Integrated Capstone project

Liver Disease prediction

D3-P9

Made by -

Akshat Lakhara Anil Kumar Dhayal Deepshikha Kumari

Disha Kamlesh Kumar Shah Swaraj Haresh Purohit Vidhi Sachin Thakare

Table of Contents

1. [Overview](#overview)
2. [Task](#Task)
3. [Installations and Packages](#installations)
4. [Dataset](#Dataset)
5. [Exploratory Data Analysis](#eda)
6. [SQL Queries](#sql)

Overview

The liver is one of the most vital organs in the human body as its operation is to flush out toxins from the body, fight infections, and regulate hormones and bile output. One of the most dangerous and chronic illnesses in the world, liver disease can have a variety of negative implications if it is not treated promptly. This project focuses on the health report of patients across India to understand their livers condition due to change in their lifestyle and living habits. This project is aimed to alleviate the burden of analyzing every single patient’s data during clinical diagnoses of liver disease.

Task

The goal of this project is to create a classifier that predicts whether a subject is healthy and does not have a liver disease or unhealthy and a patient of liver disease. These inferences are bases on the medical record of 30 thousand patients which include various clinical records acquired through medical tests.

Installations and Packages

This project works on jupyter notebook which can be accesses from [Project Jupyter | Installing Jupyter](https://jupyter.org/install) . Along with this please ensure that the required libraries are installed in your environment.

1. Conda Packages

* Pandas
* Numpy
* Matplotlib.pyplot
* Seaborn

Dataset

This project will use data from the Liver Disease dataset, which includes over 30,000 cases. According to statistics, the number of individuals with liver disease has increased recently because of certain lifestyle choices and behaviors, including abusing drugs, consuming excessive amounts of alcohol, breathing in hazardous gasses, gaining too much weight, and consuming tainted food.

Certain tests are conducted on the patient to find out the attributes:

1. ID
2. Age of the patient
3. Gender of the patient
4. Total Bilirubin (TB)
5. Direct Bilirubin (DB)
6. Alkphos Alkaline Phosphotase
7. Sgpt Alamine Aminotransferase
8. Sgot Aspartate Aminotransferase
9. Total Proteins
10. Albumin (ALB)
11. Albumin and Globulin Ratio
12. City Code
13. State Code

The state code name file has:

1. State Code
2. State Name

The city code name file has:

1. City Code
2. State Code
3. City Name

The status file has:

1. ID
2. Status

Exploratory Data Analysis

Data Merging

The data of the patient is distributed among different files so the first step was to collect data and put it in a clean csv format for analysis. Liver patient, state code name, city code name and status dataset were merged to form a complete dataset which gives all the information regarding the patient. The datasets were merged based on the common columns like city code, state code, id etc.

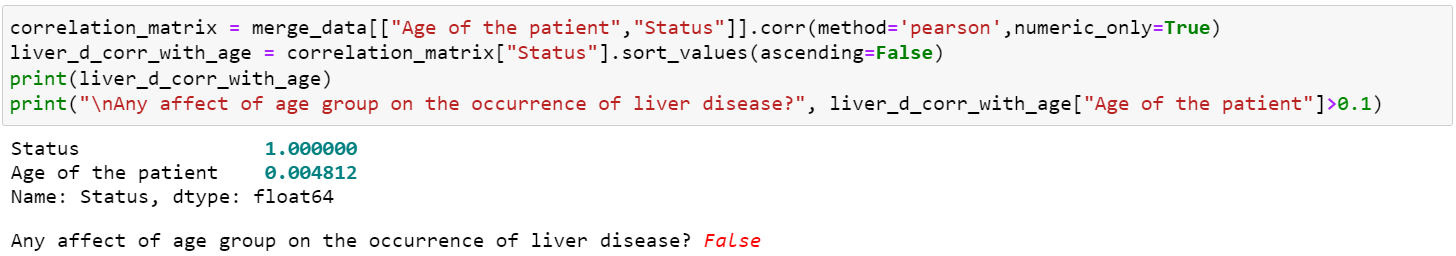


Q1. Missing Values

The dataset had missing values which must be taken care of after the data's behavior is understood. To get the missing values in each column, we use:



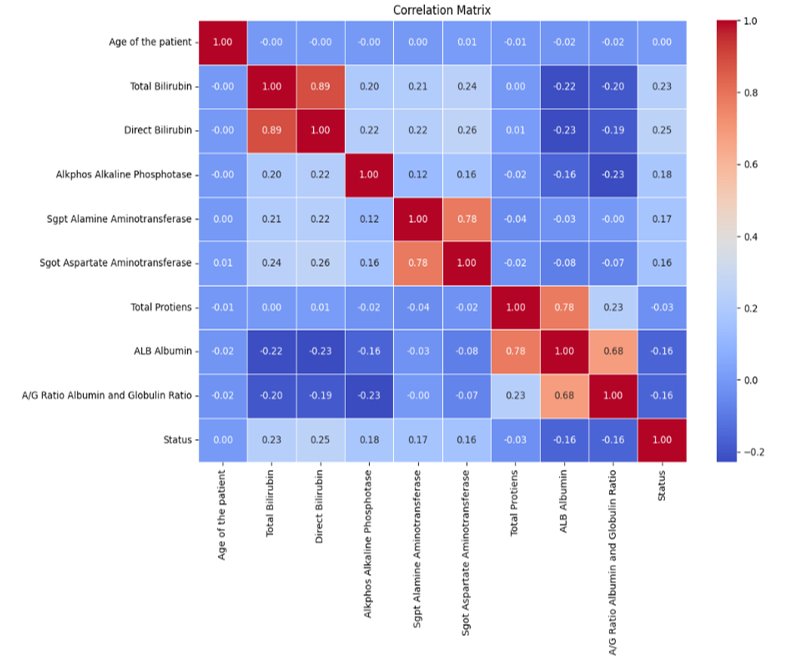
Q2. Any effect of age group on the occurrence of liver disease?



