Team ID	PNT2022TMID41146
Project Name	Project - Statistical Machine Learning Approaches to Liver Disease Prediction.

SPRINT 2

Data Collection and Preprocessing

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pickle
from sklearn.model_selection import train_test_split, StratifiedKFold, GridSearchCV
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, confusion_matrix,classification_report
```

Reading Dataset

```
data=pd.read_csv('/content/indian_liver_patient.csv')
```

Data visualization

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1636 entries, 0 to 1635
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Age                         1636 non-null   int64  
 1   Gender                      1636 non-null   object 
 2   Total_Bilirubin             1636 non-null   float64
 3   Direct_Bilirubin            1636 non-null   float64
 4   Alkaline_Phosphotase        1636 non-null   int64  
 5   Alamine_Aminotransferase    1636 non-null   int64  
 6   Aspartate_Aminotransferase  1636 non-null   int64  
 7   Total_Protiens              1636 non-null   float64
 8   Albumin                     1636 non-null   float64
 9   Albumin_and_Globulin_Ratio  1624 non-null   float64
 10  Dataset                     1636 non-null   int64  
dtypes: float64(5), int64(5), object(1)
memory usage: 140.7+ KB
```

```
[5] data.head(10)
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Ami
0	65	Female	0.7	0.1	187	16	
1	62	Male	10.9	5.5	699	64	
2	62	Male	7.3	4.1	490	60	
3	58	Male	1.0	0.4	182	14	
4	72	Male	3.9	2.0	195	27	
5	46	Male	1.8	0.7	208	19	
6	26	Female	0.9	0.2	154	16	
7	29	Female	0.9	0.3	202	14	
8	17	Male	0.9	0.3	202	22	
9	55	Male	0.7	0.2	290	53	

```
[7] data.tail(10)
```

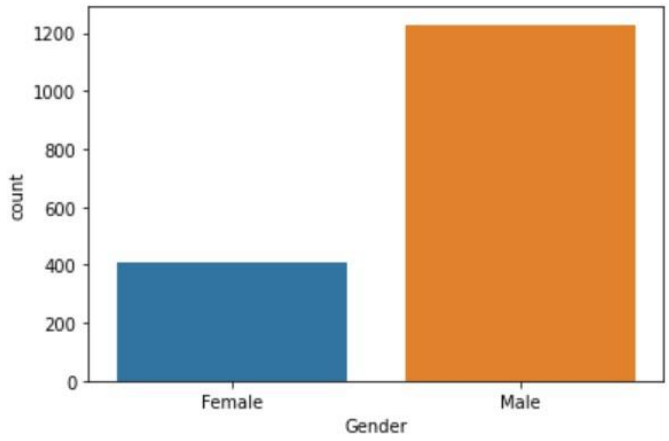
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_
1626	22	Female	2.2	1.0	215	159	
1627	28	Female	0.8	0.2	309	55	
1628	38	Male	0.7	0.2	110	22	
1629	25	Male	0.8	0.1	130	23	
1630	45	Female	0.7	0.2	164	21	
1631	45	Female	0.6	0.1	270	23	
1632	28	Female	0.6	0.1	137	22	
1633	28	Female	1.0	0.3	90	18	
1634	66	Male	1.0	0.3	190	30	
1635	66	Male	0.8	0.2	165	22	

