

## Data Collection and Preprocessing Phase

Date	13 july 2024
Team ID	739805
Project Title	Prediction and Analysis of Liver Patient Data Using Machine Learning
Maximum Marks	6 Marks

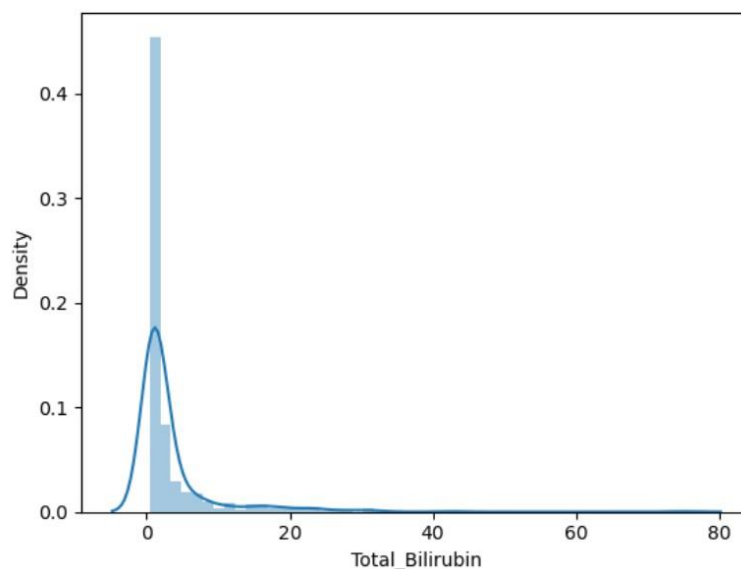
### Data Exploration and Preprocessing Template

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

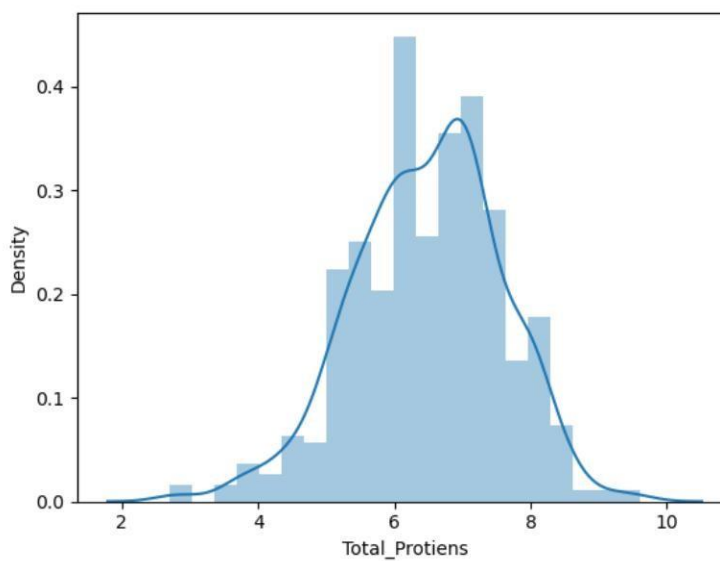
Section	Description
Data Overview	583 rows × 11 columns

