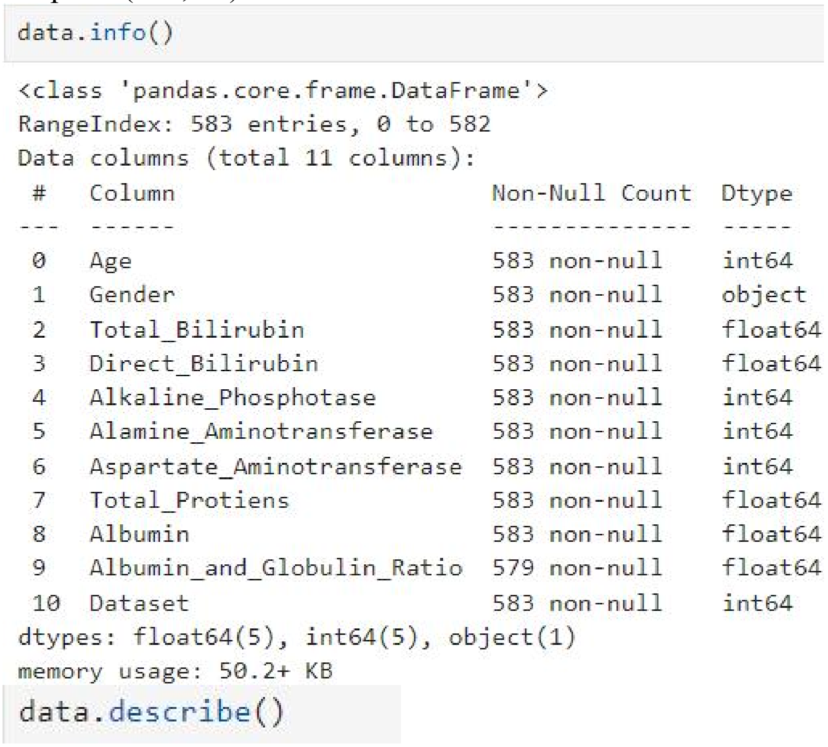
Data Collection and Preprocessing Phase

Date	09 July 2024
Team ID	SWTID1720013031
Project Title	Prediction and Analysis of Liver Patient Data Using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing

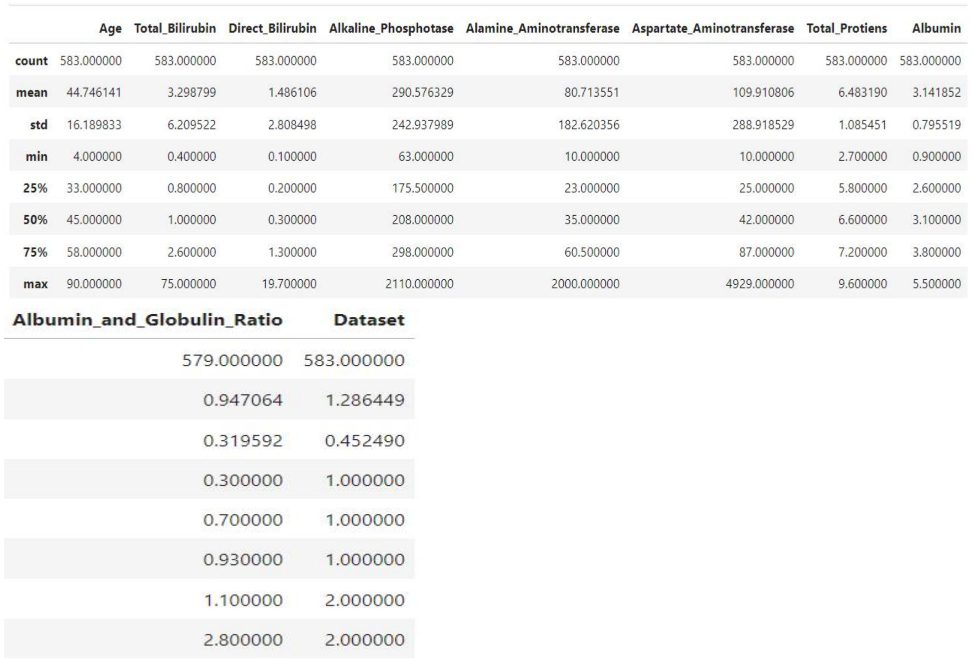
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Shape = (583, 11) 

