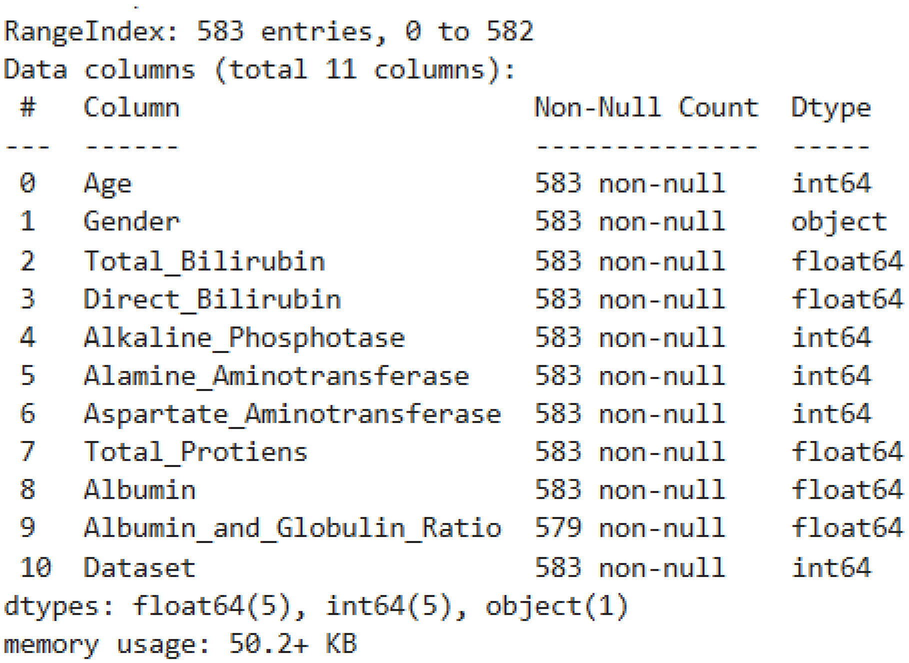
# Data Collection and Preprocessing Phase

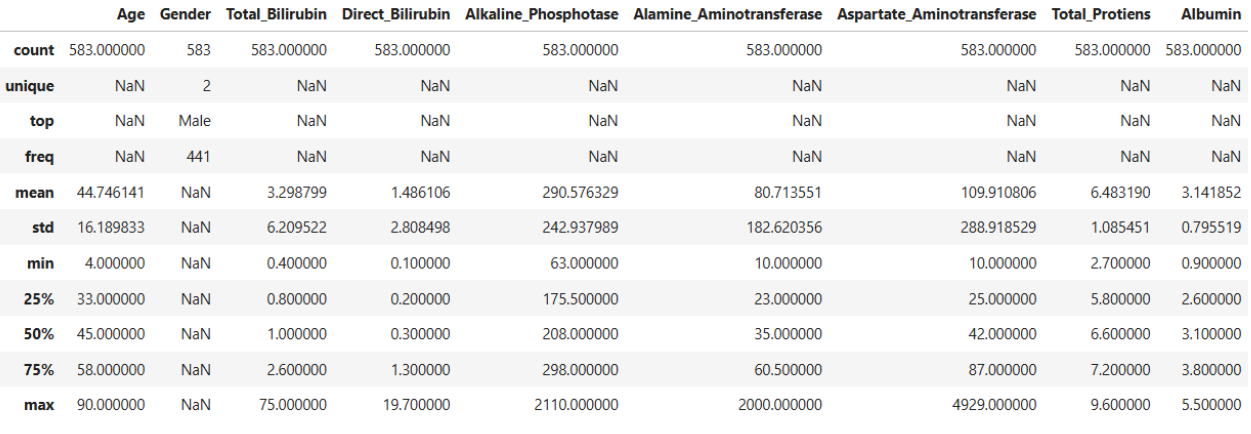
| | |
|---|---|
| Date | 6th june 2024 |
| Team ID | SWTID1720175375 |
| Project Title | Prediction and analysis of liver patient data using ML |
| Maximum Marks | 6 Marks |

