



Data Collection and Preprocessing Phase

Date	12 JULY 2024
ID	740069
Project Title	Lymphography Classification using ML
Maximum Marks	6 Marks

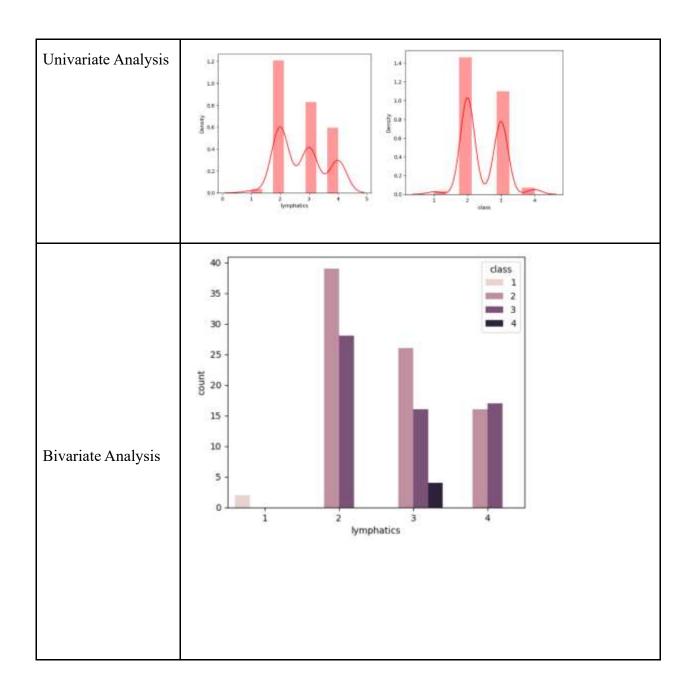
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Desc	ription	1													
	<u>Dimension:</u> 148 rows × 19 columns															
	Descriptive statistics:															
		class	lymphatics	block of affere	bl. of lymph. c	bl. of lymph, s	by pass	extravasates	regeneration of	early uptake in	lym.nodes dimin	lym.nodes enlar	changes in lyw.	defect in node	changes in node	changes in stru
Data	count	148,000000	148.000000	148.000000	148.000000	148.000000	148.000000	148,000000	148.000000	148.000000	148.000000	148.000000	148,000000	148,000000	148.000000	148.000000
Overview	mean	2.452703	2.743243	1.554054	1.175676	1.047297	1.243243	1.506757	1.067568	1.702703	1.060811	2.472973	2.398649	2.965216	2.804054	5.216216
	std	0.575396	0.817509	0.498757	0.381836	0.212996	0.430498	0.501652	0.251855	0.458621	0.313557	0.836827	0.568323	0.868305	0.761834	2.171368
	min	1,000000	1.000000	1.000000	1.000000	1.000000	1,000000	1,000000	1.000000	1,000000	1.000000	1,000000	1,000000	1,000000	1.000000	1,000000
	25%	2.000000	2.000000	1.000000	1,000000	1.000000	1,000000	1,000000	1,000000	1,000000	1.000000	2.000000	2.000000	2.000000	2.000000	4.000000
	50%	2.000000	3.000000	2.000000	1.000000	1.000000	1.000000	2,000000	1.000000	2.000000	1.000000	2.000000	2.000000	3.000000	3.000000	5.000000
	75%	3.000000	3.000000	2.000000	1.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	3.000000	3.000000	4.000000	3.000000	8,000000
	max	4.000000	4.000000	2.000000	2.000000	2.000000	2.000000	2,000000	2.000000	2.000000	3.000000	4.000000	3,000000	4.000000	4.000000	8,000000

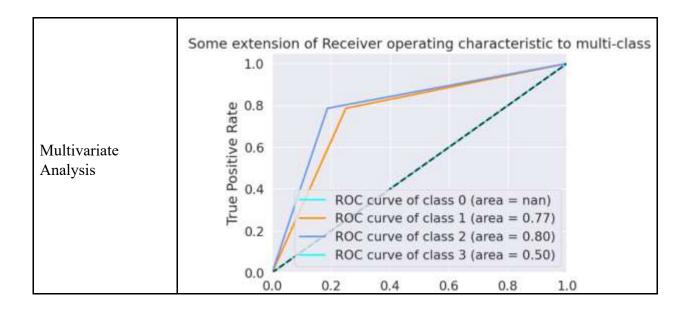












Outliers and Anomalies

Data Preprocessing Code Screenshots

Loading Data

	at .teas	fulse of data 4)																	
Ú.	e of a	dant: (18)	100																
	rlens	lymphatics	block of affere	til, af lymph, a	bl. of lymph. s	by pers	extravalates	regeneration of	early uptake in	lym.nodes dissin	lym.nodes enlar	changes in Tym.	defect in node	changes in rode	changes in ctru	uperial Forms	dislocation of	exclusion of no	hodi
0	3	+	2		1	1			- 2	- 1	2	- 2	- 2	+		- 1		- 2	
1	z		- 2			- 2	- 2		2		3	- 1	2	3	+	- 2	2	2	
2	3	- 3	- 2	2	2	- 7	2	2	- 2	- 1			- 3	+		- 1	2	7	
3	3	- 3				1	2					- 3			4		ŧ	2	
4	2		- 1		1	11	1	1	1	1	2	- 2	4	3		- 1	2	- 2	





```
1 for col in df.columns:
                           q1 = np.quantile(df[col],0.25)
                           3
                              q3 = np.quantile(df[col],0.75)
                          4 iqr = q3-q1
                          5 lower_bound = q1 - (1.5*iqr)
                           6 upper_bound = q3 + (1.5*iqr)
                               df[col] = np.where(df[col]> upper_bound,upper_bound,df[col])
                               df[col] = np.where(df[col]< lower_bound,lower_bound,df[col])</pre>
                          8
Handling
                          9 sns.boxplot(df[col])
Outliers
                         10 print("")
                               plt.show()
                         11
                  [23] I # Assuming 'class' is your target variable and the rest are features
                       2 y = df['class'] # Create y to hold your target variable
3 x - df.drop('class', axis-1) # Create x to bold your features
                       # x_train,x_test,y_train,y_test-train_test_split(x,y,test_size=0.2,random_state=42)
Training
                                                                                                                ↑↓回回•□□:
                       1 ###check shape to make sure it Is all in order
and Testing
                       2 print("size of x_train: {} \t size of x_test: {} \nsize of y_train:{} \t size of y_test: {}".formut(x_train.shape,x_test.shape,y_train.shape
Feature
                 Attached the codes in final submission.
Engineering
Save
Processed
Data
```