Data Collection and Preprocessing Phase

Date	16July 2024
Team ID	SWTID1720074204
Project Title	prediction and analysis of liver patient data using ml
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

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	Basic statistics, dimensions, and structure of the data.
Univariate Analysis	Median: 0.93 Mean: 44.746141 3.298799 1.486106 290.576329 80.713551 109.910806 6.
Bivariate Analysis	Histogram countplot
Multivariate Analysis	Usually we drop that feature which has above 0.85% multicollinearity between two independent feature. Here we have only 'Total_Bilirubin' and 'Direct_Bilirubin' feature which has 0.87% mutlicollinearity. So we drop one of the feature from them and other independent feature has less multicollinearity, less than 0.80% So we keep that feature.
Outliers and Anomalies	There is no need of Standardization and Normalization of our dataset, as we using Ensemble Technique.

Data Preprocessing Code Screenshots	
Loading Data	

Handling Missing Data	# Cheaking Missing (NaN) Values:
	dataset.isnull().sum()
	 

Data Transformation	There is no need of Standardization and Normalization of our dataset, as we using Ensemble Technique.

Feature Engineering	# Target feature:
	print("Liver Disease Patients :",
	dataset['Dataset'].value_counts()[1])
	print("Non Liver Disease Patients :",
	dataset['Dataset'].value_counts()[2])
	
	# Visualization:
	sns.countplot(dataset['Dataset'])
	plt.show()
	# Target feature:
	print("Liver Disease Patients :",
	dataset['Dataset'].value_counts()[1])
	print("Non Liver Disease Patients :",
	dataset['Dataset'].value_counts()[2])
	
	# Visualization:
	sns.countplot(dataset['Dataset'])
	plt.show()

Save Processed Data