Report IE-509

Rubul Gogoi
Roll number: 23n0462

November 25, 2023

Data Source

In the exploratory data analysis, the dataset used originates from the **kaggle website (https://www.kaggle.com/code/sidjsdd/diabetes-model).**The dataset consisted of the following columns:

- **id , chol, ,stab.glu ,hdl ,ratio ,glyhb ,location ,age ,gender ,height ,weight frame ,bp.1s ,bp.1d ,bp.2s ,bp.2d ,waist ,hip ,time.ppn**

Analysis Steps

The analysis process involved several key steps to derive meaningful insights from the dataset:

- **Data Cleanup:** The raw dataset underwent a through cleanup process to handle missing values.In the data set, out of 402 columns ,262 values of columns **bp.2s and bp.2d** (systolic and diastolic pressure on day 2) were missing. **Since almost two-third of the values were missing , both the columns were removed.** After that , the rows which had even one missing value were removed.Also the column **time.ppn(Peripheral parenteral nutrition: a short term nutritional support (ideally 5 – 7 days) for appropriate patients)** was ignored as it involved a time factor and also I couldn't interpret the parameter properly.And similarly the **ratio** column was ignored.
- **Feature Selection:** The following features were selected in order to analyze the data set:

1. **chol:** The chol column represented **the total amount of cholesterol in the blood of the patients.** It is an important parameter in order to get an overview of the cholesterol levels of the diabetic patients. Using this column , a new parameter was defined which showed whether the cholesterol level is **normal, boderline high or high.**
2. **stb.glu:** It is a kind of **blood test before eating**. Its normal range is 70-100 milligrams per deciliter.
3. **hdl:** This column represents **high density liporotein(good cholesterol)** of the individuals.
4. **glyhb:** This column represents **glycosylated hemoglobin** and it is used as a parameter **to check whether a person has diabetes or not.** A new column was defined using this column and named **status** which showed whether the person has **diabetes, prediabetes or he is normal.**
5. **age**
6. **gender**
7. **height, weight:** Using the height and weight column , the body mass index of the individuals were calculated and then categorized as **underweight , normal , overweight and obese** by introducing a new column **bmistatus.**
8. **bp.1s:** It represents the systolic pressure of the individuals on day 1
9. **bp.1d:** It represents the diastolic pressure of the individuals on day 1

- **Visualization:** Two **bar plots** were genereated:
  1. **Gender vs Glycosylated hemoglobin grouped by cholesterol**
  2. **Gender vs Glycosylated hemoglobin grouped by BMI status**
     Also a **scatter plot** was generated between **age and glycosylated hemoglobin** and then **grouped** by the **diabetic condition of the individual**

- **Heat Maps:** A Heat map was generated to visualize **correlations** among variables **chol, stab.glu, hdl,glyhb and BMI** in the dataset.

Tools Used

The analysis was conducted using the following tools and libraries in Python:

- **Pandas:** Used for data manipulation and cleaning.
- **Matplotlib and Seaborn:** Employed for data visualization.

Conclusions

1. A person with **higher cholesterol , body mass index and stabilized glucose** has a **higher chance** of **getting diabetes.**
2. **With age** , the **risk of getting diabetes becomes higher** even though one can have diabetes at an younger age
3. From the **heat plot**, it was seen that **hdl(high densty lipoprotein) has a negative correlation with glyhb(glycosylated hemoglobin)**. Therefore a person with **higher good cholesterol** has a **lower chance of getting diabetes**
4. **Stabilized Glucose** is **highly correlated** to **Glycosylated Hemoglobin** and so it can be also used as a **measure to test diabetes**
5. A person having **unusual blood preesure levels** has a **22% of getting diabetes**
6. A person with **high cholesterol** has **31% of getting diabetes**

Contribution

The entirety of the analysis and conclusions presented in the report is the result of my own work. While I may have consulted external sources like google and help from my sister Khusbu Gogoi(for knowing the medical terms) for inspiration and guidance, every piece of code and interpretation presented here is original.