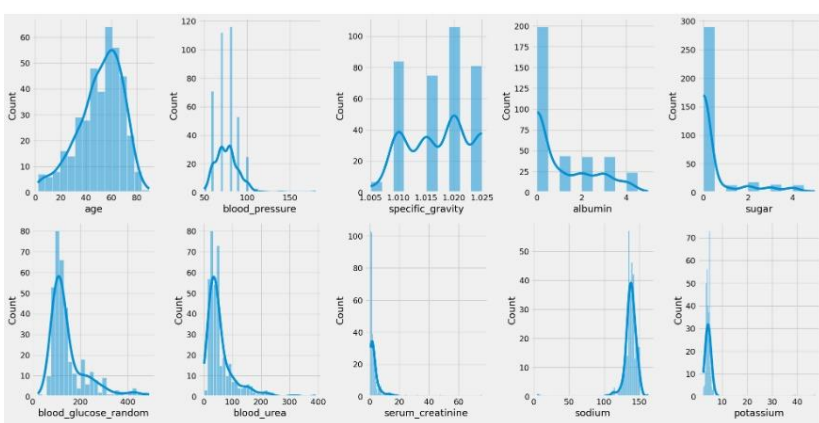
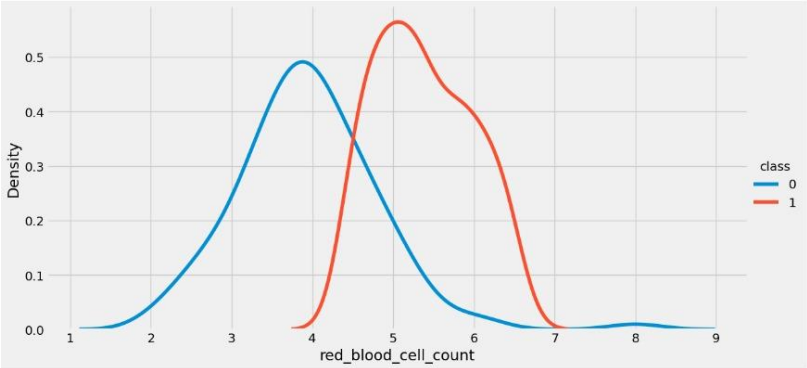
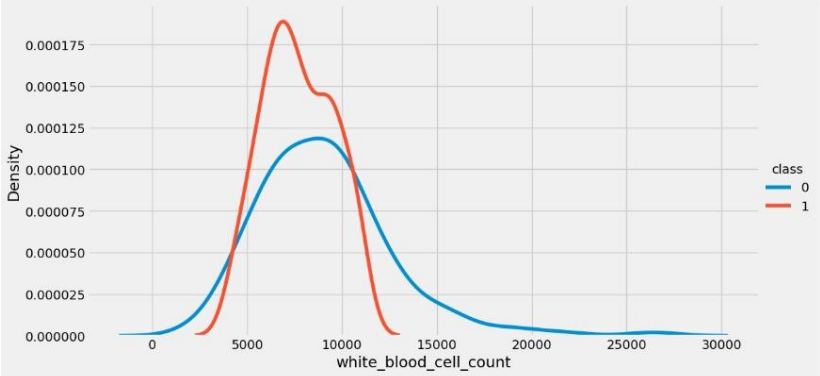
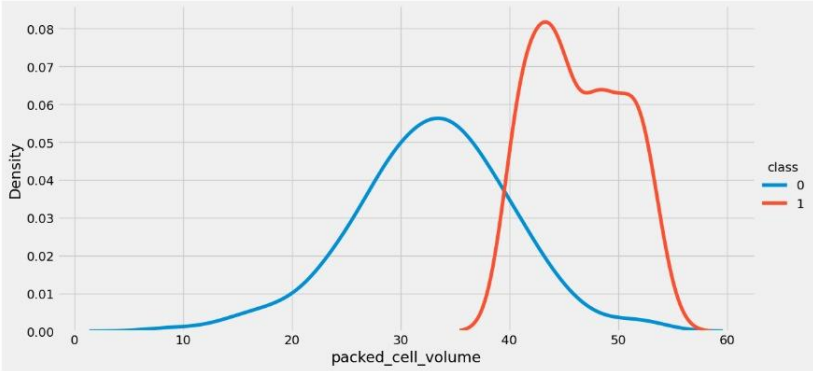
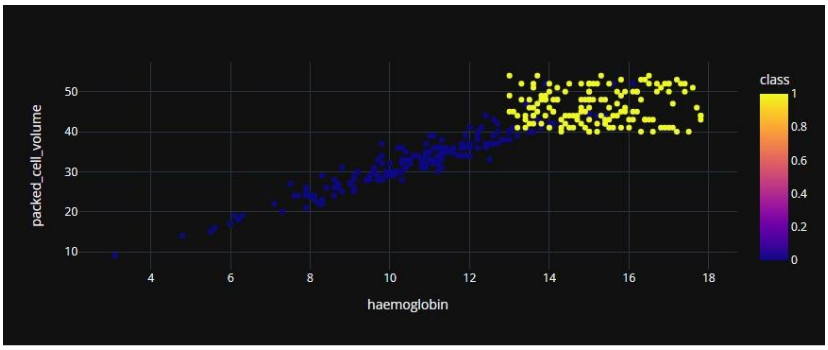
Data Collection and Preprocessing Phase

Date	4th June 2024
Team ID	SWTID1720164961
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	400 rows 26 columns [5]: data.head() [5]: |	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	cla
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
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
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
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
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
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	 5 rows × 26 columns
Univariate Analysis	

Bivariate Analysis	  
Multivariate Analysis	

Outliers and Anomalies	NA
Data Preprocessing Code Screenshots	
Loading Data	[5]: `data.head()` [5]: |	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	cla
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
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
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
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
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
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	 5 rows × 26 columns

Handling Missing Data	[66]:
	`# filling null values, we will use two methods, random sampling for higher null values and`
	`# mean/mode sampling for lower null values`
	`def random_value_imputation(feature):`
	`random_sample = data[feature].dropna().sample(data[feature].isna().sum())`
	`random_sample.index = data[data[feature].isnull()].index`
	`data.loc[data[feature].isnull(), feature] = random_sample`
	`def impute_mode(feature):`
	`mode = data[feature].mode()[0]`
	`data[feature] = data[feature].fillna(mode)`
	[67]:
	`# filling num_cols null values using random sampling method`
	`for col in num_cols:`
	`random_value_imputation(col)`
	[68]: `data[num_cols].isnull().sum()`
	[68]:
	`age 0`
	`blood_pressure 0`
	`specific_gravity 0`
	`albumin 0`
	`sugar 0`
	`blood_glucose_random 0`
	`blood_urea 0`
	`serum_creatinine 0`
	`sodium 0`
	`potassium 0`
	`haemoglobin 0`
	`packed_cell_volume 0`
	`white_blood_cell_count 0`
	`red_blood_cell_count 0`
	`dtype: int64`
	[69]:
	`# filling "red_blood_cells" and "pus_cell" using random sampling method and rest of cat_cols using mode imputation`
	`random_value_imputation('red_blood_cells')`
	`random_value_imputation('pus_cell')`
	`for col in cat_cols:`
	`impute_mode(col)`
	[71]: `data[cat_cols].isnull().sum()`

Data Transformation	NA
Feature Engineering	NA
Save Processed Data	NA