Project Development Phase
SPRINT 1

Date	29.10.2022
Team ID	PNT2022TMID17773
Project Name	Project -Statistical Machine Learning Approaches To Liver Disease Prediction

Executable Program

```
import numpy as np
# for dataframes
import pandas as pd
# for easier visualization
import seaborn as sns
# for visualization and to display plots
from matplotlib import pyplot as plt
# import color maps
from matplotlib.colors import ListedColormap
```

```
df=pd.read_csv('indian_liver_patient.csv')
df
```

Out[113]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albu
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	
...	...	...	...	...	...	...	...	...	...	...
578	60	Male	0.5	0.1	500	20	34	5.9	1.6	
579	40	Male	0.6	0.1	98	35	31	6.0	3.2	
580	52	Male	0.8	0.2	245	48	49	6.4	3.2	
581	31	Male	1.3	0.5	184	29	32	6.8	3.4	
582	38	Male	1.0	0.3	216	21	24	7.3	4.4	

583 rows × 11 columns

```
df.shape
```

```
(583, 11)
```

```
df.columns
```

```
Index(['Age', 'Gender', 'Total_Bilirubin', 'Direct_Bilirubin',
       'Alkaline_Phosphotase', 'Alamine_Aminotransferase',
       'Aspartate_Aminotransferase', 'Total_Protiens', 'Albumin',
       'Albumin_and_Globulin_Ratio', 'Dataset'],
      dtype='object')
```

```
df.head()
```

In [116]:
```
df.head()
```

Out[116]:

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

Exploratory analysis

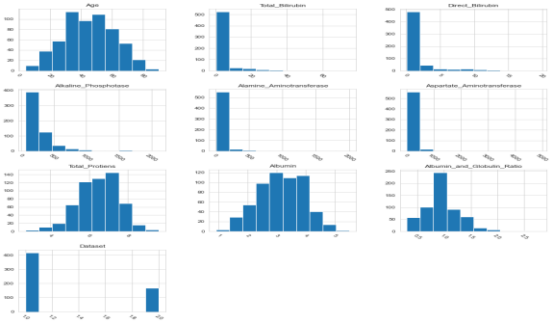
filtering categorical data

df.dtypes[df.dtypes=='object']

Distribution of Numerical Features

df.hist(figsize=(15,15), xrot=-45, bins=10)

plt.show()



df.describe()

```
In [119]: df.describe()
```

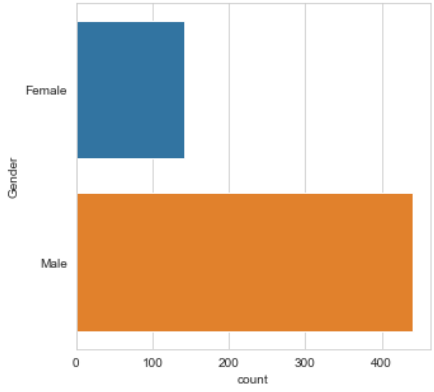
Out[119]:

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Al
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	

Dataset i.e output value has '1' for liver disease and '2' for no liver disease so let's make it 0 for no disease to make it convineint

```
def partition(x):
    if x == 2:
        return 0
    return 1

df['Dataset'] = df['Dataset'].map(partition)
plt.figure(figsize=(5,5))
sns.countplot(y='Gender', data=df)
```



sns.countplot(data=df, x = 'Gender', label='Count')

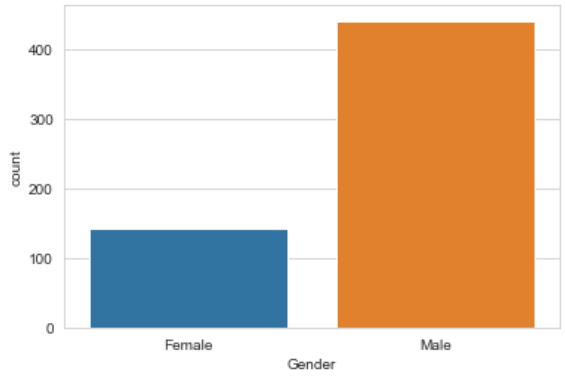
M, F = df['Gender'].value_counts()

print('Number of patients that are male: ',M)

print('Number of patients that are female: ',F)

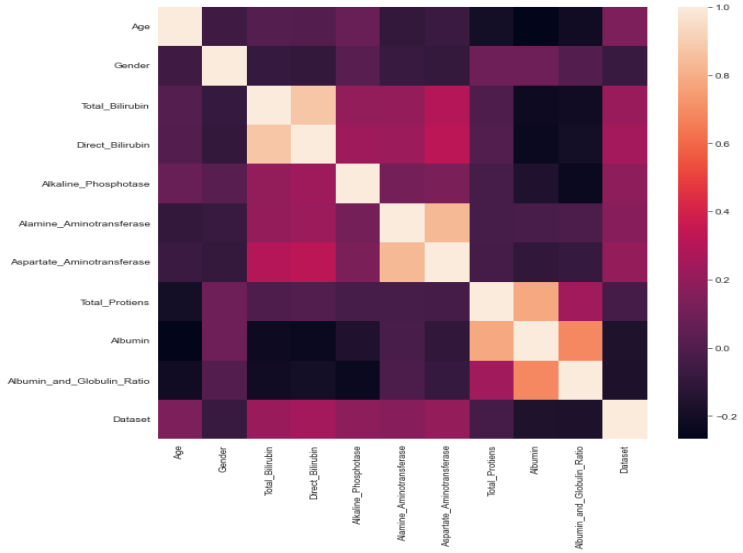
```
Number of patients that are male:  441
```

```
Number of patients that are female:  142
```



Label Male as 0 and Female as 1

```
def partition(x):
   if x =='Male':
      return 0
   return 1
df['Gender'] = df['Gender'].map(partition)
df.corr()
plt.figure(figsize=(10,10))
sns.heatmap(df.corr())
```



Data cleaning

```
df = df.drop_duplicates()
print( df.shape )
```

```
(564, 11)
```

Removing outliers

```
df = df[df.Aspartate_Aminotransferase <=3000 ]
df.shape
```

```
(564, 11)
```

```
df = df[df.Aspartate_Aminotransferase <=2500 ]
df.shape
```

```
(564, 11)
```

Dropping null values

```
df=df.dropna(how='any')
df.shape
```

```
(564, 11)
```