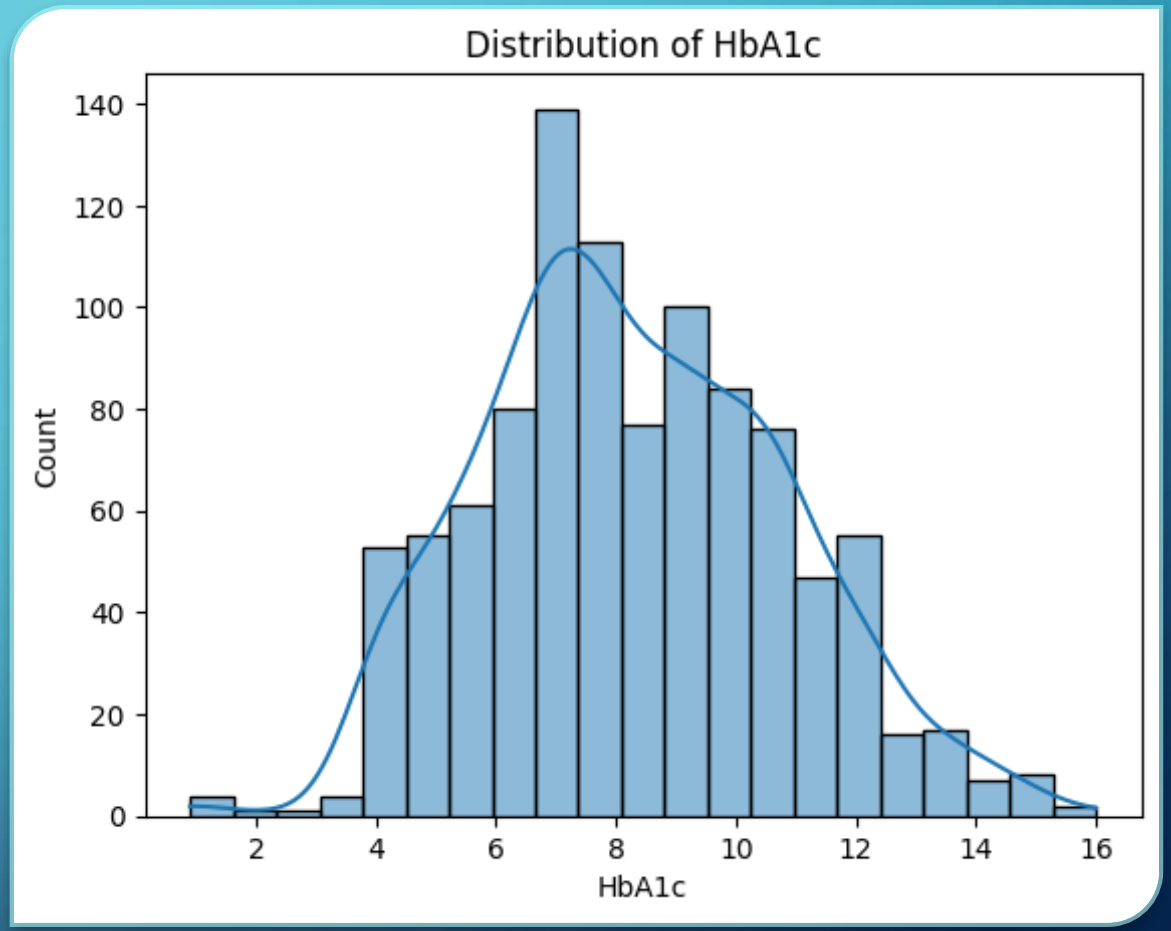
- I will present some of the visualization I performed before PCA analysis:
- Starting with the Classification VS Proportions bar plot.
- We have about 85% of patients within the dataset who are predicted as Diabetic(Y), about 10% as Not Diabetic (N) and about 5% as Predicted Diabetics (P) within our dataset.



