

Data Collection and Preprocessing Phase

Date	25 th June 2025
Team ID	LTVIP2025TMID36224
Project Title	Revolutionizing Liver Care: Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques.
Maximum Marks	2 Marks

Data Quality Report Template

This report summarizes the data quality issues identified in the liver cirrhosis dataset, along with their severity levels and proposed resolution plans. The goal is to systematically identify and rectify discrepancies to ensure high-quality data for accurate predictions.

Data Source	Data Quality Issue	Severity	Resolution Plan																																																												
Kaggle Dataset	<p>Missing values in all the columns of the dataset. (42 columns)</p> <pre>df.isnull().sum()</pre> <table> <tr><td>S.NO</td><td>0</td></tr> <tr><td>Age</td><td>0</td></tr> <tr><td>Gender</td><td>0</td></tr> <tr><td>Place(location where the patient lives)</td><td>134</td></tr> <tr><td>Duration of alcohol consumption(years)</td><td>0</td></tr> <tr><td>Quantity of alcohol consumption (quarters/day)</td><td>0</td></tr> <tr><td>Type of alcohol consumed</td><td>0</td></tr> <tr><td>Hepatitis B infection</td><td>0</td></tr> <tr><td>Hepatitis C infection</td><td>0</td></tr> <tr><td>Diabetes Result</td><td>0</td></tr> <tr><td>Blood pressure (mmhg)</td><td>0</td></tr> <tr><td>Obesity</td><td>0</td></tr> <tr><td>Family history of cirrhosis/ hereditary</td><td>0</td></tr> <tr><td>TCH</td><td>359</td></tr> <tr><td>TG</td><td>359</td></tr> <tr><td>LDL</td><td>359</td></tr> <tr><td>HDL</td><td>368</td></tr> <tr><td>Hemoglobin (g/dl)</td><td>0</td></tr> <tr><td>PCV (%)</td><td>30</td></tr> <tr><td>RBC (million cells/microliter)</td><td>552</td></tr> <tr><td>MCV (femtoliters/cell)</td><td>9</td></tr> <tr><td>MCH (picograms/cell)</td><td>658</td></tr> <tr><td>MCHC (grams/deciliter)</td><td>672</td></tr> <tr><td>Total count</td><td>10</td></tr> <tr><td>Polymorphs (%)</td><td>0</td></tr> <tr><td>...</td><td></td></tr> <tr><td>SGOT/AST (U/L)</td><td>0</td></tr> <tr><td>SGPT/ALT (U/L)</td><td>0</td></tr> <tr><td>USG Abdomen (diffuse liver or not)</td><td>0</td></tr> <tr><td>Outcome</td><td>54</td></tr> </table> <p>dtype: int64</p>	S.NO	0	Age	0	Gender	0	Place(location where the patient lives)	134	Duration of alcohol consumption(years)	0	Quantity of alcohol consumption (quarters/day)	0	Type of alcohol consumed	0	Hepatitis B infection	0	Hepatitis C infection	0	Diabetes Result	0	Blood pressure (mmhg)	0	Obesity	0	Family history of cirrhosis/ hereditary	0	TCH	359	TG	359	LDL	359	HDL	368	Hemoglobin (g/dl)	0	PCV (%)	30	RBC (million cells/microliter)	552	MCV (femtoliters/cell)	9	MCH (picograms/cell)	658	MCHC (grams/deciliter)	672	Total count	10	Polymorphs (%)	0	...		SGOT/AST (U/L)	0	SGPT/ALT (U/L)	0	USG Abdomen (diffuse liver or not)	0	Outcome	54	High	Use mean/median imputation.
S.NO	0																																																														
Age	0																																																														
Gender	0																																																														
Place(location where the patient lives)	134																																																														
Duration of alcohol consumption(years)	0																																																														
Quantity of alcohol consumption (quarters/day)	0																																																														
Type of alcohol consumed	0																																																														
Hepatitis B infection	0																																																														
Hepatitis C infection	0																																																														
Diabetes Result	0																																																														
Blood pressure (mmhg)	0																																																														
Obesity	0																																																														
Family history of cirrhosis/ hereditary	0																																																														
TCH	359																																																														
TG	359																																																														
LDL	359																																																														
HDL	368																																																														
Hemoglobin (g/dl)	0																																																														
PCV (%)	30																																																														
RBC (million cells/microliter)	552																																																														
MCV (femtoliters/cell)	9																																																														
MCH (picograms/cell)	658																																																														
MCHC (grams/deciliter)	672																																																														
Total count	10																																																														
Polymorphs (%)	0																																																														
...																																																															
SGOT/AST (U/L)	0																																																														
SGPT/ALT (U/L)	0																																																														
USG Abdomen (diffuse liver or not)	0																																																														
Outcome	54																																																														
Kaggle Dataset	Categorical data in the dataset	Moderate	Perform encoding (e.g., Label Encoding or One-Hot Encoding).																																																												