Data Exploration and Preprocessing Report

Team ID: SWTID1720067156

July 10, 2024

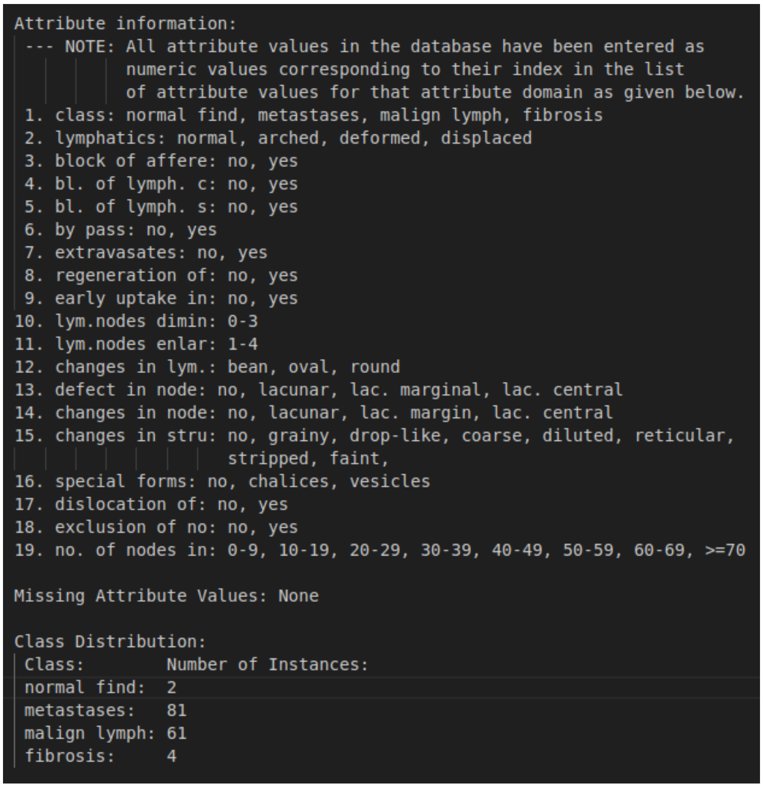
Data Collection and Preprocessing Phase

Team ID: SWTID1720067156
Project Title: Lymphography Classification Tool
Maximum Marks: 6 Marks

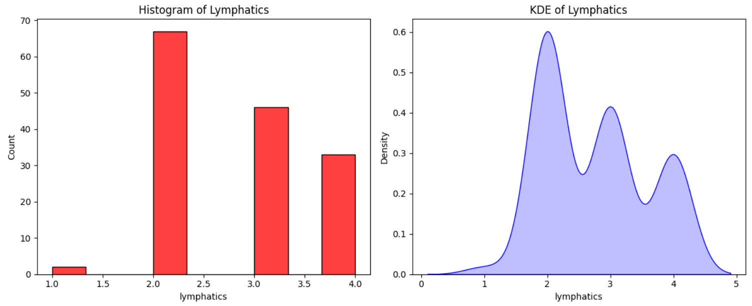
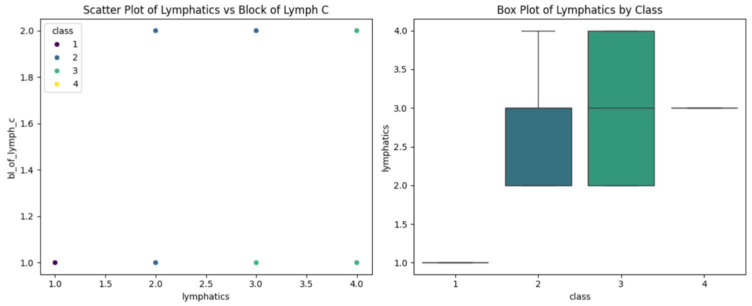
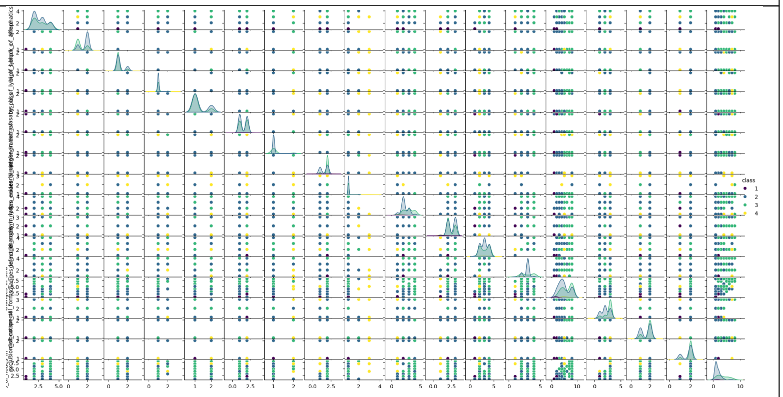
Data Exploration and Preprocessing Report

