

Team ID PNT2022TMID52707
Project Name 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
```

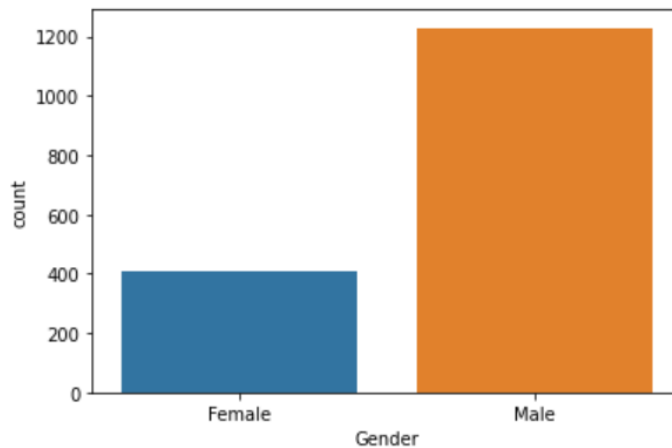

0s

0s



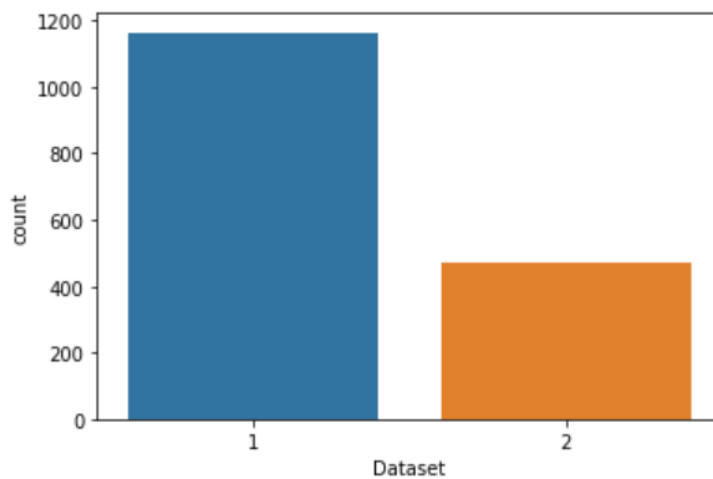
```
✓ [13] sns.countplot(data=data,x='Gender',label='Count')
1s      m,f=data['Gender'].value_counts()
      print("No of Males:",m)
      print("no of Females:",f)
```

No of Males: 1229
no of Females: 407



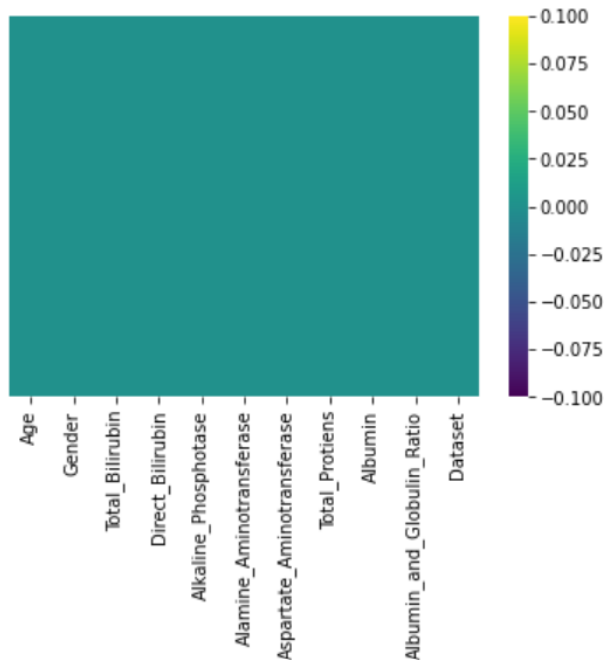
```
✓ [14] sns.countplot(data=data,x='Dataset')
1s      LD,NLD=data['Dataset'].value_counts()
      print("liver disease patients:",LD)
      print("non-liver disease patients:",NLD)
```

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
```

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

0s



```
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 AnalysisUni – variate Analysis:

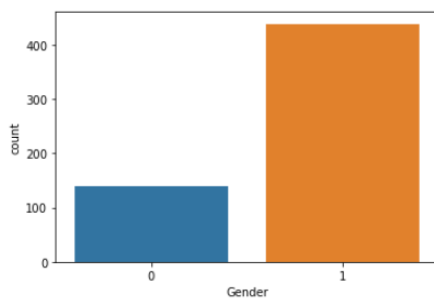
```
4]: df.head()
```

```
4]:
```

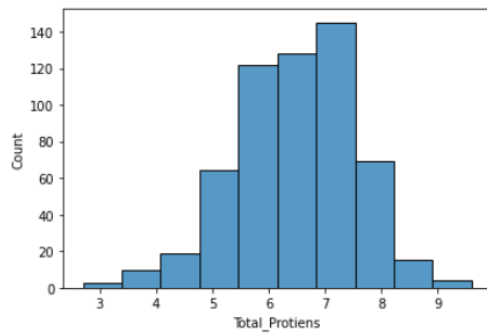
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195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

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

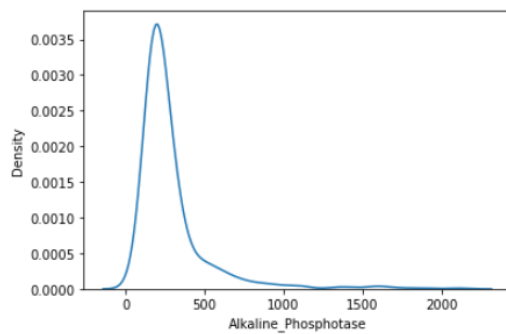
```
5]: <AxesSubplot:xlabel='Gender', ylabel='count'>
```



```
: 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)
```

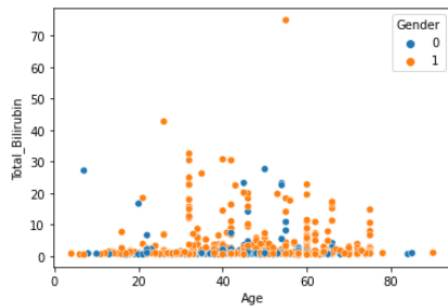
Bi – variate Analysis:

```
: 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')
```

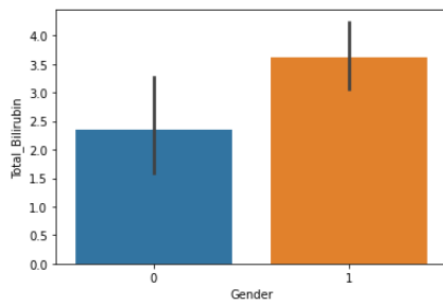
```
: <AxesSubplot:xlabel='Age', ylabel='Total_Bilirubin'>
```



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

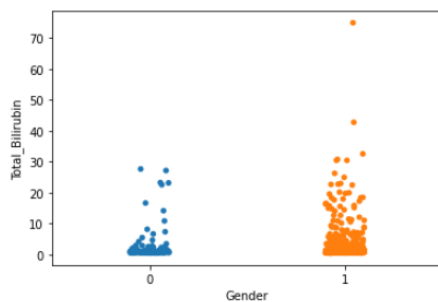
```
: sns.barplot(x='Gender',y='Total_Bilirubin',data=df)
```

```
: <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>
```



