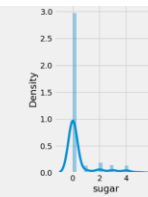
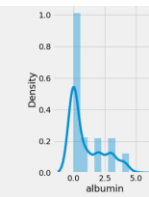
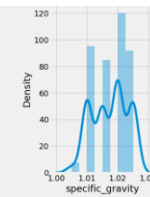
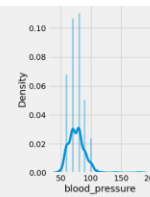
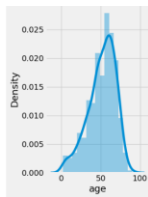
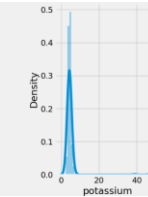
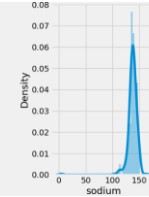
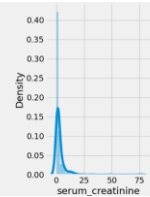
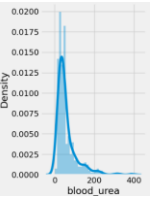
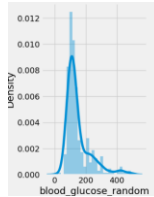


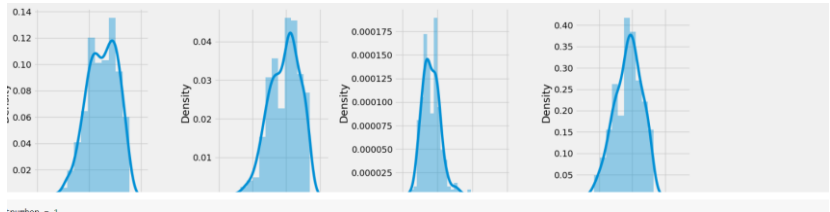
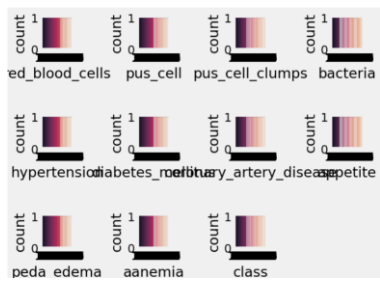
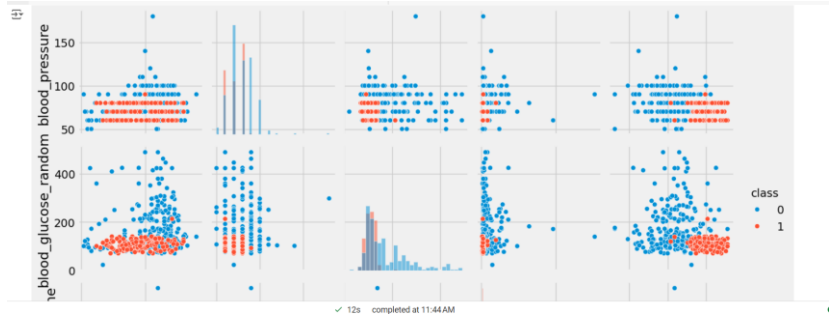
## Data Collection and Preprocessing Phase

Date	16 JULY 2024
Team ID	SWTID1720075199
Project Title	Early Prediction Of Chronic Kidney Disease Using Machine Learning
Maximum Marks	6 Marks

### Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																												
Data Overview	<div>400 rows x 26 columns.</div> <table><thead><tr><th></th><th>age</th><th>blood_pressure</th><th>specific_gravity</th><th>albumin</th><th>sugar</th><th>blood_glucose_random</th><th>blood_urea</th><th>serum_creatinine</th><th>sodium</th><th>potassium</th><th>haemoglobin</th></tr></thead><tbody><tr><td>count</td><td>391.000000</td><td>388.000000</td><td>353.000000</td><td>354.000000</td><td>351.000000</td><td>356.000000</td><td>381.000000</td><td>383.000000</td><td>313.000000</td><td>312.000000</td><td>348.000000</td></tr><tr><td>mean</td><td>51.483376</td><td>76.469072</td><td>1.017408</td><td>1.016949</td><td>0.450142</td><td>148.036517</td><td>57.425722</td><td>3.072454</td><td>137.528754</td><td>4.627244</td><td>12.526437</td></tr><tr><td>std</td><td>17.169714</td><td>13.683637</td><td>0.005717</td><td>1.352679</td><td>1.099191</td><td>79.281714</td><td>50.503006</td><td>5.741126</td><td>10.408752</td><td>3.193904</td><td>2.912587</td></tr><tr><td>min</td><td>2.000000</td><td>50.000000</td><td>1.005000</td><td>0.000000</td><td>0.000000</td><td>22.000000</td><td>1.500000</td><td>0.400000</td><td>4.500000</td><td>2.500000</td><td>3.100000</td></tr><tr><td>25%</td><td>42.000000</td><td>70.000000</td><td>1.010000</td><td>0.000000</td><td>0.000000</td><td>99.000000</td><td>27.000000</td><td>0.900000</td><td>135.000000</td><td>3.800000</td><td>10.300000</td></tr><tr><td>50%</td><td>55.000000</td><td>80.000000</td><td>1.020000</td><td>0.000000</td><td>0.000000</td><td>121.000000</td><td>42.000000</td><td>1.300000</td><td>138.000000</td><td>4.400000</td><td>12.650000</td></tr><tr><td>75%</td><td>64.500000</td><td>80.000000</td><td>1.020000</td><td>2.000000</td><td>0.000000</td><td>163.000000</td><td>66.000000</td><td>2.800000</td><td>142.000000</td><td>4.900000</td><td>15.000000</td></tr><tr><td>max</td><td>90.000000</td><td>180.000000</td><td>1.025000</td><td>5.000000</td><td>5.000000</td><td>490.000000</td><td>391.000000</td><td>76.000000</td><td>163.000000</td><td>47.000000</td><td>17.800000</td></tr></tbody></table>		age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin	count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000	mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437	std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587	min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000	25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000	50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000	75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000	max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000
	age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin																																																																																																		
count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000																																																																																																		
mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437																																																																																																		
std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587																																																																																																		
min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000																																																																																																		
25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000																																																																																																		
50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000																																																																																																		
75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000																																																																																																		
max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000																																																																																																		
Univariate Analysis	<div></div> <div></div>																																																																																																												

	
Bivariate Analysis	
Multivariate Analysis	
Outliers and Anomalies	None
<b>Data Preprocessing Code Screenshots</b>	
Loading Data	<pre>import pandas as pd  path="/content/drive/MyDrive/chronickidneydisease/chronickidneydisease.csv" df=pd.read_csv(path) #we read the data here</pre>
Handling Missing Data	<pre>  def random_value_imputation(feature):       random_sample = df[feature].dropna().sample(df[feature].isna().sum())       random_sample.index = df[df[feature].isnull()].index       df.loc[df[feature].isnull(), feature] = random_sample    def impute_mode(feature):       mode = df[feature].mode()[0]       df[feature] = df[feature].fillna(mode)    for col in num_cols:       random_value_imputation(col)</pre>

	<pre>7] random_value_imputation('red_blood_cells')    random_value_imputation('pus_cell')     for col in cat_cols:        impute_mode(col)</pre>
Data Transformation	Code for transforming variables (scaling, normalization).
Feature Engineering	<pre>cat_cols = [col for col in df.columns if df[col].dtype == 'object'] num_cols = [col for col in df.columns if df[col].dtype != 'object']  for col in cat_cols:     print(f"{col} has {df[col].unique()} values\n")</pre>
Save Processed Data	<pre>import pandas as pd  # Assuming df is your cleaned and processed DataFrame # Perform your preprocessing steps before this point  # Save to CSV file df.to_csv('processed_data.csv', index=False)</pre>