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```

```
data.describe()
```



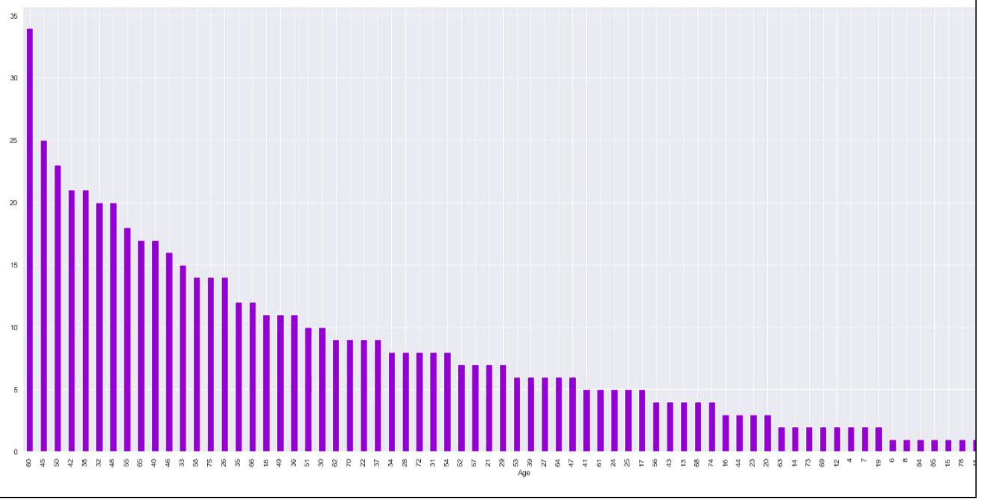
	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	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

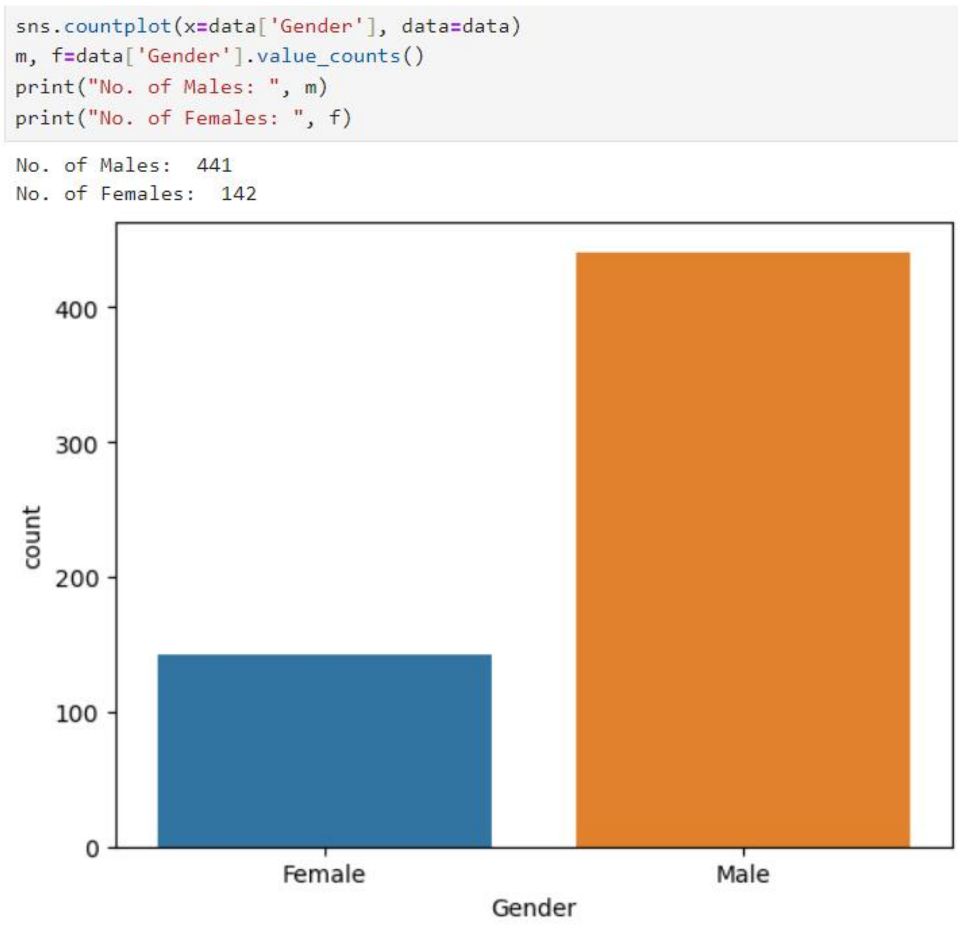
Albumin_and_Globulin_Ratio	Dataset
579.000000	583.000000
0.947064	1.286449
0.319592	0.452490
0.300000	1.000000
0.700000	1.000000
0.930000	1.000000
1.100000	2.000000
2.800000	2.000000