```
[8] data.describe()
```

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate
count	1636.000000	1636.000000	1636.000000	1636.000000	1636.000000	
mean	44.727995	3.114792	1.387286	293.103912	80.944377	
std	16.295775	5.955451	2.631630	248.412910	186.409237	
min	4.000000	0.400000	0.100000	63.000000	10.000000	
25%	33.000000	0.800000	0.200000	175.000000	23.000000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	
75%	58.000000	2.400000	1.200000	298.000000	60.000000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	

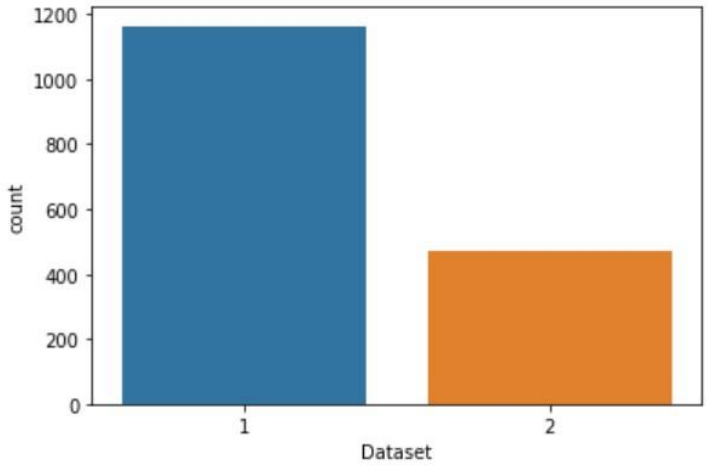
```
[13] sns.countplot(data=data,x='Gender',label='Count')
     m,f=data['Gender'].value_counts()
     print("No of Males:",m)
     print("no of Females:",f)
```

```
No of Males: 1229
no of Females: 407
```



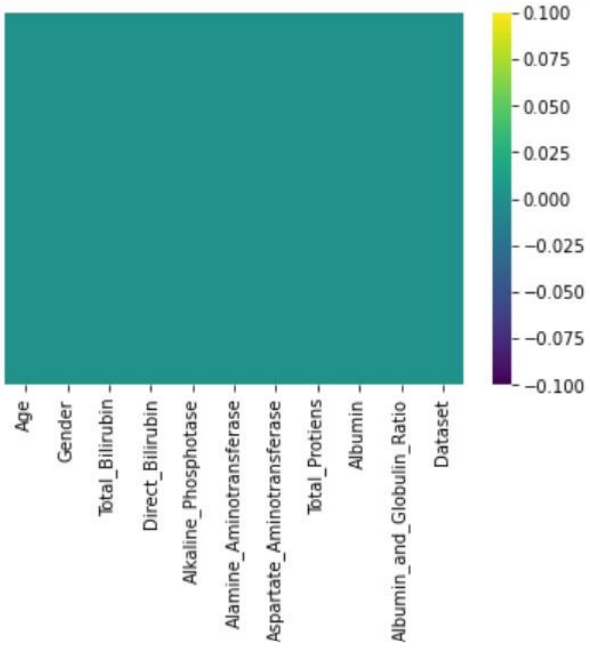

```
[14] sns.countplot(data=data,x='Dataset')
     LD,NLD=data['Dataset'].value_counts()
     print("liver disease patients:",LD)
     print("non-liver disease patients:",NLD)
```

```
liver disease patients: 1164
non-liver disease patients: 472
```



```
[16] sns.heatmap(data.isnull(),yticklabels=False,cmap='viridis')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f504195ba50>
```



Checking for Null values and handling the Null values

```
[9] data.isnull().any()
```

```
Age                          False
Gender                       False
Total_Bilirubin              False
Direct_Bilirubin             False
Alkaline_Phosphotase         False
Alamine_Aminotransferase     False
Aspartate_Aminotransferase   False
Total_Protiens               False
Albumin                      False
Albumin_and_Globulin_Ratio    True
Dataset                      False
dtype: bool
```

```
[10] 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  12
Dataset                      0
dtype: int64
```

```
[11] data['Albumin_and_Globulin_Ratio']=data['Albumin_and_Globulin_Ratio'].fillna(data['Albumin_and_Globulin_Ratio'].mode()[0])
```

```
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  0
Dataset                     0
dtype: int64
```

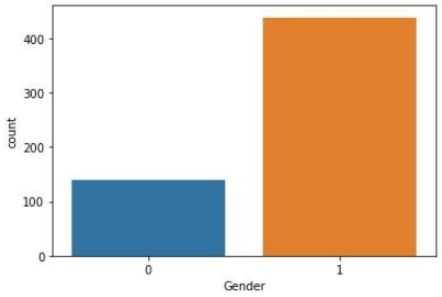
EDA : Exploratory Data Analysis Uni-variate Analysis

