

Data Collection and Preprocessing Phase

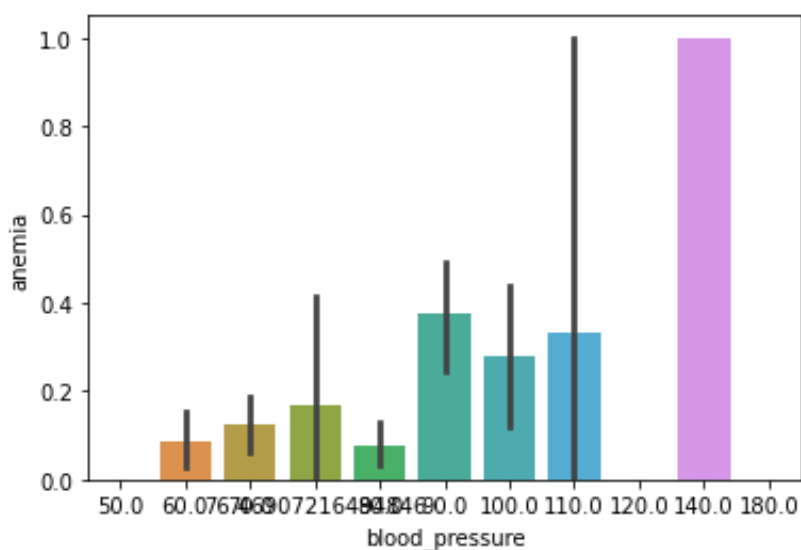
Date	15 July 2024
Team ID	739846
Project Title	Early Prediction Of Chronic Kidney Disease
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

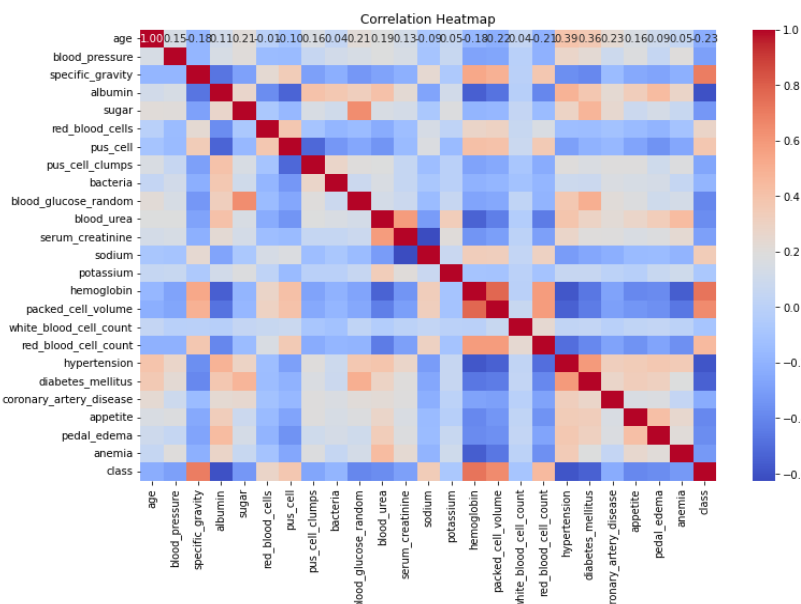
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																																																				
Data Overview	<div>Dimensions 399rows x 26columns</div> <table><tr><th></th><th>id</th><th>age</th><th>bp</th><th>sg</th><th>al</th><th>su</th><th>rbc</th><th>pc</th><th>pcc</th><th>ba</th><th>...</th><th>pcv</th><th>wc</th><th>rc</th><th>htn</th><th>dm</th><th>cad</th><th>appet</th><th>pe</th><th>ane</th><th>classification</th></tr><tr><td>0</td><td>0</td><td>48.0</td><td>80.0</td><td>1.020</td><td>1.0</td><td>0.0</td><td>NaN</td><td>normal</td><td>notpresent</td><td>notpresent</td><td>...</td><td>44</td><td>7800</td><td>5.2</td><td>yes</td><td>yes</td><td>no</td><td>good</td><td>no</td><td>no</td><td>ckd</td></tr><tr><td>1</td><td>1</td><td>7.0</td><td>50.0</td><td>1.020</td><td>4.0</td><td>0.0</td><td>NaN</td><td>normal</td><td>notpresent</td><td>notpresent</td><td>...</td><td>38</td><td>6000</td><td>NaN</td><td>no</td><td>no</td><td>no</td><td>good</td><td>no</td><td>no</td><td>ckd</td></tr><tr><td>2</td><td>2</td><td>62.0</td><td>80.0</td><td>1.010</td><td>2.0</td><td>3.0</td><td>normal</td><td>normal</td><td>notpresent</td><td>notpresent</td><td>...</td><td>31</td><td>7500</td><td>NaN</td><td>no</td><td>yes</td><td>no</td><td>poor</td><td>no</td><td>yes</td><td>ckd</td></tr><tr><td>3</td><td>3</td><td>48.0</td><td>70.0</td><td>1.005</td><td>4.0</td><td>0.0</td><td>normal</td><td>abnormal</td><td>present</td><td>notpresent</td><td>...</td><td>32</td><td>6700</td><td>3.9</td><td>yes</td><td>no</td><td>no</td><td>poor</td><td>yes</td><td>yes</td><td>ckd</td></tr><tr><td>4</td><td>4</td><td>51.0</td><td>80.0</td><td>1.010</td><td>2.0</td><td>0.0</td><td>normal</td><td>normal</td><td>notpresent</td><td>notpresent</td><td>...</td><td>35</td><td>7300</td><td>4.6</td><td>no</td><td>no</td><td>no</td><td>good</td><td>no</td><td>no</td><td>ckd</td></tr></table>		id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification	0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd	1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd	2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd	3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd	4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification																																																																																																																
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd																																																																																																																
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd																																																																																																																
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd																																																																																																																
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd																																																																																																																
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd																																																																																																																
Univariate Analysis	<div><div><div>Distribution of age</div></div><div><div>Distribution of blood pressure</div></div><div><div>Distribution of sugar</div></div></div>																																																																																																																																				

Bivariate Analysis



Multivariate Analysis



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data

```
# Load dataset
filename = "C:\\Users\\soumy\\OneDrive\\Desktop\\mini p\\data.csv"
data = pd.read_csv(filename)
```

Handling Missing Data	<pre># Handle missing values data = data.apply(lambda x: x.fillna(x.mean()) if x.dtype.kind in 'biufc' else x.fillna(x.mode().iloc[0]))</pre>
Data Transformation	<pre># Clean categorical variables with incorrect values data['class'] = data['class'].replace("ckd\t", "ckd") data['coronary_artery_disease'] = data['coronary_artery_disease'].replace("\tno", 'no') data['diabetes_mellitus'] = data['diabetes_mellitus'].replace(to_replace={"\tno": 'no', '\tyes': 'yes', 'yes': 'yes'})</pre> <pre># Encode categorical variables cat_cols = data.select_dtypes(include=['object']).columns label_encoders = {} for col in cat_cols: le = LabelEncoder() data[col] = le.fit_transform(data[col]) label_encoders[col] = le</pre>
Feature Engineering	Attached code in final submission.
Save Processed Data	-