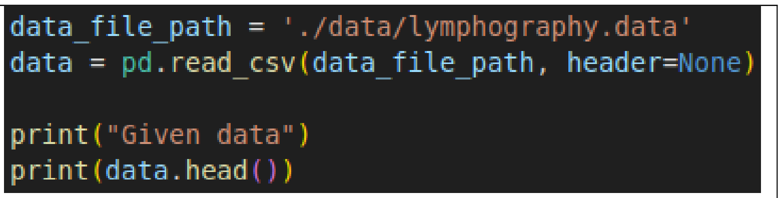
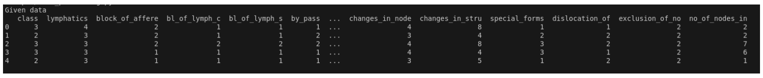
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

Section Description

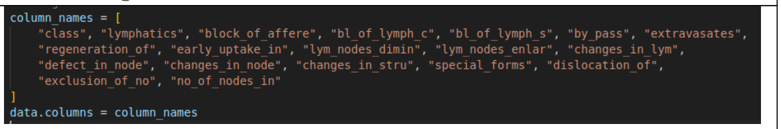
Section	**Description**
Data Overview	Dimension: 614 rows × 13 columns Descriptive statistics: 

```
Attribute information:
 --- NOTE: All attribute values in the database have been entered as
           numeric values corresponding to their index in the list
           of attribute values for that attribute domain as given below.
  1. class: normal find, metastases, malign lymph, fibrosis
  2. lymphatics: normal, arched, deformed, displaced
  3. block of affere: no, yes
  4. bl. of lymph. c: no, yes
  5. bl. of lymph. s: no, yes
  6. by pass: no, yes
  7. extravasates: no, yes
  8. regeneration of: no, yes
  9. early uptake in: no, yes
 10. lym.nodes dimin: 0-3
 11. lym.nodes enlar: 1-4
 12. changes in lym.: bean, oval, round
 13. defect in node: no, lacunar, lac. marginal, lac. central
 14. changes in node: no, lacunar, lac. margin, lac. central
 15. changes in stru: no, grainy, drop-like, coarse, diluted, reticular,
                      stripped, faint,
 16. special forms: no, chalices, vesicles
 17. dislocation of: no, yes
 18. exclusion of no: no, yes
 19. no. of nodes in: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70

Missing Attribute Values: None

Class Distribution:
  Class:        Number of Instances:
  normal find:  2
  metastases:   81
  malign lymph: 61
  fibrosis:     4
```

Univariate Analysis	
Bivariate Analysis	
Multivariate Analysis	
Data Preprocessing Code Screenshots	

Loading Data	`data_file_path = './data/lymphography.data'`
	`data = pd.read_csv(data_file_path, header=None)`
	
	`print("Given data")`
	`print(data.head())`
	`Given data`
	`   class  lymphatics  block_of_affere  bl_of_lymph_c  bl_of_lymph_s  by_pass  ...  changes_in_node  changes_in_stru  special_forms  dislocation_of  exclusion_of_no  no_of_nodes_in`
	`0      3           4                2              1              1        1  ...                4                8              1               1                2               2`
	`1      2           3                2              1              1        2  ...                3                4              2               2                2               2`
	`2      3           3                2              2              2        2  ...                4                8              3               2                2               7`
	`3      3           3                1              1              1        1  ...                4                4              3               1                2               6`
	`4      2           3                1              1              1        1  ...                3                5              1               2                2               1`

Handling Missing Data	No missing attributes

Data Transformation	`column_names = [`
	`    "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_lym",`
	`    "defect_in_node", "changes_in_node", "changes_in_stru", "special_forms", "dislocation_of",`
	`    "exclusion_of_no", "no_of_nodes_in"`
	`]`
	`data.columns = column_names`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	Done