Data Overview



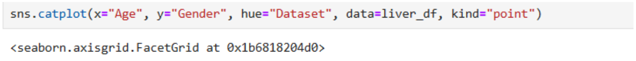
```
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Age                         583 non-null    int64  
 1   Gender                      583 non-null    object 
 2   Total_Bilirubin             583 non-null    float64
 3   Direct_Bilirubin            583 non-null    float64
 4   Alkaline_Phosphotase        583 non-null    int64  
 5   Alamine_Aminotransferase    583 non-null    int64  
 6   Aspartate_Aminotransferase  583 non-null    int64  
 7   Total_Protiens              583 non-null    float64
 8   Albumin                     583 non-null    float64
 9   Albumin_and_Globulin_Ratio  579 non-null    float64
 10  Dataset                     583 non-null    int64  
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ KB
```

Univariate Analysis:



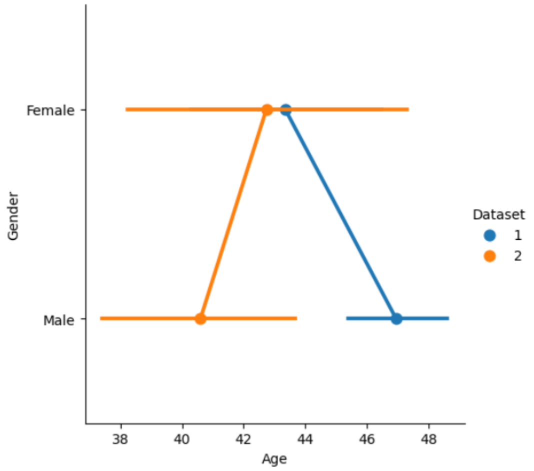
|  | Age | Gender | Total_Bilirubin | Direct_Bilirubin | Alkaline_Phosphotase | Alamine_Aminotransferase | Aspartate_Aminotransferase | Total_Protiens | Albumin |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 583.000000 | 583 | 583.000000 | 583.000000 | 583.000000 | 583.000000 | 583.000000 | 583.000000 | 583.000000 |
| **unique** | NaN | 2 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **top** | NaN | Male | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **freq** | NaN | 441 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **mean** | 44.746141 | NaN | 3.298799 | 1.486106 | 290.576329 | 80.713551 | 109.910806 | 6.483190 | 3.141852 |
| **std** | 16.189833 | NaN | 6.209522 | 2.808498 | 242.937989 | 182.620356 | 288.918529 | 1.085451 | 0.795519 |
| **min** | 4.000000 | NaN | 0.400000 | 0.100000 | 63.000000 | 10.000000 | 10.000000 | 2.700000 | 0.900000 |
| **25%** | 33.000000 | NaN | 0.800000 | 0.200000 | 175.500000 | 23.000000 | 25.000000 | 5.800000 | 2.600000 |
| **50%** | 45.000000 | NaN | 1.000000 | 0.300000 | 208.000000 | 35.000000 | 42.000000 | 6.600000 | 3.100000 |
| **75%** | 58.000000 | NaN | 2.600000 | 1.300000 | 298.000000 | 60.500000 | 87.000000 | 7.200000 | 3.800000 |
| **max** | 90.000000 | NaN | 75.000000 | 19.700000 | 2110.000000 | 2000.000000 | 4929.000000 | 9.600000 | 5.500000 |

