

Data Collection and Preprocessing Phase

Date	12 july 2024
Team ID	739952
Project Title	Prediction and Analysis of Liver Patient Data Using Machine Learning
Maximum Marks	6 Marks

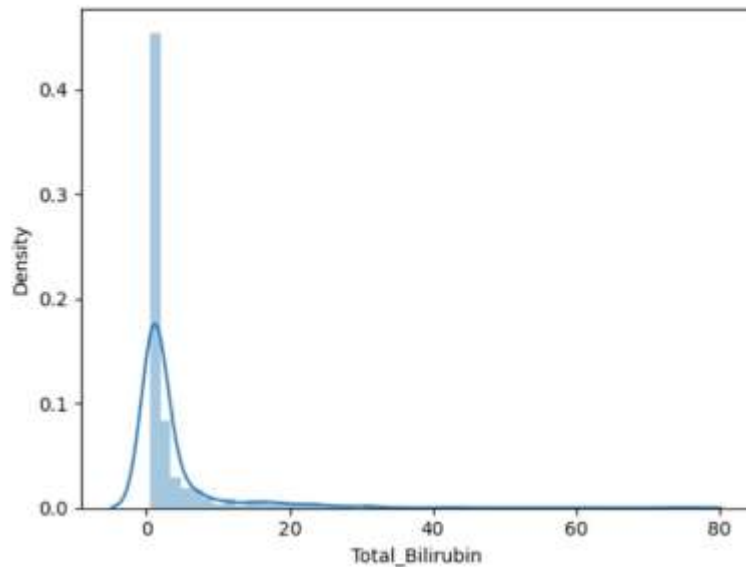
Data Exploration and Preprocessing Template

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

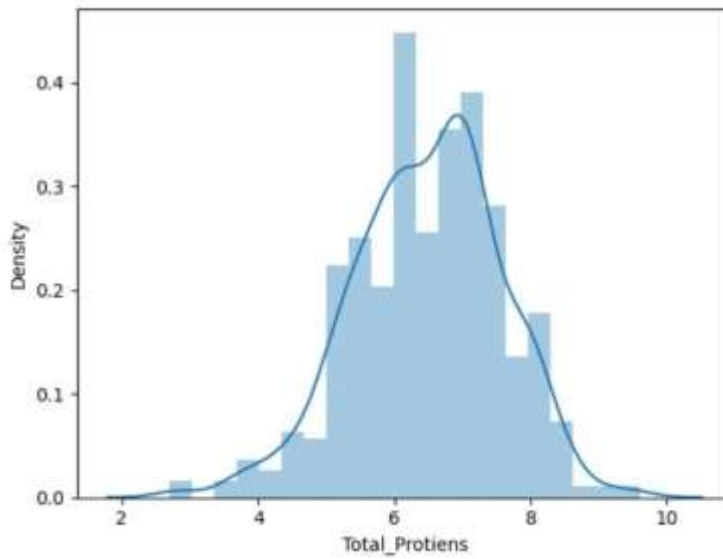
Section	Description
Data Overview	583 rows × 11 columns

Univariate Analysis

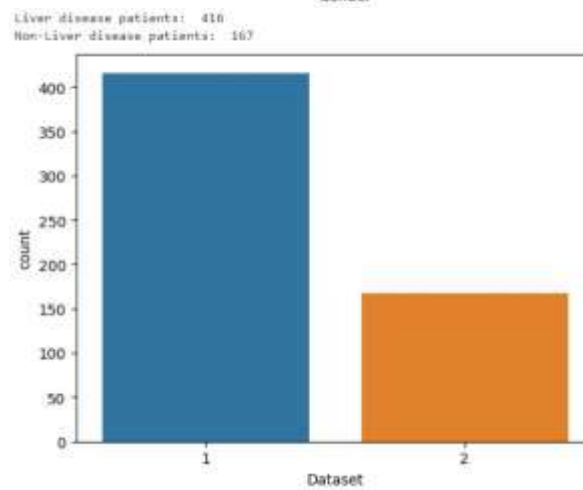
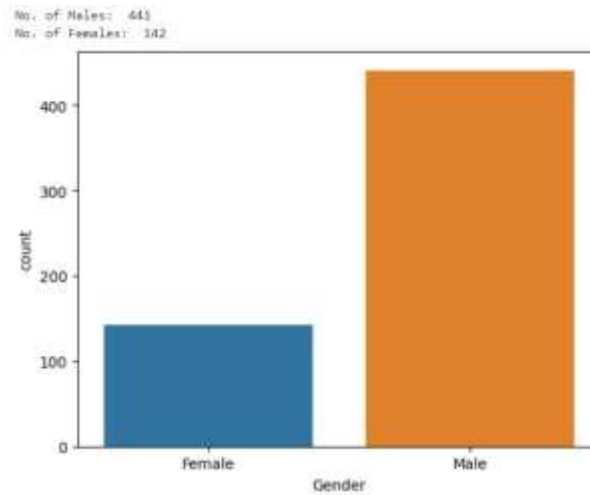
<Axes: xlabel='Total_Bilirubin', ylabel='Density'>



