Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740097
Project Title	Predicting baseline histoligical stage in HCV patients using ML
Maximum Marks	6 Marks

Section	Description

Data Overview	Now that we know the nature of the data, let's preprocess the information that was gathered.
	The downloaded data set may contain too much randomness to be used for training a machine learning model, so to get good results, the dataset must be carefully cleaned. The following steps are involved in this activity.
	? Handling missing values
	? Handling categorical data
	There are no categorical variables in our datasets
	The general procedures for pre-processing data before applying it to machine learning are as follows. Your dataset's state will determine whether or not you need to follow each of these stages.
	The first step will be to find the shape i.e. dimensions of the dataset. To find the shape of our data, the df. shape method is used. To find the data type, the df.info() function is used.
	`df.shape`
	`(1385, 29)`
	Thus, our dataset contains 1385 rows and 29 columns.

```
: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1385 entries, 0 to 1384
Data columns (total 29 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 1385 non-null   int64
 1   Gender                              1385 non-null   int64
 2   BMI                                 1385 non-null   int64
 3   Fever                               1385 non-null   int64
 4   Nausea/Vomting                      1385 non-null   int64
 5   Headache                            1385 non-null   int64
 6   Diarrhea                            1385 non-null   int64
 7   Fatigue & generalized bone ache     1385 non-null   int64
 8   Jaundice                            1385 non-null   int64
 9   Epigastric pain                     1385 non-null   int64
 10  WBC                                 1385 non-null   int64
 11  RBC                                 1385 non-null   float64
 12  HGB                                 1385 non-null   int64
 13  Plat                                1385 non-null   float64
 14  AST 1                               1385 non-null   int64
 15  ALT 1                               1385 non-null   int64
 16  ALT4                                1385 non-null   float64
 17  ALT 12                              1385 non-null   int64
 18  ALT 24                              1385 non-null   int64
 19  ALT 36                              1385 non-null   int64
  20  ALT 48                             1385 non-null   int64
  21  ALT after 24 w                     1385 non-null   int64
  22  RNA Base                           1385 non-null   int64
  23  RNA 4                              1385 non-null   int64
  24  RNA 12                             1385 non-null   int64
  25  RNA EOT                            1385 non-null   int64
  26  RNA EF                             1385 non-null   int64
  27  Baseline histological Grading      1385 non-null   int64
  28  Baselinehistological staging       1385 non-null   int64
dtypes: float64(3), int64(26)
memory usage: 313.9 KB
```

From the above, we can see that there are no null values in this dataset.

Data exploration and preprocessing report

Data set variables will be stastically analyzed to identify patterns and outliers with python employed for preprocessing tasks like normalization and feature engineering.data cleaning will address missing values and outliers,ensuring quality for subsequent analysis and modelling and forming a strong foundation for insights and predictions.

images will be