```
df.head()
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

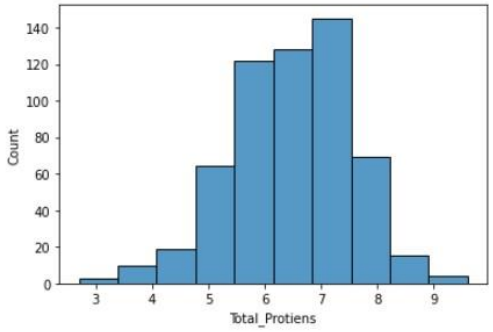
```
sns.countplot(x='Gender',data=df, dodge=True)
```

```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```



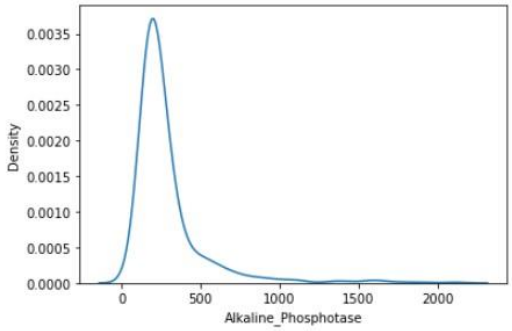
```
sns.histplot(x='Total_Protiens',data=df,bins=10)
```

```
<AxesSubplot:xlabel='Total_Protiens', ylabel='Count'>
```




```
sns.kdeplot(x='Alkaline_Phosphotase', data=df)
```

```
<AxesSubplot:xlabel='Alkaline_Phosphotase', ylabel='Density'>
```




```
sns.boxplot(x='Albumin_and_Globulin_Ratio',data=df)
```

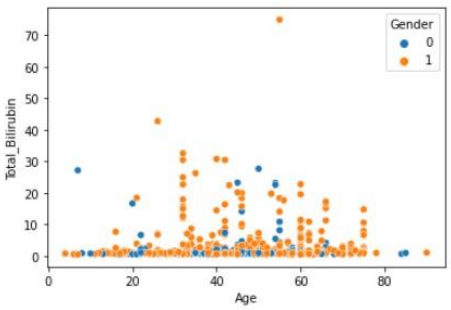
Bi-variate Analysis :

```
df.head()
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

```
sns.scatterplot(x='Age',y='Total_Bilirubin',data=df,hue='Gender')
```

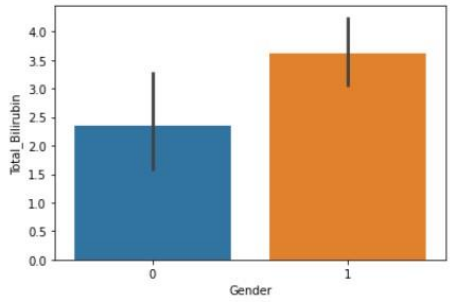
```
<AxesSubplot:xlabel='Age', ylabel='Total_Bilirubin'>
```




```
sns.barplot(x='Gender',y='Total_Bilirubin',data=df)
```

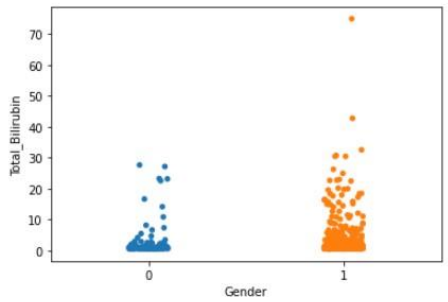
```
sns.barplot(x='Gender',y='Total_Bilirubin',data=df)
```

```
<AxesSubplot:xlabel='Gender', ylabel='Total_Bilirubin'>
```




```
sns.stripplot(x='Gender',y='Total_Bilirubin',data=df)
```

```
<AxesSubplot:xlabel='Gender', ylabel='Total_Bilirubin'>
```



Multi – variate Analysis :

```
sns.pairplot(data=df,hue='Gender')
```

```
<seaborn.axisgrid.PairGrid at 0x22f6cbfda90>
```