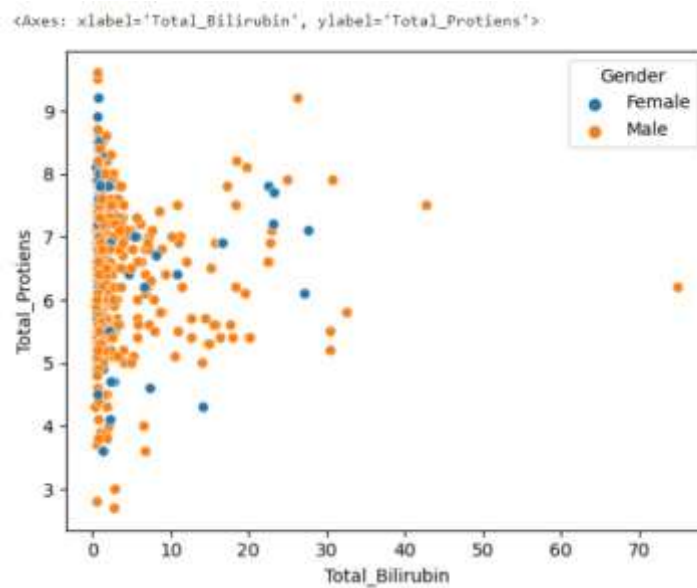
<Axes: xlabel='Total_Protiens', ylabel='Density'>



Bivariate Analysis



Multivariate Analysis



	<p><Axes: ></p> <p>Heatmap showing correlations between variables: Age, Total_Bilirubin, Direct_Bilirubin, Total_Protiens, Albumin, Albumin_and_Globulin_Ratio, and Dataset. The color scale ranges from -0.2 (dark purple) to 1.0 (yellow).</p>
Outliers and Anomalies	<pre>sns.boxplot(data.Albumin_and_Globulin_Ratio,orient='h')</pre> <p><Axes: ></p> <p>Horizontal boxplot for Albumin_and_Globulin_Ratio showing a distribution with a median around 0.9 and several outliers extending to the right.</p>
Data Preprocessing Code Screenshots	
Loading Data	<pre># Loading the dataset data = pd.read_csv("indian_liver_patient.csv")</pre>