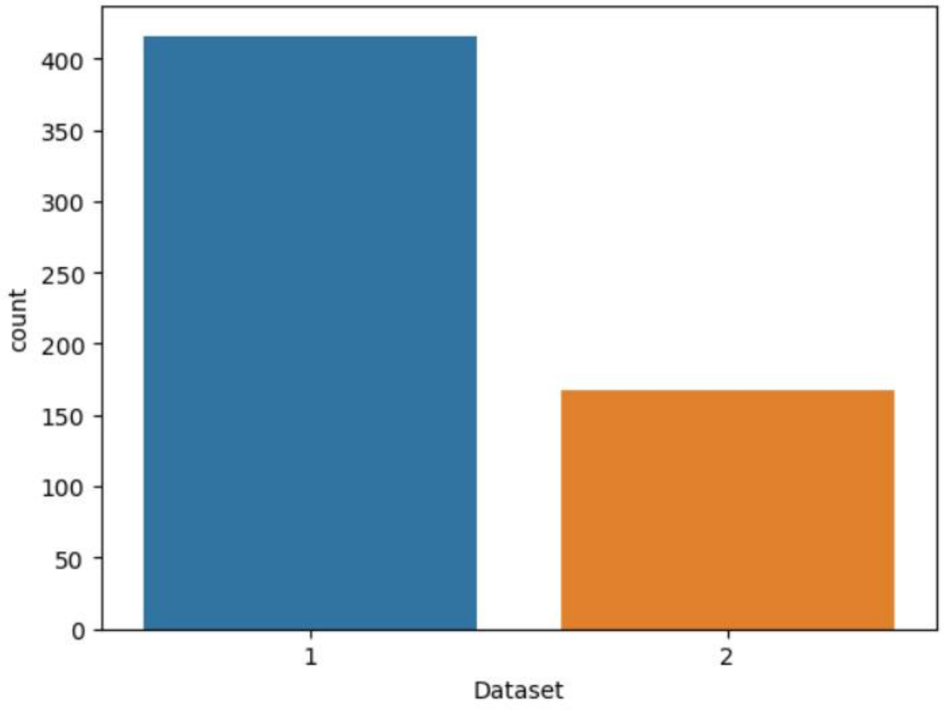
Univariate Analysis

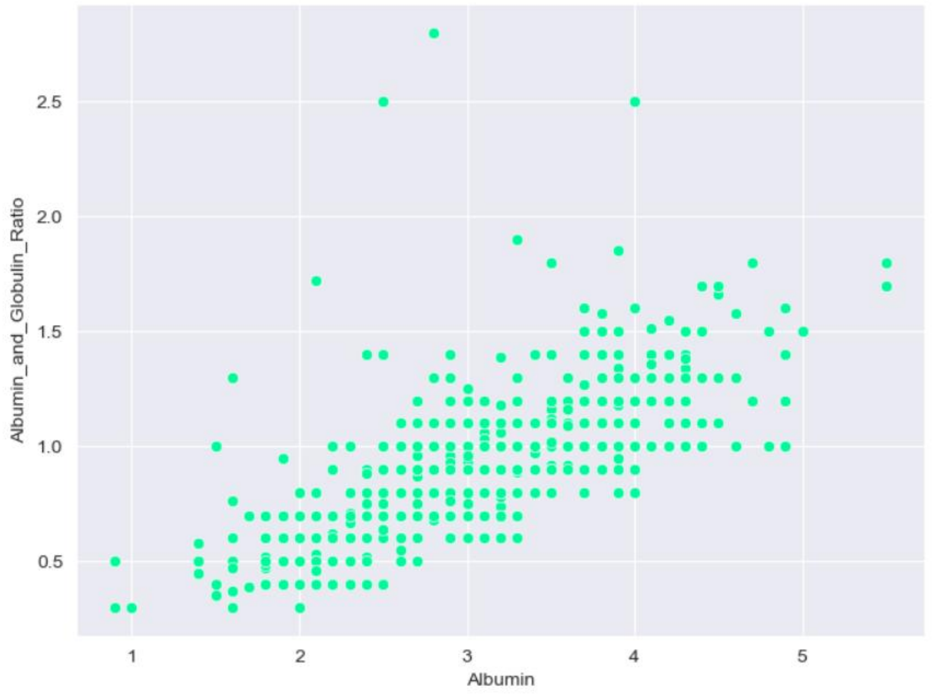
```
sns.set_style('darkgrid')
plt.figure(figsize=(25,10))
data['Age'].value_counts().plot.bar(color='darkviolet')
```