## Univariate Analysis

<Axes: xlabel='Total\_Bilirubin', ylabel='Density'>

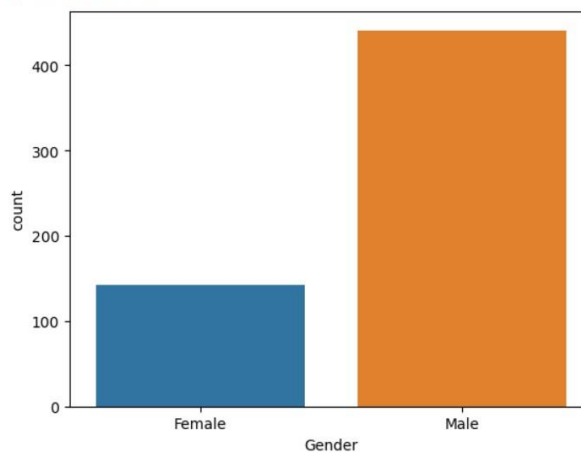


<Axes: xlabel='Total\_Protiens', ylabel='Density'>

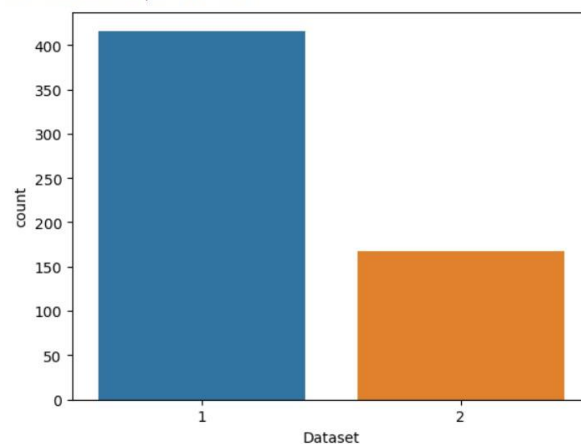


## Bivariate Analysis

No. of Males: 441  
 No. of Females: 142

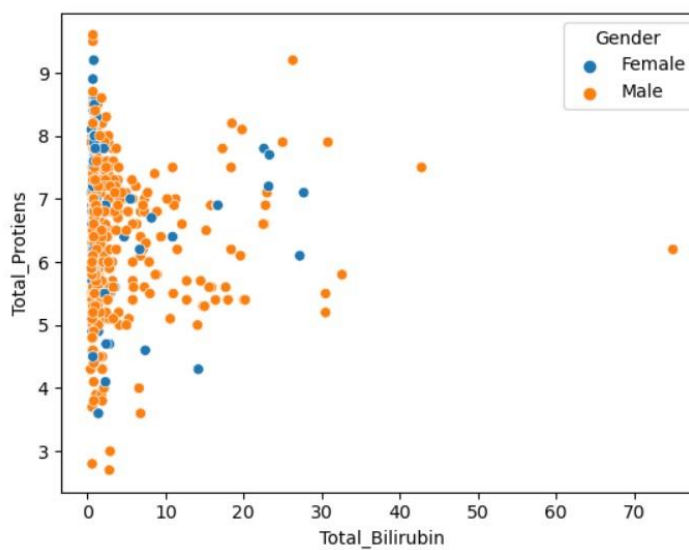


Liver disease patients: 416  
 Non-Liver disease patients: 167



## Multivariate Analysis

<Axes: xlabel='Total\_Bilirubin', ylabel='Total\_Protiens'>



	<div><div>&lt;Axes: &gt;</div><div><table><tr><th>Age</th><td>1</td><td>0.012</td><td>0.0075</td><td>-0.19</td><td>-0.27</td><td>-0.22</td><td>-0.14</td></tr><tr><th>Total_Bilirubin</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Direct_Bilirubin</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Total_Protiens</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Albumin</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Albumin_and_Globulin_Ratio</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Dataset</th><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></div></div>	Age	1	0.012	0.0075	-0.19	-0.27	-0.22	-0.14	Total_Bilirubin								Direct_Bilirubin								Total_Protiens								Albumin								Albumin_and_Globulin_Ratio								Dataset							
Age	1	0.012	0.0075	-0.19	-0.27	-0.22	-0.14																																																		
Total_Bilirubin																																																									
Direct_Bilirubin																																																									
Total_Protiens																																																									
Albumin																																																									
Albumin_and_Globulin_Ratio																																																									
Dataset																																																									
Outliers and Anomalies	<div><div><pre>sns.boxplot(data.Albumin_and_Globulin_Ratio,orient='h')</pre></div><div><div>&lt;Axes: &gt;</div><div></div></div></div>																																																								
Data Preprocessing Code Screenshots																																																									
Loading Data	<div><pre># Loading the dataset data = pd.read_csv("indian_liver_patient.csv")</pre></div>																																																								

