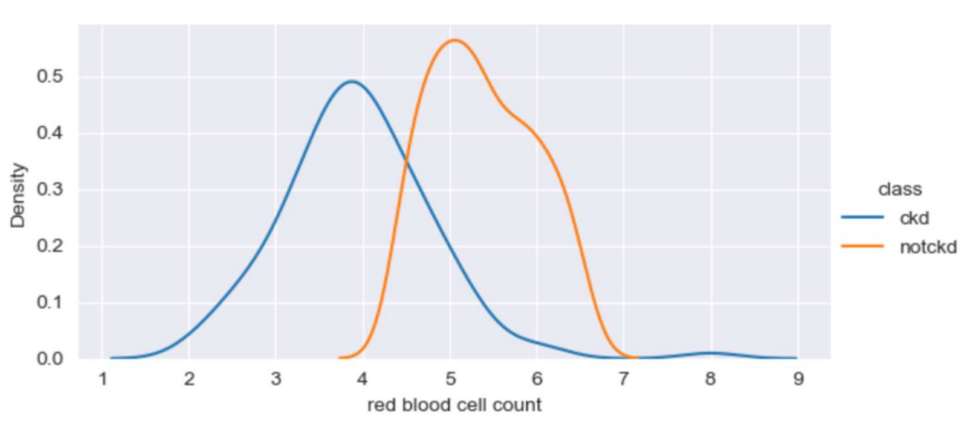


Data Collection and Preprocessing Phase

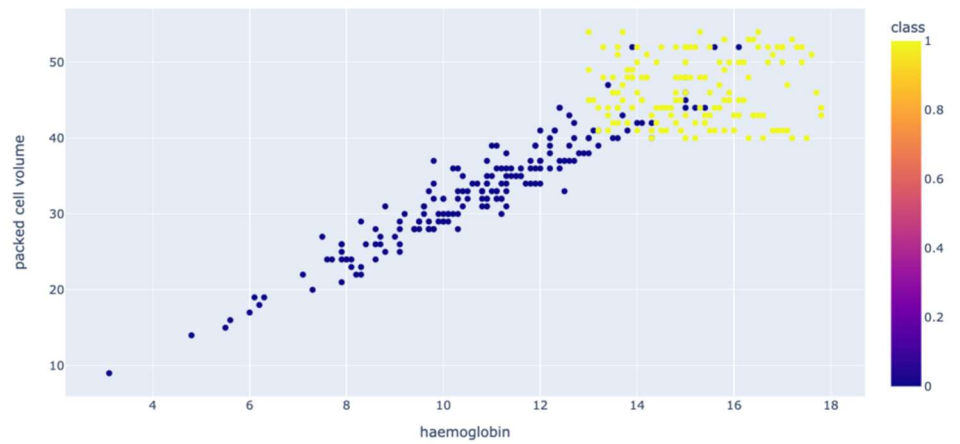
| | |
|---------------|---|
| Date | 08 July 2024 |
| Team ID | SWTID1720190579 |
| Project Title | Early Prediction Of Chronic Kidney Disease Using Machine Learning |
| Maximum Marks | 6 Marks |

Data Exploration and Preprocessing

The variables of the dataset will be statistically examined to find general trends and extremes, and for this, a tool such as Python used for preprocessing like normalization and feature engineering activities. Data cleaning will find missing value analysis it determines the ways of handling missing values and outliers to improve the quality of the data in the upcoming analysis or modeling process.

| Section | Description |
|---------------------|---|
| Data Overview | <pre> : id age bp sg al su bgr bu sc sod pot hemo count 400.000000 391.000000 388.000000 353.000000 354.000000 351.000000 356.000000 381.000000 383.000000 313.000000 312.000000 348.000000 mean 199.500000 51.483376 76.469072 1.017408 1.016949 0.450142 148.036517 57.425722 3.072454 137.528754 4.627244 12.526437 std 115.614301 17.169714 13.683637 0.005717 1.352679 1.099191 79.281714 50.503006 5.741126 10.408752 3.193904 2.912587 min 0.000000 2.000000 50.000000 1.005000 0.000000 0.000000 22.000000 1.500000 0.400000 4.500000 2.500000 3.100000 25% 99.750000 42.000000 70.000000 1.010000 0.000000 0.000000 99.000000 27.000000 0.900000 135.000000 3.800000 10.300000 50% 199.500000 55.000000 80.000000 1.020000 0.000000 0.000000 121.000000 42.000000 1.300000 138.000000 4.400000 12.650000 75% 299.250000 64.500000 80.000000 1.020000 2.000000 0.000000 163.000000 66.000000 2.800000 142.000000 4.900000 15.000000 max 399.000000 90.000000 180.000000 1.025000 5.000000 5.000000 490.000000 391.000000 76.000000 163.000000 47.000000 17.800000 </pre> |
| Univariate Analysis |  |