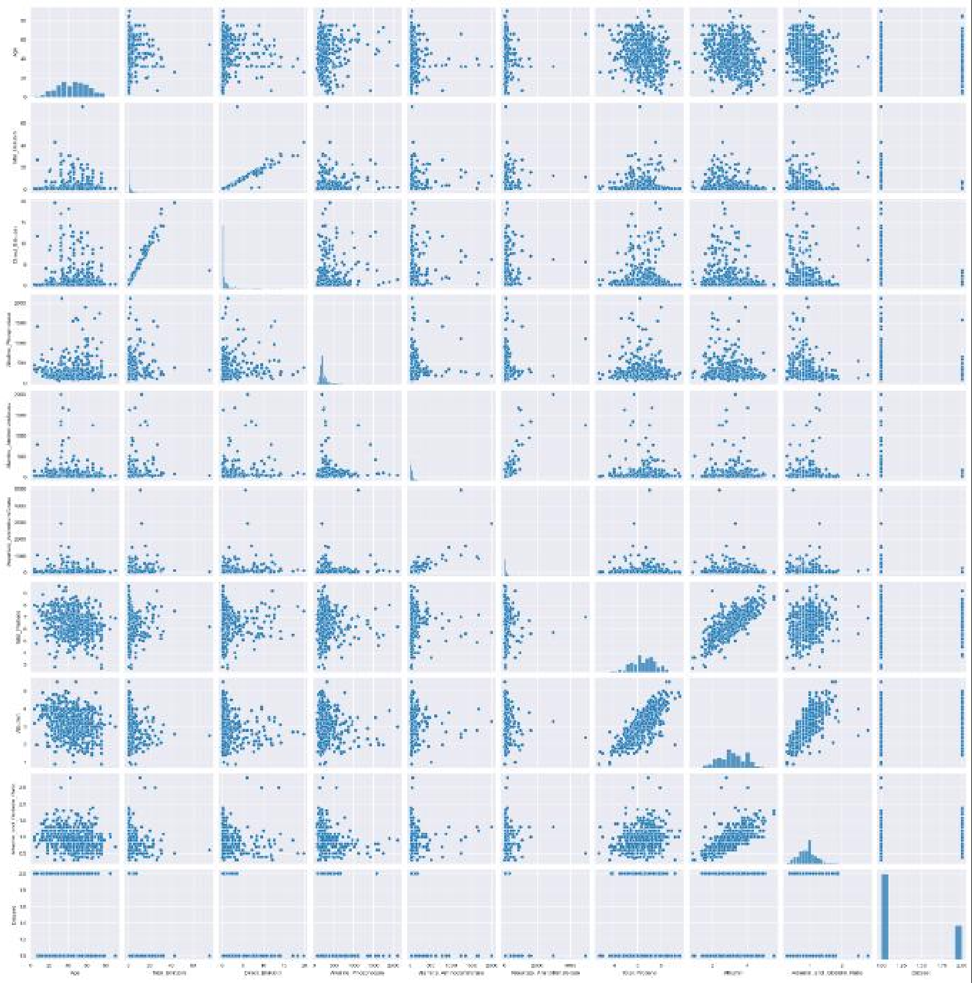
```
<Axes: xlabel='Age'>
```