	<table><tr><th></th><th>Age</th><th>Gender</th><th>Total_Bilirubin</th><th>Direct_Bilirubin</th><th>Alkaline_Phosphotase</th><th>Alamine_Aminotransferase</th><th>Aspartate_Aminotransferase</th><th>Total_Protiens</th><th>Albumin</th><th>Albumin_and_G</th></tr><tr><td>0</td><td>65</td><td>Female</td><td>0.7</td><td>0.1</td><td>187</td><td>16</td><td>18</td><td>6.8</td><td>3.3</td><td></td></tr><tr><td>1</td><td>62</td><td>Male</td><td>10.9</td><td>5.5</td><td>699</td><td>64</td><td>100</td><td>7.5</td><td>3.2</td><td></td></tr><tr><td>2</td><td>62</td><td>Male</td><td>7.3</td><td>4.1</td><td>490</td><td>60</td><td>68</td><td>7.0</td><td>3.3</td><td></td></tr><tr><td>3</td><td>58</td><td>Male</td><td>1.0</td><td>0.4</td><td>182</td><td>14</td><td>20</td><td>6.8</td><td>3.4</td><td></td></tr><tr><td>4</td><td>72</td><td>Male</td><td>3.9</td><td>2.0</td><td>195</td><td>27</td><td>59</td><td>7.3</td><td>2.4</td><td></td></tr></table>		Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_G	0	65	Female	0.7	0.1	187	16	18	6.8	3.3		1	62	Male	10.9	5.5	699	64	100	7.5	3.2		2	62	Male	7.3	4.1	490	60	68	7.0	3.3		3	58	Male	1.0	0.4	182	14	20	6.8	3.4		4	72	Male	3.9	2.0	195	27	59	7.3	2.4	
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_G																																																									
0	65	Female	0.7	0.1	187	16	18	6.8	3.3																																																										
1	62	Male	10.9	5.5	699	64	100	7.5	3.2																																																										
2	62	Male	7.3	4.1	490	60	68	7.0	3.3																																																										
3	58	Male	1.0	0.4	182	14	20	6.8	3.4																																																										
4	72	Male	3.9	2.0	195	27	59	7.3	2.4																																																										
Handling Missing Data	<pre>data.isnull().sum()  Age                0 Gender             0 Total_Bilirubin    0 Direct_Bilirubin   0 Alkaline_Phosphotase 0 Alamine_Aminotransferase 0 Aspartate_Aminotransferase 0 Total_Protiens     0 Albumin            0 Albumin_and_Globulin_Ratio 4 Dataset            0 dtype: int64  data['Albumin_and_Globulin_Ratio'].fillna(data['Albumin_and_Globulin_Ratio'].mode()[0],inplace=True)  data.isna().sum()  Age                0 Gender             0 Total_Bilirubin    0 Direct_Bilirubin   0 Alkaline_Phosphotase 0 Alamine_Aminotransferase 0 Aspartate_Aminotransferase 0 Total_Protiens     0 Albumin            0 Albumin_and_Globulin_Ratio 0 Dataset            0 dtype: int64</pre>																																																																		
Data Transformation	<pre>from sklearn.preprocessing import StandardScaler  sc=StandardScaler()  x=sc.fit_transform(x)  x  array([[ 1.25209764, -1.76228085, -0.41887783, ...,  0.29211961,          0.19896867, -0.14789798],        [ 1.06663704,  0.56744644,  1.22517135, ...,  0.93756634,          0.07315659, -0.65069686],        [ 1.06663704,  0.56744644,  0.6449187 , ...,  0.47653296,          0.19896867, -0.17932291],        ...,        [ 0.44843504,  0.56744644, -0.4027597 , ..., -0.0767071 ,          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  le=LabelEncoder()  x['Gender']=le.fit_transform(x['Gender'])  x['Gender']  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'))
```