## Bivariate Analysis:

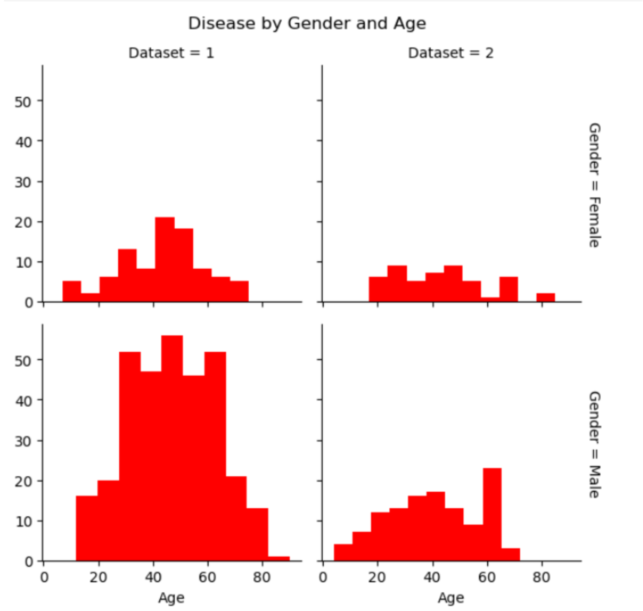


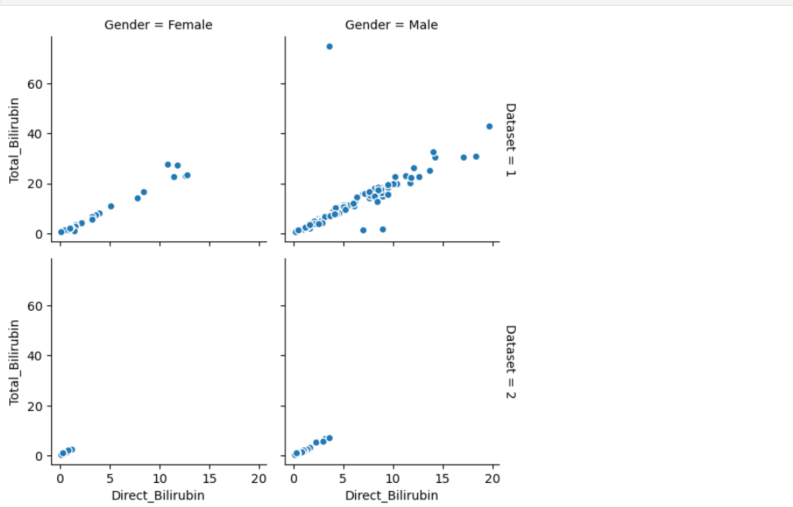
```
sns.catplot(x="Age", y="Gender", hue="Dataset", data=liver_df, kind="point")
```

```
<seaborn.axisgrid.FacetGrid at 0x1b6818204d0>
```

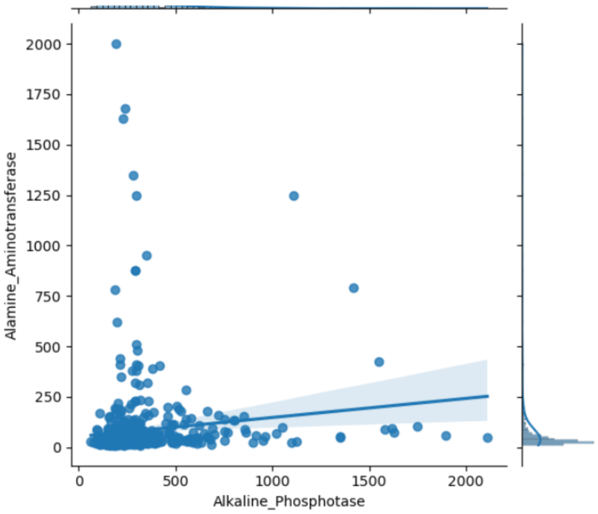


- Age seems to be a factor for liver disease for both male and female genders

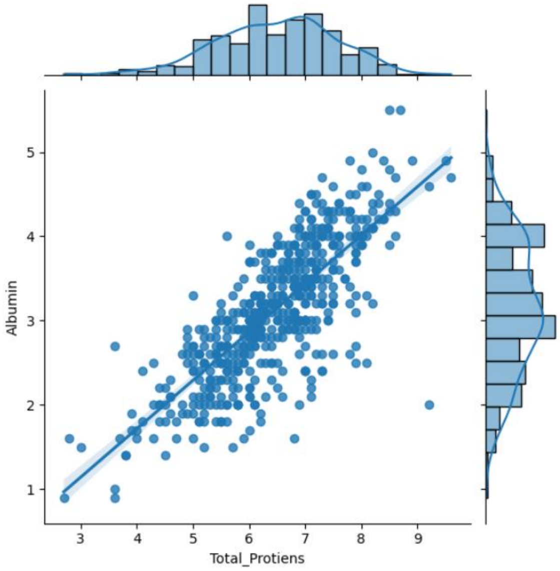




- There seems to be direct relationship between Total_Bilirubin and Direct_Bilirubin. We have the possibility of removing one of this feature.