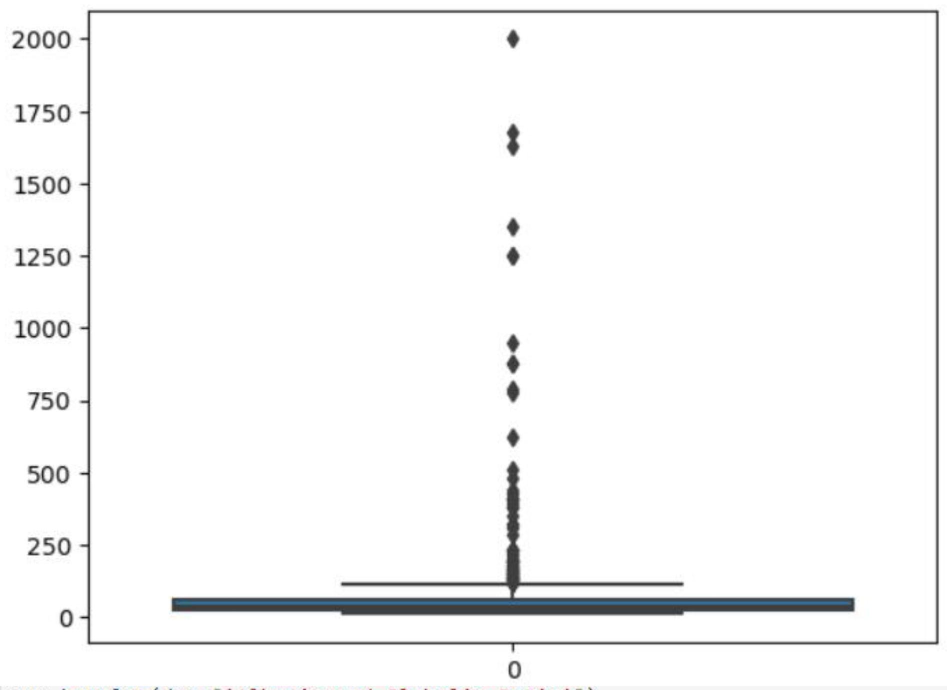
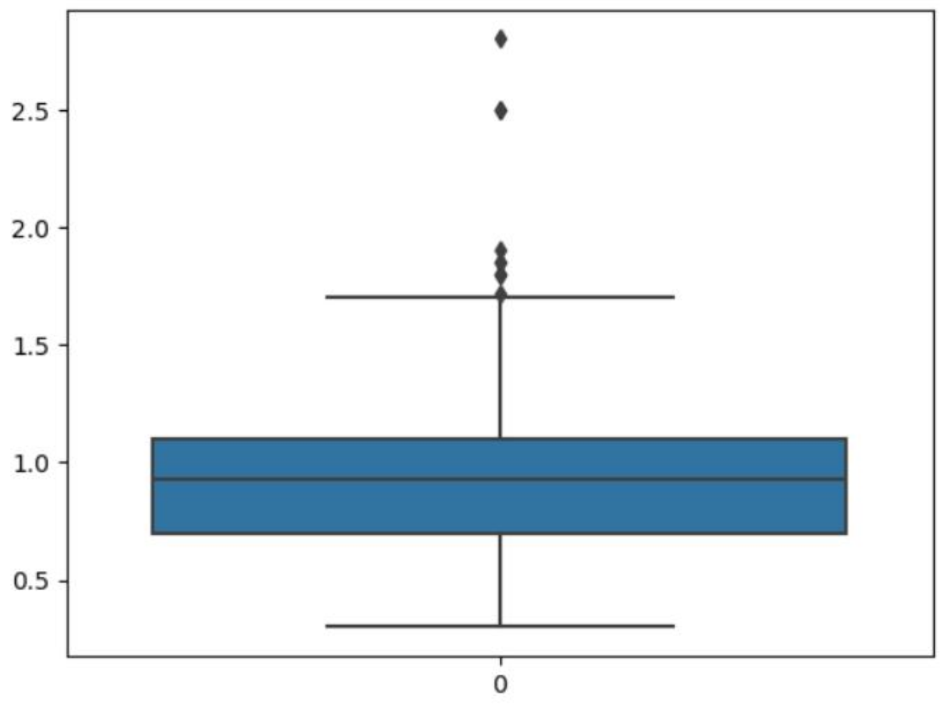


