DATA PRE-PROCESSING

READING THE DATASET

Date	09 November 2022
Team ID	PNT2022TMID05432
Project Name	Statistical Machine LearningApproaches To Liver Disease Prediction

READING THE DATASET :

Firstly we will be loading the dataset from the folders using the pandas library. While reading the dataset we will be dropping the null column. This dataset is a clean dataset with no null values and all the features consist of 0's and 1's. Whenever we are solving a classification task it is necessary to check whether our target column is balanced or not. We will be using a bar plot, to check whether the dataset is balanced or not.

PYTHON

```
# Reading the train.csv by removing the

# last column since it's an empty column

DATA_PATH = "dataset/Training.csv"

data = pd.read_csv(DATA_PATH).dropna(axis = 1)


# Checking whether the dataset is balanced or not

disease_counts = data["prognosis"].value_counts()

temp_df = pd.DataFrame({

        "Disease": disease_counts.index,
```

```
    "Counts": disease_counts.values
})

plt.figure(figsize = (18,8))
sns.barplot(x = "Disease", y = "Counts", data = temp_df)
plt.xticks(rotation=90)
plt.show()
```

From the above plot, we can observe that the dataset is a balanced dataset i.e. there are exactly 120 samples for each disease, and no further balancing is required. We can notice that our target column i.e. prognosis column is of object datatype, this format is not suitable to train a machine learning model. So, we will be using a label encoder to convert the prognosis column to the numerical datatype. Label Encoder converts the labels into numerical form by assigning a unique index to the labels. IF the total number of labels is n, then the numbers assigned to each label will be between 0 to n-1.

PYTHON :

```
# Encoding the target value into numerical
# value using LabelEncoder
E ncoder = LabelEncoder()
data["prognosis"] = encoder.fit_transform(data["prognosis"])
```