- No linear correlation between Alkaline_Phosphotase and Alamine_Aminotransferase



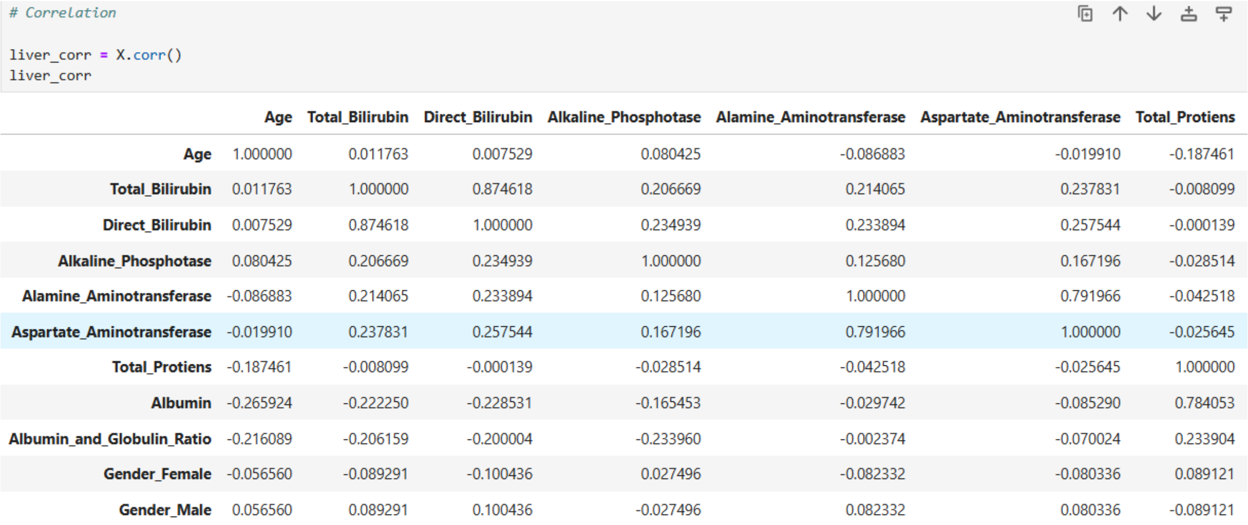
# Observation:

From the above jointplots and scatterplots, we find direct relationship between the following features:

- Direct_Bilirubin & Total_Bilirubin
- Aspartate_Aminotransferase & Alamine_Aminotransferase
- Total_Protiens & Albumin
- Albumin_and_Globulin_Ratio & Albumin

Hence, we can very well find that we can omit one of the features. I'm going to keep the follwing features:

- Total_Bilirubin
- Alamine_Aminotransferase
- Total_Protiens
- Albumin_and_Globulin_Ratio
- Albumin



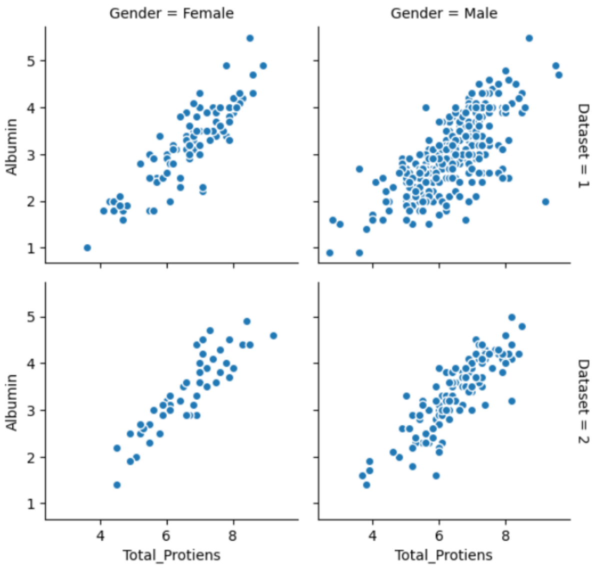
```
# Correlation

liver_corr = X.corr()
liver_corr
```