Bivariate Analysis	 ```sns.countplot(x=data['Gender'], data=data)``` ```m, f=data['Gender'].value_counts()``` ```print("No. of Males: ", m)``` ```print("No. of Females: ", f)``` ```No. of Males: 441``` ```No. of Females: 142``` 400 300 200 100 0 count Female Male Gender

	```
	sns.countplot(x=data['Dataset'], data=data)
	LD, NLD=data['Dataset'].value_counts()
	print("Liver disease patients: ", LD)
	print("Non-Liver disease patients: ", NLD)
	```
	```
	Liver disease patients:  416
	Non-Liver disease patients:  167
	```
	

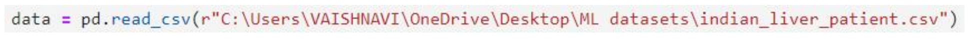
	```
	f, ax = plt.subplots(figsize=(8, 6))
	sns.scatterplot(x="Albumin", y="Albumin_and_Globulin_Ratio",color='mediumspringgreen',data=data);
	plt.show()
	```
	 

Multivariate Analysis	`sns.pairplot(data)`
	`plt.show()`
	

Outliers and Anomalies	`sns.boxplot(data['Alamine_Aminotransferase'])`
	 `<Axes: >`
	 
	 `sns.boxplot(data['Albumin_and_Globulin_Ratio'])`
	 `<Axes: >`
	 

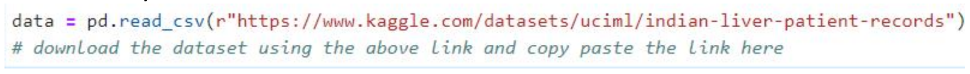
Data Preprocessing Code Screenshots

Loading Data



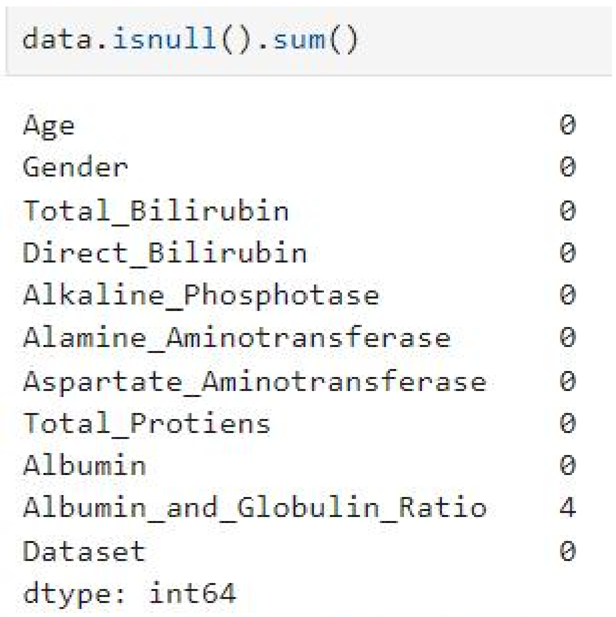
```
data = pd.read_csv(r"C:\Users\VAISHNAVI\OneDrive\Desktop\ML datasets\indian_liver_patient.csv")
```

General way:



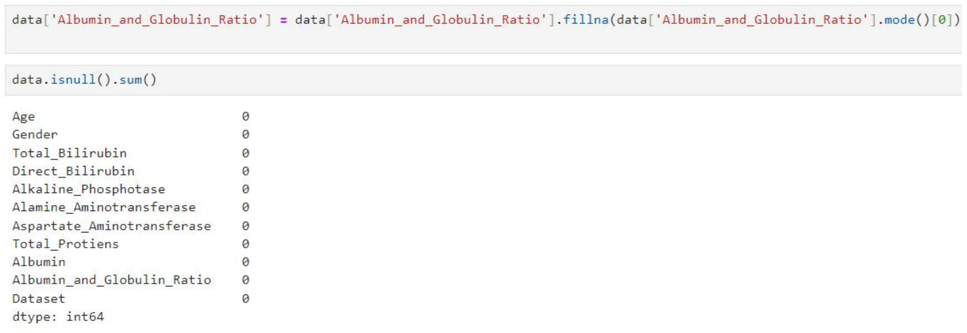
```
data = pd.read_csv(r"https://www.kaggle.com/datasets/uciml/indian-liver-patient-records")
# download the dataset using the above link and copy paste the link here
```

Handling Missing Data



```
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    4
Dataset                       0
dtype: int64
```

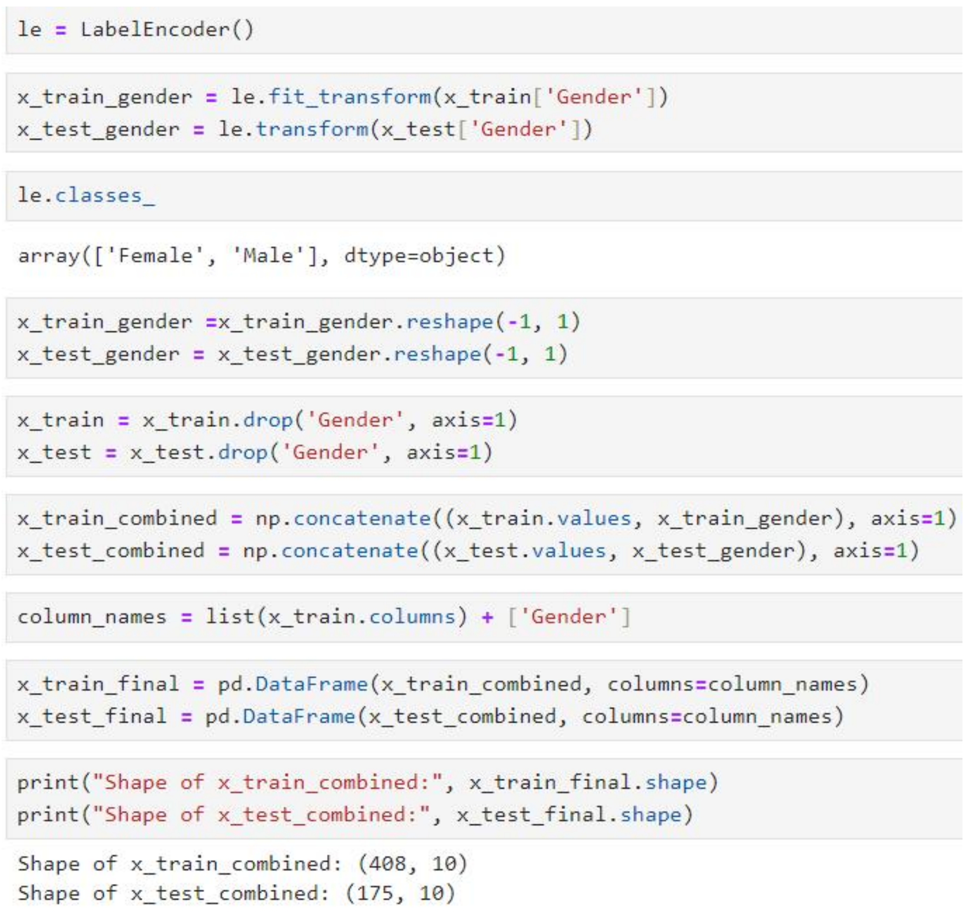


```
data['Albumin_and_Globulin_Ratio'] = data['Albumin_and_Globulin_Ratio'].fillna(data['Albumin_and_Globulin_Ratio'].mode()[0])

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
```

Data Transformation	wasn't required
Feature Engineering	LabelEncoder for Gender column



```
le = LabelEncoder()
```

```
x_train_gender = le.fit_transform(x_train['Gender'])
x_test_gender = le.transform(x_test['Gender'])
```

```
le.classes_
```

```
array(['Female', 'Male'], dtype=object)
```

```
x_train_gender =x_train_gender.reshape(-1, 1)
x_test_gender = x_test_gender.reshape(-1, 1)
```

```
x_train = x_train.drop('Gender', axis=1)
x_test = x_test.drop('Gender', axis=1)
```

