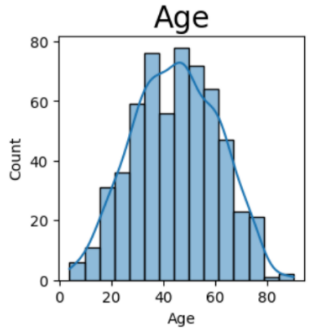
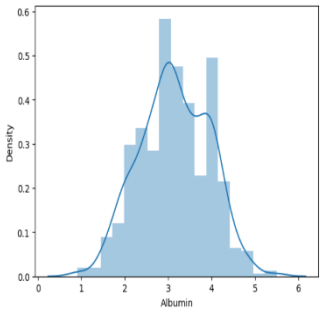
Data Collection and Preprocessing Phase

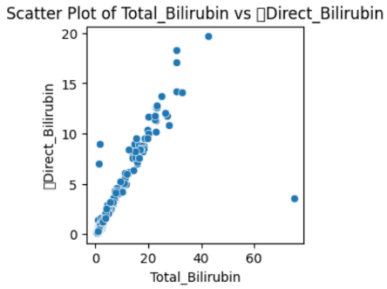
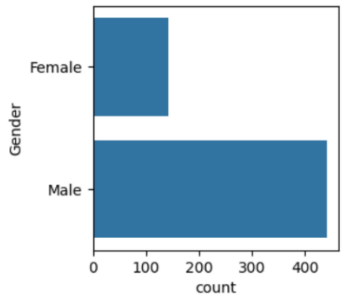
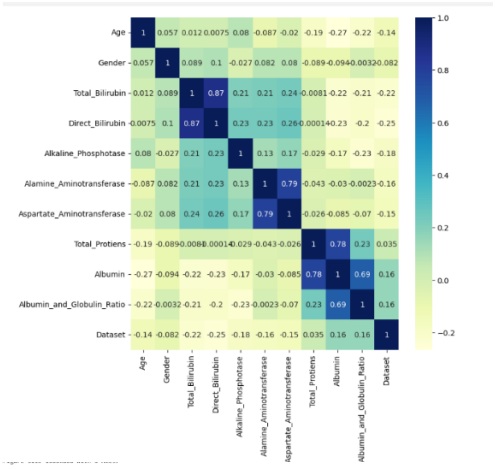
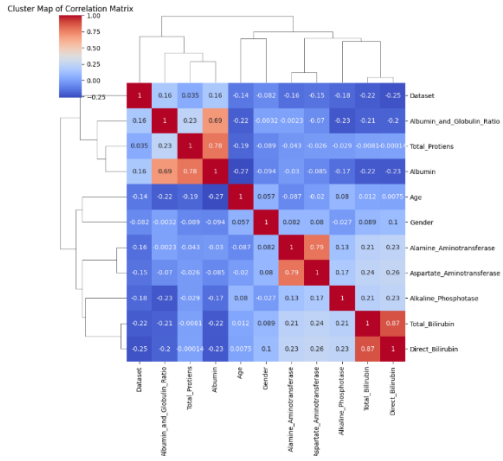
Date	03-10-2024
Team ID	LTVIP2024TMID24892
Project Title	Liver Patient Identification – prediction of liver patient
Maximum Marks	6 Marks

Data Exploration and Preprocessing :

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.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:
584rows X 12columns
Descriptive statistics:
<table>
<tr><th></th><th>Age</th><th>Total_Bilirubin</th><th>Direct_Bilirubin</th><th>Alkaline_Phosphotase</th><th>Alamine_Aminotransferase</th><th>Aspartate_Aminotransferase</th></tr>
<tr><td>count</td><td>583.000000</td><td>583.000000</td><td>583.000000</td><td>583.000000</td><td>583.000000</td><td>583.000000</td></tr>
<tr><td>mean</td><td>44.746141</td><td>3.298799</td><td>1.486106</td><td>290.576329</td><td>80.713551</td><td>109.910806</td></tr>
<tr><td>std</td><td>16.189833</td><td>6.209522</td><td>2.808498</td><td>242.937989</td><td>182.620356</td><td>288.918529</td></tr>
<tr><td>min</td><td>4.000000</td><td>0.400000</td><td>0.100000</td><td>63.000000</td><td>10.000000</td><td>10.000000</td></tr>
<tr><td>25%</td><td>33.000000</td><td>0.800000</td><td>0.200000</td><td>175.500000</td><td>23.000000</td><td>25.000000</td></tr>
<tr><td>50%</td><td>45.000000</td><td>1.000000</td><td>0.300000</td><td>208.000000</td><td>35.000000</td><td>42.000000</td></tr>
<tr><td>75%</td><td>58.000000</td><td>2.600000</td><td>1.300000</td><td>298.000000</td><td>60.500000</td><td>87.000000</td></tr>
<tr><td>max</td><td>90.000000</td><td>75.000000</td><td>19.700000</td><td>2110.000000</td><td>2000.000000</td><td>4929.000000</td></tr>
</table></td></tr>
<tr><td>Univariate Analysis</td><td> 


</td></tr>
</table>

Bivariate Analysis	 
Multivariate Analysis	 

Outliers and Anomalies	```
	#Calculate the boundaries of Total_Protiens feature which differentiates the outliers:
	uppper_boundary=df['Total_Protiens'].mean() + 3* df['Total_Protiens'].std()
	lower_boundary=df['Total_Protiens'].mean() - 3* df['Total_Protiens'].std()
	
	print(df['Total_Protiens'].mean())
	print(lower_boundary)
	print(uppper_boundary)
	``
	``
	6.483190394511149
	3.22683594244075
	9.739544846581548
	``

Data Preprocessing Code Screenshots

Loading Data

```
# Reading Dataset:
df = pd.read_csv("/content/Liver_data.csv")
# Top 5 records:
df.head()
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
0	65	Female	0.7	0.1	187	16	18
1	62	Male	10.9	5.5	699	64	100
2	62	Male	7.3	4.1	490	60	68
3	58	Male	1.0	0.4	182	14	20
4	72	Male	3.9	2.0	195	27	59

Handling Missing Data

```
# Mean & Median of "Albumin_and_Globulin_Ratio" feature:
print(df['Albumin_and_Globulin_Ratio'].median())
print(df['Albumin_and_Globulin_Ratio'].mean())
```

```
0.93
0.9470639032815197
```

```
# Filling NaN Values of "Albumin_and_Globulin_Ratio" feature with Median :
df['Albumin_and_Globulin_Ratio'] = df['Albumin_and_Globulin_Ratio'].fillna(df['Albumin_and_Globulin_Ratio'].median())
```

Data Transformation

There is no need of Standardization and Normalization of our dataset, as we using Ensemble Technique.

Feature Engineering

```
# SMOTE Technique:
from imblearn.combine import SMOTETomek
num_bins = 3
y = pd.cut(y, bins=num_bins, labels=False)

smote = SMOTETomek()
X_smote, y_smote = smote.fit_resample(X, y)
```

```
# Counting before and after SMOTE:
from collections import Counter
print('Before SMOTE : ', Counter(y))
print('After SMOTE  : ', Counter(y_smote))
```

```
Before SMOTE :  Counter({0: 416, 2: 167})
After SMOTE  :  Counter({0: 394, 2: 394})
```

Save Processed Data

-