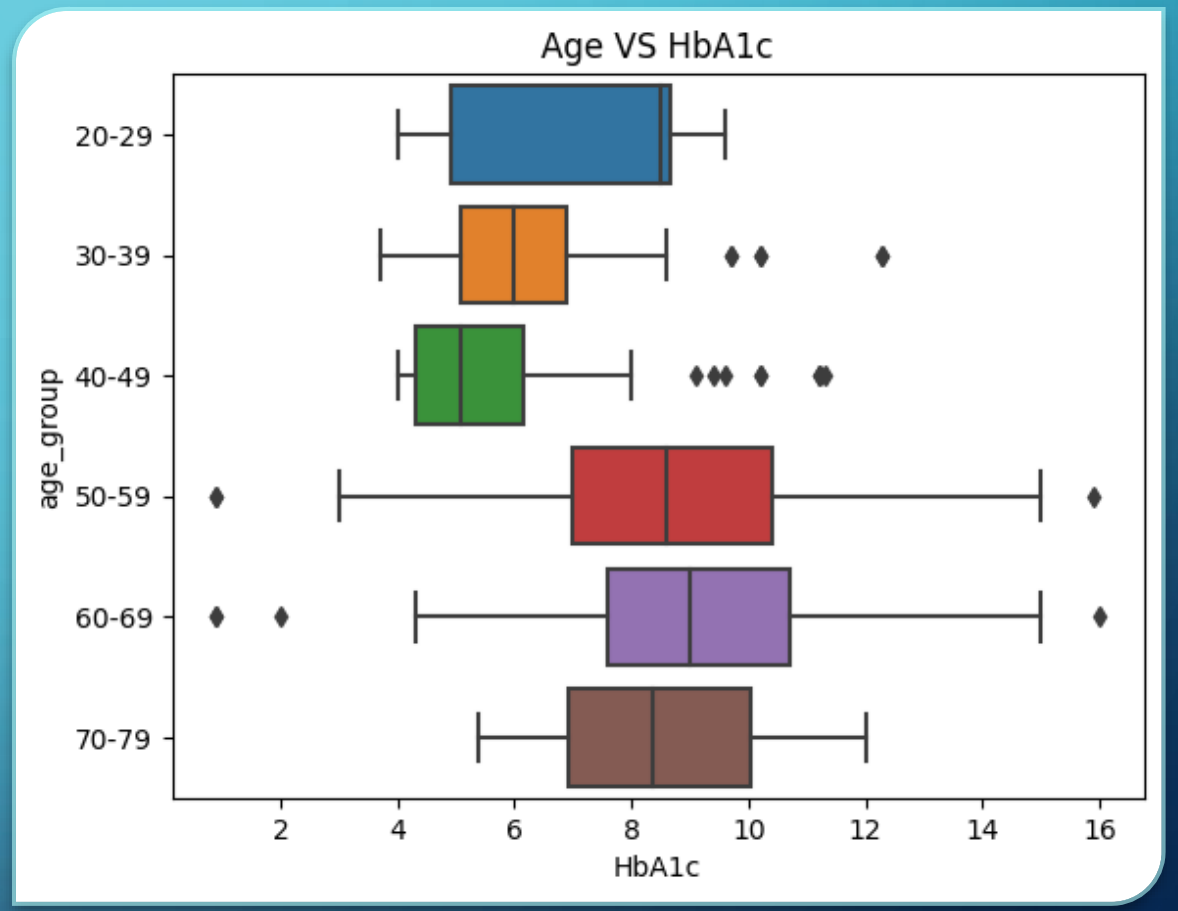
- HbA1c VS Counts histogram plot.
- From the plot we can see that the peak is between 6% and 8%, indicating that a lot of people within the dataset are diabetic and according to:

Normal		**Below 6.0%**
Prediabetes		**6.0% to 6.4%**
Diabetes		**6.5% or over**

(Diabetes Uk, 2019).



- HbA1c VS Age box and whisker plot.
- From the plot we can see that most patients above the age of 50 years are seem to have a Hb1Ac of about 6% and above, indicating that they are Predicted Diabetic and Diabetic.
- For the age groups below 50 years the patients are distributed across the Non-Diabetic, Predicted Diabetic and Diabetic range.