	<table><tr><th></th><th>Age</th><th>Gender</th><th>Total_Bilirubin</th><th>Direct_Bilirubin</th><th>Alkaline_Phosphatase</th><th>Alanine_Aminotransferase</th><th>Aspartate_Aminotransferase</th><th>Total_Proteins</th><th>Albumin</th><th>Albumin_and_Globulin_Ratio</th></tr><tr><td>0</td><td>81</td><td>Female</td><td>0.7</td><td>0.1</td><td>187</td><td>56</td><td>56</td><td>68</td><td>5.0</td><td>0.0</td></tr><tr><td>1</td><td>82</td><td>Male</td><td>33.9</td><td>9.1</td><td>889</td><td>84</td><td>190</td><td>73</td><td>5.0</td><td>0.0</td></tr><tr><td>2</td><td>82</td><td>Male</td><td>9.2</td><td>4.1</td><td>490</td><td>90</td><td>89</td><td>72</td><td>5.0</td><td>0.0</td></tr><tr><td>3</td><td>58</td><td>Male</td><td>5.0</td><td>0.4</td><td>192</td><td>54</td><td>20</td><td>68</td><td>2.4</td><td>0.0</td></tr><tr><td>4</td><td>72</td><td>Male</td><td>0.9</td><td>0.0</td><td>109</td><td>27</td><td>38</td><td>73</td><td>2.4</td><td>0.0</td></tr></table>		Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	Alanine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	0	81	Female	0.7	0.1	187	56	56	68	5.0	0.0	1	82	Male	33.9	9.1	889	84	190	73	5.0	0.0	2	82	Male	9.2	4.1	490	90	89	72	5.0	0.0	3	58	Male	5.0	0.4	192	54	20	68	2.4	0.0	4	72	Male	0.9	0.0	109	27	38	73	2.4	0.0
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	Alanine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio																																																									
0	81	Female	0.7	0.1	187	56	56	68	5.0	0.0																																																									
1	82	Male	33.9	9.1	889	84	190	73	5.0	0.0																																																									
2	82	Male	9.2	4.1	490	90	89	72	5.0	0.0																																																									
3	58	Male	5.0	0.4	192	54	20	68	2.4	0.0																																																									
4	72	Male	0.9	0.0	109	27	38	73	2.4	0.0																																																									
Handling Missing Data	<pre>data.isnull().sum()</pre> <table><tr><td>Age</td><td>0</td></tr><tr><td>Gender</td><td>0</td></tr><tr><td>Total_Bilirubin</td><td>0</td></tr><tr><td>Direct_Bilirubin</td><td>0</td></tr><tr><td>Alkaline_Phosphatase</td><td>0</td></tr><tr><td>Alanine_Aminotransferase</td><td>0</td></tr><tr><td>Aspartate_Aminotransferase</td><td>0</td></tr><tr><td>Total_Proteins</td><td>0</td></tr><tr><td>Albumin</td><td>0</td></tr><tr><td>Albumin_and_Globulin_Ratio</td><td>4</td></tr><tr><td>Dataset</td><td>0</td></tr></table> <pre>dtype: int64</pre> <pre>data['Albumin_and_Globulin_Ratio'].fillna(data['Albumin_and_Globulin_Ratio'].mode()[0],inplace=True)</pre> <pre>data.isna().sum()</pre> <table><tr><td>Age</td><td>0</td></tr><tr><td>Gender</td><td>0</td></tr><tr><td>Total_Bilirubin</td><td>0</td></tr><tr><td>Direct_Bilirubin</td><td>0</td></tr><tr><td>Alkaline_Phosphatase</td><td>0</td></tr><tr><td>Alanine_Aminotransferase</td><td>0</td></tr><tr><td>Aspartate_Aminotransferase</td><td>0</td></tr><tr><td>Total_Proteins</td><td>0</td></tr><tr><td>Albumin</td><td>0</td></tr><tr><td>Albumin_and_Globulin_Ratio</td><td>0</td></tr><tr><td>Dataset</td><td>0</td></tr></table> <pre>dtype: int64</pre>	Age	0	Gender	0	Total_Bilirubin	0	Direct_Bilirubin	0	Alkaline_Phosphatase	0	Alanine_Aminotransferase	0	Aspartate_Aminotransferase	0	Total_Proteins	0	Albumin	0	Albumin_and_Globulin_Ratio	4	Dataset	0	Age	0	Gender	0	Total_Bilirubin	0	Direct_Bilirubin	0	Alkaline_Phosphatase	0	Alanine_Aminotransferase	0	Aspartate_Aminotransferase	0	Total_Proteins	0	Albumin	0	Albumin_and_Globulin_Ratio	0	Dataset	0																						
Age	0																																																																		
Gender	0																																																																		
Total_Bilirubin	0																																																																		
Direct_Bilirubin	0																																																																		
Alkaline_Phosphatase	0																																																																		
Alanine_Aminotransferase	0																																																																		
Aspartate_Aminotransferase	0																																																																		
Total_Proteins	0																																																																		
Albumin	0																																																																		
Albumin_and_Globulin_Ratio	4																																																																		
Dataset	0																																																																		
Age	0																																																																		
Gender	0																																																																		
Total_Bilirubin	0																																																																		
Direct_Bilirubin	0																																																																		
Alkaline_Phosphatase	0																																																																		
Alanine_Aminotransferase	0																																																																		
Aspartate_Aminotransferase	0																																																																		
Total_Proteins	0																																																																		
Albumin	0																																																																		
Albumin_and_Globulin_Ratio	0																																																																		
Dataset	0																																																																		
Data Transformation	<pre>from sklearn.preprocessing import StandardScaler</pre> <pre>sc=StandardScaler()</pre> <pre>X=sc.fit_transform(X)</pre> <pre>X</pre> <pre>array([[1.25289764, -1.76228085, 0.41887783, ..., 0.29211961, 0.19896867, -0.14789798], [1.06663784, 0.56744644, 1.22517135, ..., 0.93756634, 0.07315659, -0.65069686], [1.06663784, 0.56744644, 0.6449187 , ..., 0.47653296, 0.19896867, -0.17932291], ..., [0.44843504, 0.56744644, -0.4027597 , ..., -0.8767071 , 0.07315659, 0.16635131], [-0.84978917, 0.56744644, -0.32216906, ..., 0.29211961, 0.32478075, 0.16635131], [-0.41704777, 0.56744644, -0.37052344, ..., 0.75315299, 1.58290153, 1.73759779]])</pre>																																																																		
Feature Engineering	<pre>from sklearn.preprocessing import LabelEncoder</pre> <pre>le=LabelEncoder()</pre> <pre>X['Gender']=le.fit_transform(X['Gender'])</pre> <pre>X['Gender']</pre> <pre>0 0 1 1 2 1 3 1 4 1 ... 578 1 579 1 580 1 581 1 582 1 Name: Gender, Length: 583, dtype: int32</pre>																																																																		

Save Processed Data

```
import pickle
pickle.dump(svm , open('model.pkl','wb'))
pickle.dump(sc , open('sc.pkl','wb'))
```