Bivariate Analysis

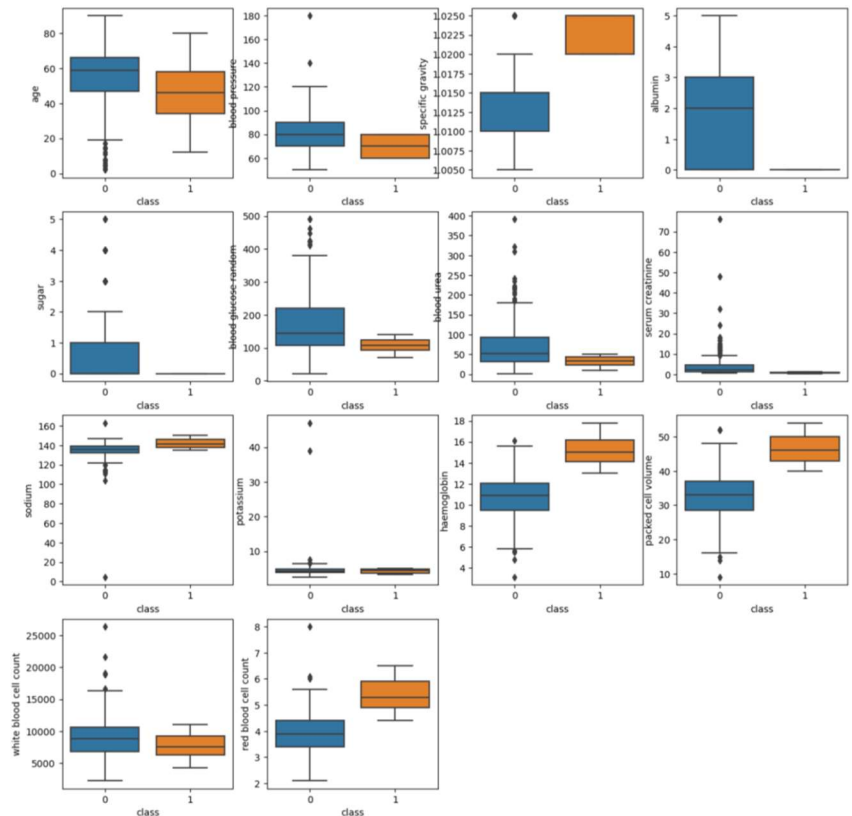


Multivariate Analysis

<Axes: xlabel='age', ylabel='blood pressure'>



Outliers and Anomalies



Data Preprocessing Code Screenshots

Loading Data

| | id | age | blood pressure | specific gravity | albumin | sugar | red blood cells | pus cell | pus cell clumps | bacteria | ... | packed cell volume | white blood cell count | red blood cell count | hypertension | diabetes mellitus | coronary artery disease | appet |
|-----|-----|------|----------------|------------------|---------|-------|-----------------|----------|-----------------|------------|-----|--------------------|------------------------|----------------------|--------------|-------------------|-------------------------|-------|
| 0 | 0 | 48.0 | 80.0 | 1.020 | 1.0 | 0.0 | NaN | normal | notpresent | notpresent | ... | 44 | 7800 | 5.2 | yes | yes | no | go |
| 1 | 1 | 7.0 | 50.0 | 1.020 | 4.0 | 0.0 | NaN | normal | notpresent | notpresent | ... | 38 | 6000 | NaN | no | no | no | go |
| 2 | 2 | 62.0 | 80.0 | 1.010 | 2.0 | 3.0 | normal | normal | notpresent | notpresent | ... | 31 | 7500 | NaN | no | yes | no | px |
| 3 | 3 | 48.0 | 70.0 | 1.005 | 4.0 | 0.0 | normal | abnormal | present | notpresent | ... | 32 | 6700 | 3.9 | yes | no | no | px |
| 4 | 4 | 51.0 | 80.0 | 1.010 | 2.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 35 | 7300 | 4.6 | no | no | no | go |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 395 | 395 | 55.0 | 80.0 | 1.020 | 0.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 47 | 6700 | 4.9 | no | no | no | go |
| 396 | 396 | 42.0 | 70.0 | 1.025 | 0.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 54 | 7800 | 6.2 | no | no | no | go |
| 397 | 397 | 12.0 | 80.0 | 1.020 | 0.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 49 | 6600 | 5.4 | no | no | no | go |
| 398 | 398 | 17.0 | 60.0 | 1.025 | 0.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 51 | 7200 | 5.9 | no | no | no | go |
| 399 | 399 | 58.0 | 80.0 | 1.025 | 0.0 | 0.0 | normal | normal | notpresent | notpresent | ... | 53 | 6800 | 6.1 | no | no | no | go |

400 rows x 26 columns

Handling Missing Data

```
df['diabetes mellitus'].replace(to_replace = {'\tno':'no', '\tyes':'yes', ' yes':'yes'}, inplace=True)
df['coronary artery disease'] = df['coronary artery disease'].replace(to_replace = '\tno', value='no')
df['class'] = df['class'].replace(to_replace = 'ckd\t', value = 'ckd')

for col in cat_col:
    print('{} has {} values'.format(col, df[col].unique()))
    print('\n')
```

| | |
|---------------------|---|
| Data Transformation | - |
| Feature Engineering | - |
| Save Processed Data | - |