| | Age | Total_Bilirubin | Direct_Bilirubin | Alkaline_Phosphotase | Alamine_Aminotransferase | Aspartate_Aminotransferase | Total_Protiens |
|---|---|---|---|---|---|---|---|
| **Age** | 1.000000 | 0.011763 | 0.007529 | 0.080425 | -0.086883 | -0.019910 | -0.187461 |
| **Total_Bilirubin** | 0.011763 | 1.000000 | 0.874618 | 0.206669 | 0.214065 | 0.237831 | -0.008099 |
| **Direct_Bilirubin** | 0.007529 | 0.874618 | 1.000000 | 0.234939 | 0.233894 | 0.257544 | -0.000139 |
| **Alkaline_Phosphotase** | 0.080425 | 0.206669 | 0.234939 | 1.000000 | 0.125680 | 0.167196 | -0.028514 |
| **Alamine_Aminotransferase** | -0.086883 | 0.214065 | 0.233894 | 0.125680 | 1.000000 | 0.791966 | -0.042518 |
| **Aspartate_Aminotransferase** | -0.019910 | 0.237831 | 0.257544 | 0.167196 | 0.791966 | 1.000000 | -0.025645 |
| **Total_Protiens** | -0.187461 | -0.008099 | -0.000139 | -0.028514 | -0.042518 | -0.025645 | 1.000000 |
| **Albumin** | -0.265924 | -0.222250 | -0.228531 | -0.165453 | -0.029742 | -0.085290 | 0.784053 |
| **Albumin_and_Globulin_Ratio** | -0.216089 | -0.206159 | -0.200004 | -0.233960 | -0.002374 | -0.070024 | 0.233904 |
| **Gender_Female** | -0.056560 | -0.089291 | -0.100436 | 0.027496 | -0.082332 | -0.080336 | 0.089121 |
| **Gender_Male** | 0.056560 | 0.089291 | 0.100436 | -0.027496 | 0.082332 | 0.080336 | -0.089121 |

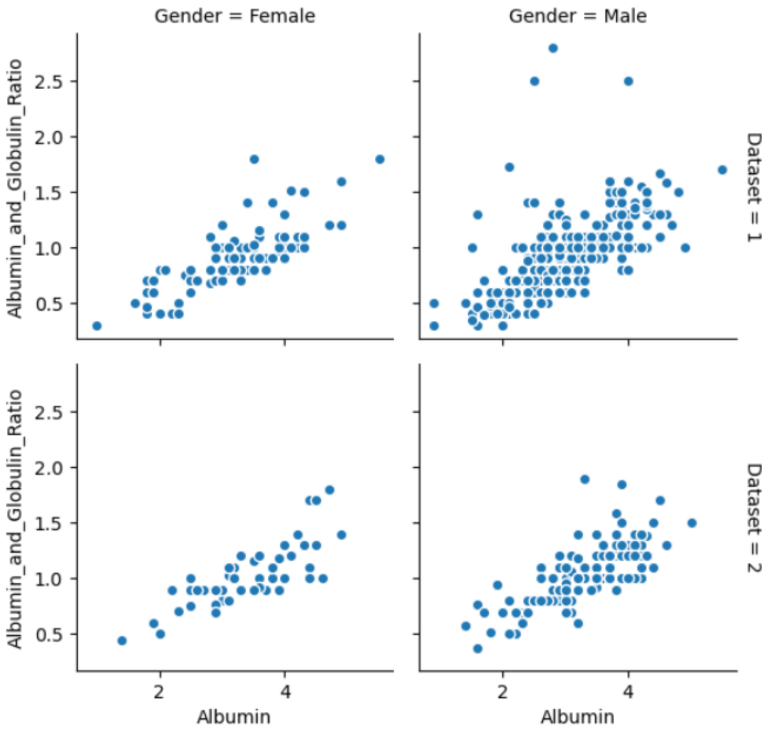
High Correlation between:

- Total_Protiens & Albumin
- Alamine_Aminotransferase & Aspartate_Aminotransferase
- Direct_Bilirubin & Total_Bilirubin
- There is some correlation between Albumin_and_Globulin_Ratio and Albumin. But its not as high as Total_Protiens & Albumin

Multivariate Analysis:



- There is linear relationship between Total_Protiens and Albumin and the gender. We have the possibility of removing one of this feature.

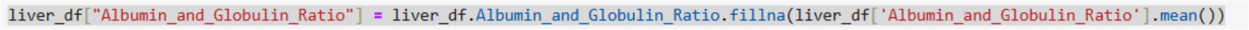


- There is linear relationship between Albumin_and_Globulin_Ratio and Albumin. We have the possibility of removing one of this feature.