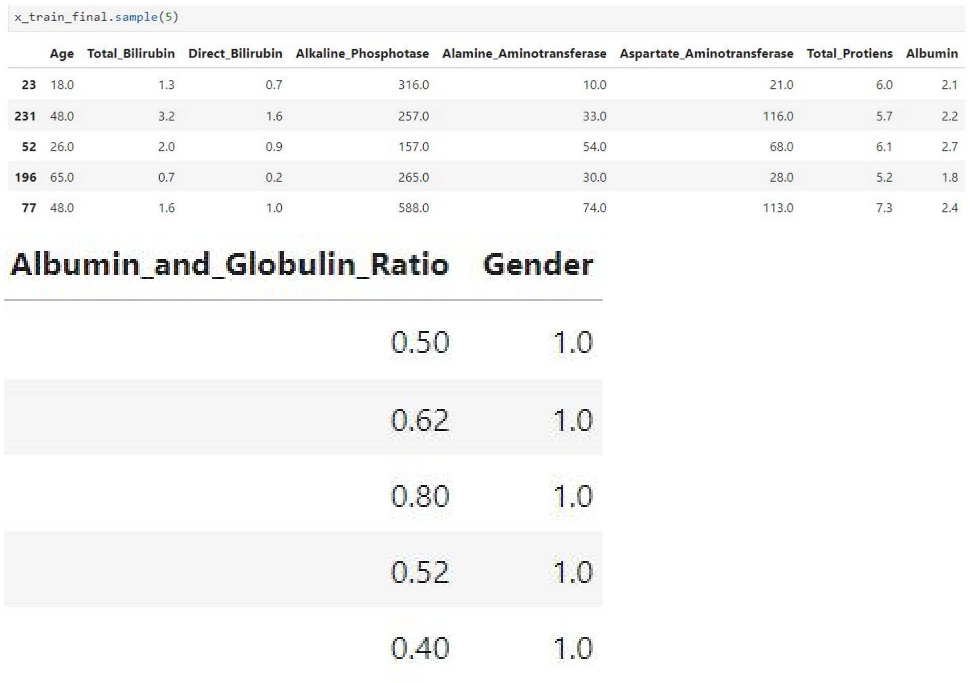
```
x_train_combined = np.concatenate((x_train.values, x_train_gender), axis=1)
x_test_combined = np.concatenate((x_test.values, x_test_gender), axis=1)
```

```
column_names = list(x_train.columns) + ['Gender']
```

```
x_train_final = pd.DataFrame(x_train_combined, columns=column_names)
x_test_final = pd.DataFrame(x_test_combined, columns=column_names)
```

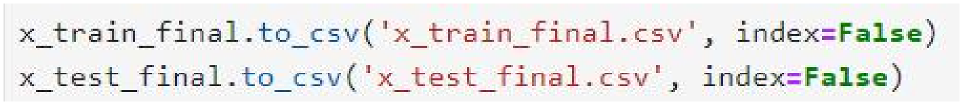
```
print("Shape of x_train_combined:", x_train_final.shape)
print("Shape of x_test_combined:", x_test_final.shape)
```

```
Shape of x_train_combined: (408, 10)
Shape of x_test_combined: (175, 10)
```



```
x_train_final.sample(5)
```

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Gender
23	18.0	1.3	0.7	316.0	10.0	21.0	6.0	2.1	0.50	1.0
231	48.0	3.2	1.6	257.0	33.0	116.0	5.7	2.2	0.62	1.0
52	26.0	2.0	0.9	157.0	54.0	68.0	6.1	2.7	0.80	1.0
196	65.0	0.7	0.2	265.0	30.0	28.0	5.2	1.8	0.52	1.0
77	48.0	1.6	1.0	588.0	74.0	113.0	7.3	2.4	0.40	1.0

Save Processed Data	 `x_train_final.to_csv('x_train_final.csv', index=False)` `x_test_final.to_csv('x_test_final.csv', index=False)`