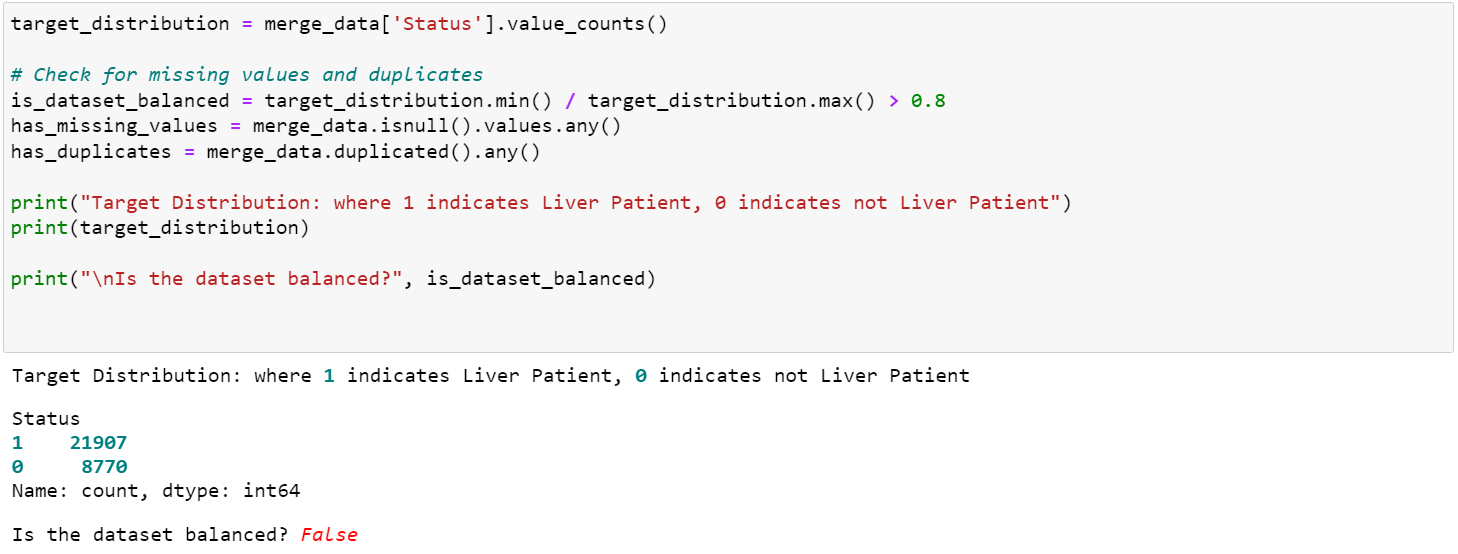
Q3. Any relationship between varied factors and with the presence of liver disease given in the dataset?

A computer code with text

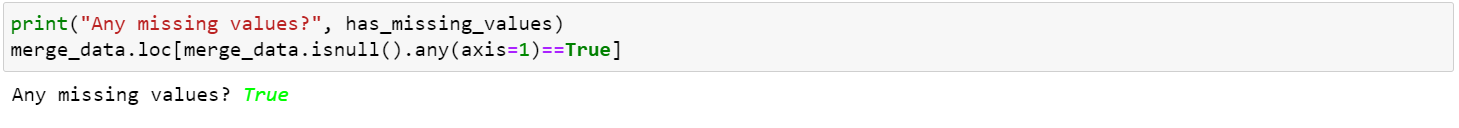
Description automatically generated with medium confidence



Q4.1. Is the dataset balanced?



Q4.2. Are there any missing values?



Q4.3. Are there any duplicate values?



SQL Queries

1. Retrieve the liver patients with a direct bilirubin level higher than the average direct bilirubin level for all patients.

A screenshot of a computer

Description automatically generated

1. Retrieve average alkaline phosphatase level for liver patients under the age of 50.

A screenshot of a computer

Description automatically generated

1. Retrieve the average age and average total proteins for each gender.

A screenshot of a computer

Description automatically generated

1. Retrieve the gender with the highest total protein level.

A screenshot of a computer

Description automatically generated

1. Retrieve the gender with the highest average total protein level among liver patients below the age of 50.

A screenshot of a computer

Description automatically generated

1. Retrieve the gender with the highest average total bilirubin and direct bilirubin levels.

A screenshot of a computer

Description automatically generated