Anomalies:

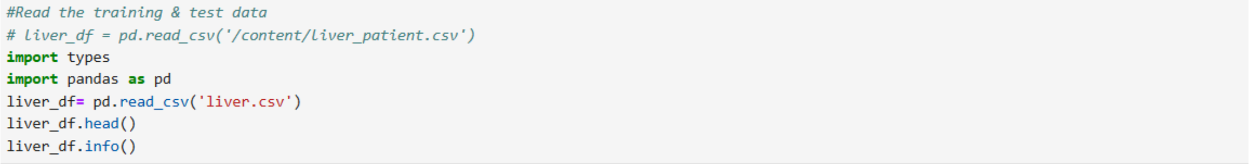
Replaced Null values with mean value.

Removed duplicate values.



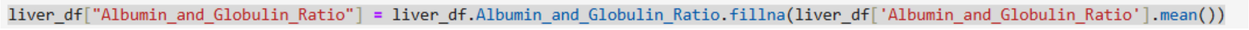
```
liver_df["Albumin_and_Globulin_Ratio"] = liver_df.Albumin_and_Globulin_Ratio.fillna(liver_df['Albumin_and_Globulin_Ratio'].mean())
```

Loading Data:



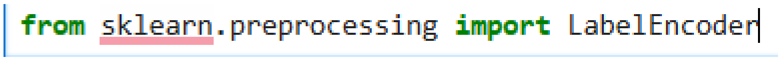
```
#Read the training & test data
# liver_df = pd.read_csv('/content/liver_patient.csv')
import types
import pandas as pd
liver_df= pd.read_csv('liver.csv')
liver_df.head()
liver_df.info()
```

Handling Missing Data:

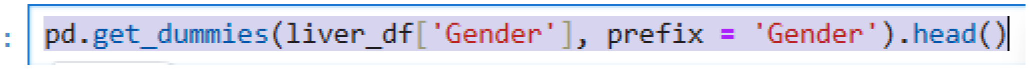


```
liver_df["Albumin_and_Globulin_Ratio"] = liver_df.Albumin_and_Globulin_Ratio.fillna(liver_df['Albumin_and_Globulin_Ratio'].mean())
```

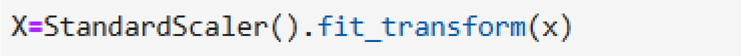
Data Transformation:



```
from sklearn.preprocessing import LabelEncoder
```

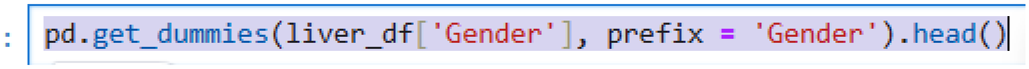


```
: pd.get_dummies(liver_df['Gender'], prefix = 'Gender').head()
```

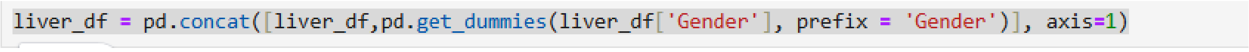


```
X=StandardScaler().fit_transform(x)
```

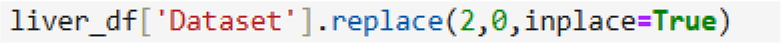
Feature Engineering:



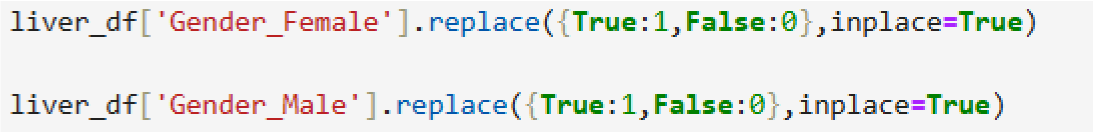
```
: pd.get_dummies(liver_df['Gender'], prefix = 'Gender').head()
```



```
liver_df = pd.concat([liver_df,pd.get_dummies(liver_df['Gender'], prefix = 'Gender')], axis=1)
```



```
liver_df['Dataset'].replace(2,0,inplace=True)
```



```
liver_df['Gender_Female'].replace({True:1,False:0},inplace=True)

liver_df['Gender_Male'].replace({True:1,False:0},inplace=True)
```