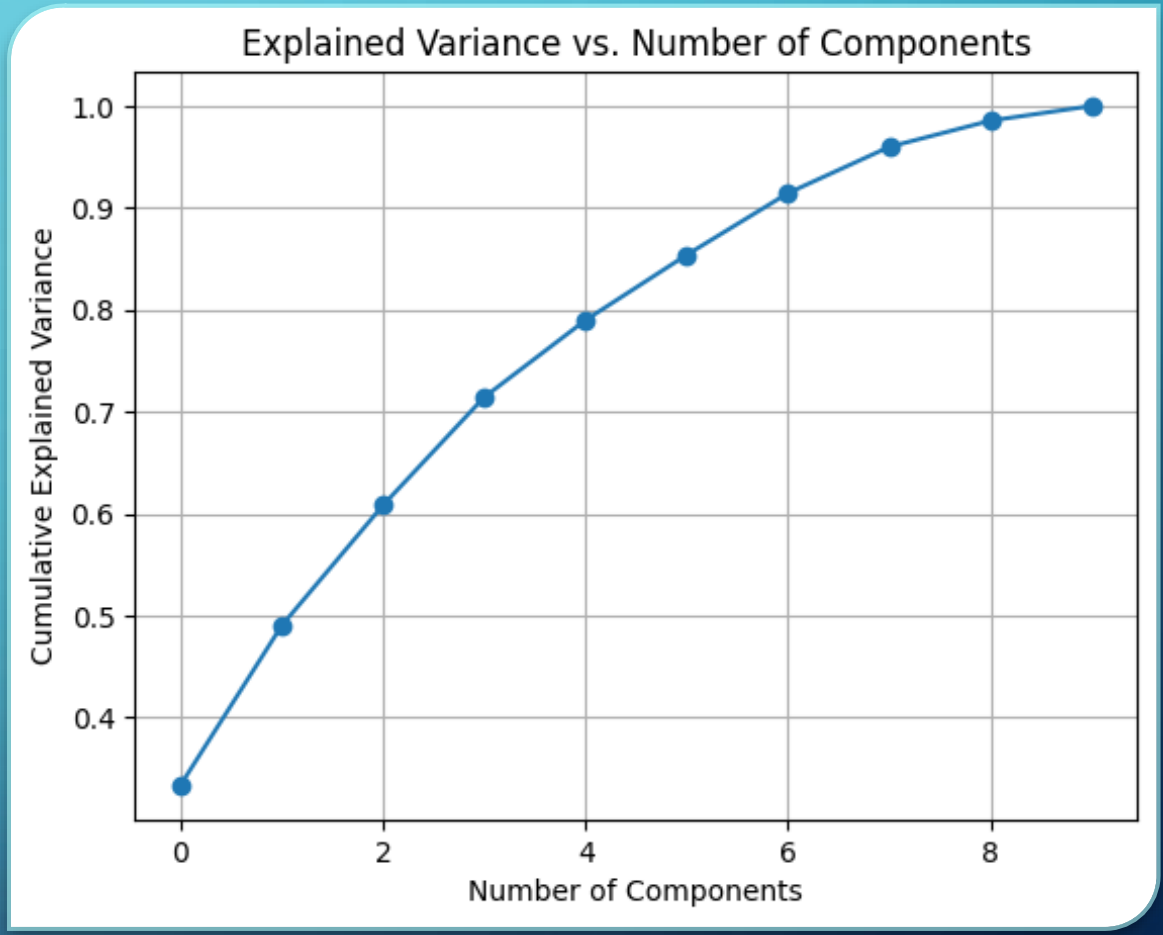


DATA PREPROCESSING

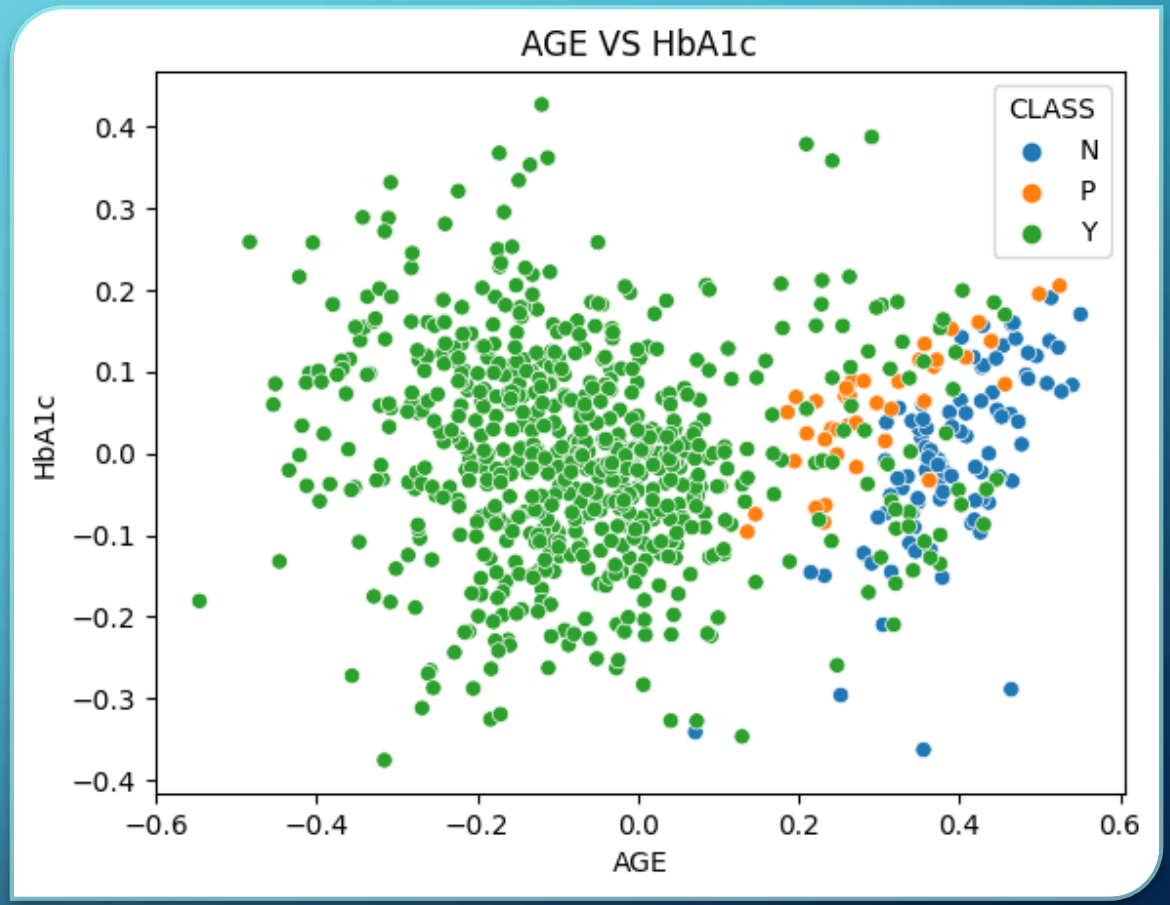
- To prepare the data for PCA analysis, I made sure to Normalize the data.
- Meaning that I scaled it between 0 and 1.
- This was done to ensure the PCA can capture the variance of the dataset more accurately.
- As some features high values within the hundreds while some were below ten.

