

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 | <div>Dimension: 614 rows × 13 columns</div> <div>Descriptive statistics:</div> <div><div>Attribute information:</div><div><div>---</div><div>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.</div><div>1. class: normal find, metastases, malign lymph, fibrosis</div><div>2. lymphatics: normal, arched, deformed, displaced</div><div>3. block of affere: no, yes</div><div>4. bl. of lymph. c: no, yes</div><div>5. bl. of lymph. s: no, yes</div><div>6. by pass: no, yes</div><div>7. extravasates: no, yes</div><div>8. regeneration of: no, yes</div><div>9. early uptake in: no, yes</div><div>10. lym.nodes dimin: 0-3</div><div>11. lym.nodes enlar: 1-4</div><div>12. changes in lym.: bean, oval, round</div><div>13. defect in node: no, lacunar, lac. marginal, lac. central</div><div>14. changes in node: no, lacunar, lac. margin, lac. central</div><div>15. changes in stru: no, grainy, drop-like, coarse, diluted, reticular, stripped, faint,</div><div>16. special forms: no, chalices, vesicles</div><div>17. dislocation of: no, yes</div><div>18. exclusion of no: no, yes</div><div>19. no. of nodes in: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70</div></div><div>Missing Attribute Values: None</div><div><div>Class Distribution:</div><table><tr><th>Class:</th><th>Number of Instances:</th></tr><tr><td>normal find:</td><td>2</td></tr><tr><td>metastases:</td><td>81</td></tr><tr><td>malign lymph:</td><td>61</td></tr><tr><td>fibrosis:</td><td>4</td></tr></table></div></div> | Class: | Number of Instances: | normal find: | 2 | metastases: | 81 | malign lymph: | 61 | fibrosis: | 4 |
| Class: | Number of Instances: | | | | | | | | | | |
| normal find: | 2 | | | | | | | | | | |
| metastases: | 81 | | | | | | | | | | |
| malign lymph: | 61 | | | | | | | | | | |
| fibrosis: | 4 | | | | | | | | | | |

| | |
|-------------------------------------|---|
| Univariate Analysis |  <p>The univariate analysis section contains two plots. On the left is a 'Histogram of Lymphatics' showing the count of lymphatics for different values, with a prominent peak at 2.0. On the right is a 'KDE of Lymphatics' plot showing a multi-modal density distribution with peaks around 2.0, 3.0, and 4.0.</p> |
| Bivariate Analysis |  <p>The bivariate analysis section contains two plots. On the left is a 'Scatter Plot of Lymphatics vs Block of Lymph C' showing the relationship between lymphatics and block of lymph c for different classes. On the right is a 'Box Plot of Lymphatics by Class' showing the distribution of lymphatics across four classes.</p> |
| Multivariate Analysis |  <p>The multivariate analysis section displays a large grid of small plots, likely a faceted plot, showing the relationship between lymphatics and various other variables across different classes.</p> |
| Data Preprocessing Code Screenshots | |
| Loading Data |  <pre>data_file_path = './data/lymphography.data' data = pd.read_csv(data_file_path, header=None) print("Given data") print(data.head())</pre> <p>The screenshot shows the code to load the data and the first few rows of the resulting DataFrame.</p> |
| Handling Missing Data | No missing attributes |
| Data Transformation |  <pre>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</pre> <p>The screenshot shows the code to define column names and assign them to the data columns.</p> |
| Feature Engineering | Attached the codes in final submission. |
| Save Processed Data | Done |