PRINCIPAL COMPONENT ANALYSIS RESULTS

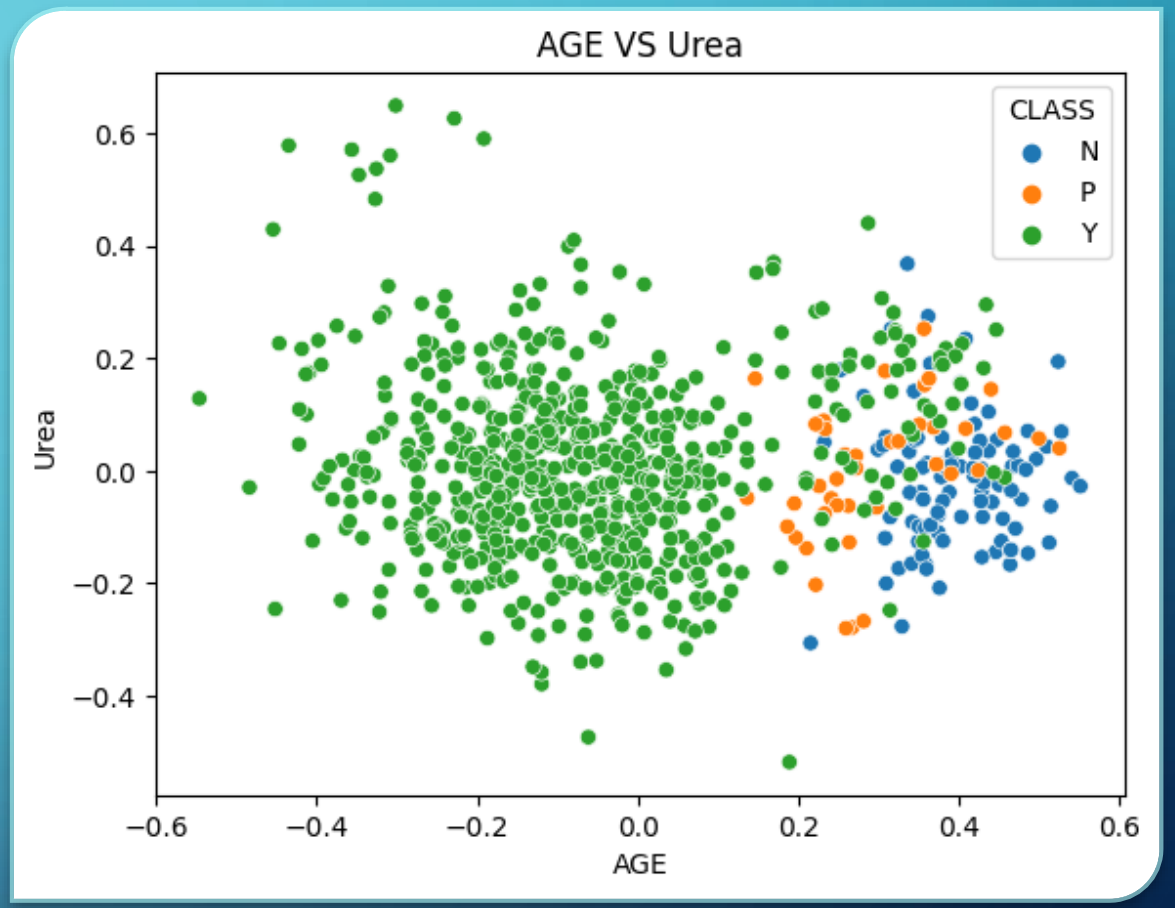
- Performing the PCA algorithm I found that at least 4 out of 8 features can explain about 80% of the data variance as seen of there Explained Variance VS No. of Components plot.
- These features were:
  - Age with 33%
  - Urea with 16%
  - Creatinine ratio with 12%
  - HbA1c with 11%
- Whereby the percentages are explained variance for each principal component.



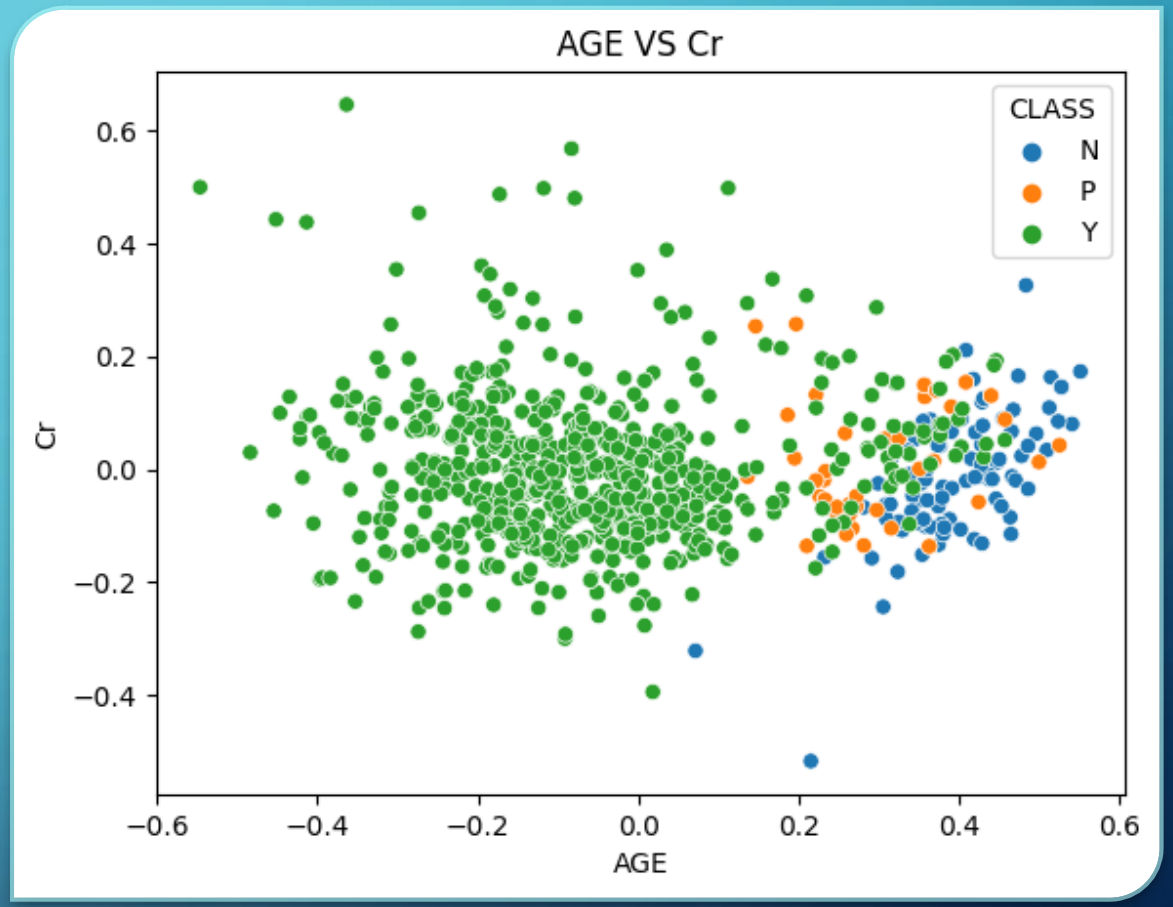
- Scatterplot of HbA1c VS Age.
- When observing the plot we can see that there are two significant clusters, one with Diabetic Patients and the second one with a mixture of Diabetic and mostly Predicted Diabetic and Not Diabetic.
- From the plot we can see that mostly patients of middle ages have higher glucose levels within their blood as observed from the outliers on the top of the plot.
- Overall HbA1c is not a good indicator "alone" of diabetes as even patients who are not or predicted Diabetic are observed at some of the same levels of HbAc1 which other patients were classified as Diabetic.



- Scatterplot of Age VS Urea
- When observing the plot we can see that there are two significant clusters as well, one with Diabetic Patients and the second one with a mixture of Diabetic and mostly Predicted Diabetic and Not Diabetic.
- From the plot we can see that mostly diabetic patients at younger ages seem to have a higher Urea Concentration , compared to older patients (Observe outliers).
- This might indicate a relationship between diabetic patients and urea concentration.
- Having a higher urea concentration means your kidneys are no longer properly filtrating your urine. (medlineplus.gov, n.d.)
- But that “alone” as well cannot tell if a patient is diabetic as we can see that distribution of Diabetic patients and Not or Predict Diabetic is centered around the same region of Urea Concentration.



- Scatterplot of Age VS Creatinine ratio
- When observing the plot we can see that there are two significant clusters as well, one with Diabetic Patients and the second one with a mixture of Diabetic and mostly Predicted Diabetic and Not Diabetic.
- From the plot we can see that mostly diabetic patients at younger ages seem to have a higher Creatinine ratio, compared to older patients (Observe outliers).
- Having a higher Creatinine ratio my indicated severe kidney damage (Cunha, 2018).
- Like the first two features Creatinine ratio "alone" cannot tell if a patient is diabetic as we can see that distribution of Diabetic patients and Not or Predict Diabetic is centered around the same region of Creatinine ratio.



CONCLUSION

- I was able to use the PCA algorithm to extract the features that can explain about 80% of the data.
- I was able to plot those features against each other to try and further understand the relationship between Diabetes and the extracted features.
- One thing I noticed is that alone these features cannot tell us if someone is diabetic or not but if we combined these features it is possible as most of the diabetic patients where found in almost a similar region within the plots as well as not and predicted diabetic patients.

REFERENCES

- Diabetes Uk (2019). *What Is HbA1c? - Definition, Units, Conversion, Testing & Control*. [online] Diabetes.co.uk. Available at: https://www.diabetes.co.uk/what-is-hba1c.html.
- medlineplus.gov. (n.d.). *BUN (Blood Urea Nitrogen): MedlinePlus Medical Test*. [online] Available at: https://medlineplus.gov/lab-tests/bun-blood-urea-nitrogen
- Cunha, J.P. (2018). *High, Low, & Normal Creatinine Levels: What This Blood Test Means*. [online] eMedicineHealth. Available at: https://www.emedicinehealth.com/creatinine_blood_tests/article_em.htm.