Chronic Diseases

The analysis of long term health diseases

Mariam Hersi
Data Science (CMP020N205S)
HER22555014
April 15, 2024

I will be using data science techniques to analyse chronic diseases. I will use that to come up with predictions and make conclusions to hopefully benefit the medical industry.

I. Introduction (*Heading 1*)

In this essay I will be analysing data on chronic diseases. Chronic Diseases are diseases that last for long periods of time and slow progression. They usually cannot be passed around from person to person. This includes cardiovascular disease, chronic respiratory diseases, cancer and diabetes [1]. It would benefit the medical industry to know what brings about this kind of illness and the best way to cure this. Therefore, I believe the collection of data surrounding these kinds of illness and then the analysis of such data would help achieve such goals. I have chosen to focus on the chronic disease diabetes to further elaborate on these long term illnesses. This is the URL for the websites for which I collected the data from:

1. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

II. Hypothesis

A. *Problem 1*

First, I would like to identify the main contributing factors to diabetes. With this information, the medical industry will be able to decipher what the causes are and how to prevent them. They can help patients put in the preventive measures necessary which will help minimise overall diagnoses.

B. *Problem 2*

Second, I would like to predict hospital readmissions in relations to diabetes. This can help understand how dangerous these diseases. It will also allow us to see how life changing these diseases. We can question whether patients can live with these diseases or do they hinder day to day life.

C. *Problem 3*

Lastly, I would like to find what treatments have worked best for the patients of diabetes. This will be very useful information to health professionals as they can help patients with the pain and symptoms of the disease. It can also potentially bring them closer to what might cure this disease or simply decrease the time it affects the patient.

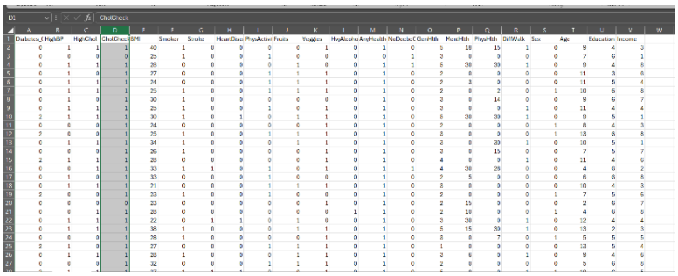
III. Preprocessing

Before we can analyse the data and make predictions or draw conclusions from it we must make sure the data we are using is the best condition and is credible. To do that I will preprocess the data. Preprocessing is the transformation of raw data into something that is understandable and useable [2]. Here are the steps that I went through to preprocess the data that I had:

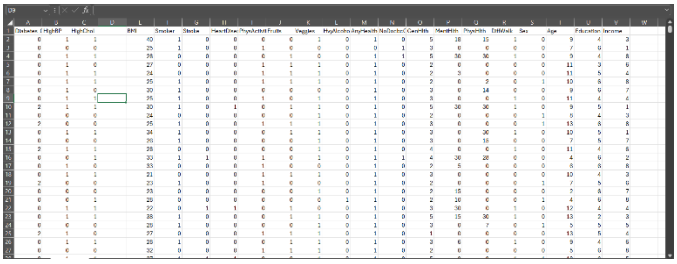
A. *Step 1 – Data Cleaning*

Here we must modify the data and remove any data that is inaccurate duplicate ,incomplete, incorrectly formatted, or corrupted[3]. Considering this, I made sure to look for any inconsistencies and removed the duplicate Chol features. This came as a result of the integration.

This is before the removal of the feature:



This is after the removal:

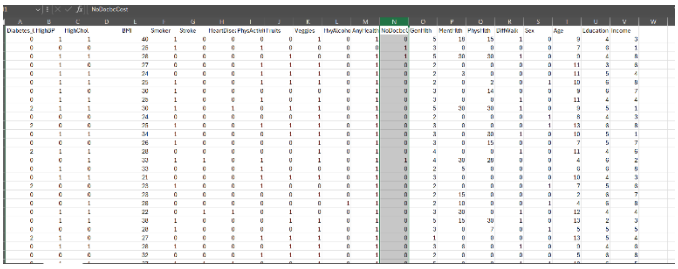


B. *Step 2 – Data Reduction*

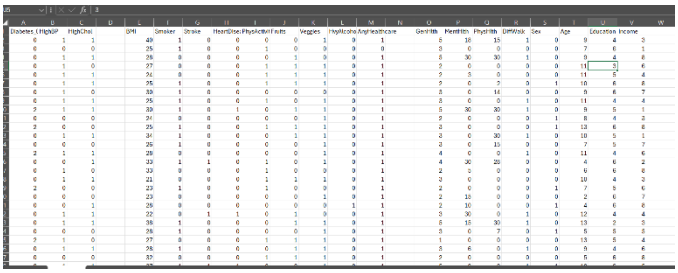
This step involves reducing the size of the data so that it can be processed whilst still making sure to keep the most important information [4]. I chose to do this by removing the NoDocsbcCost feature because it wouldn't be necessary to make my solution .

This before the removal of the feature:

Identify applicable funding agency here. If none, delete this text box.



This is after the removal:

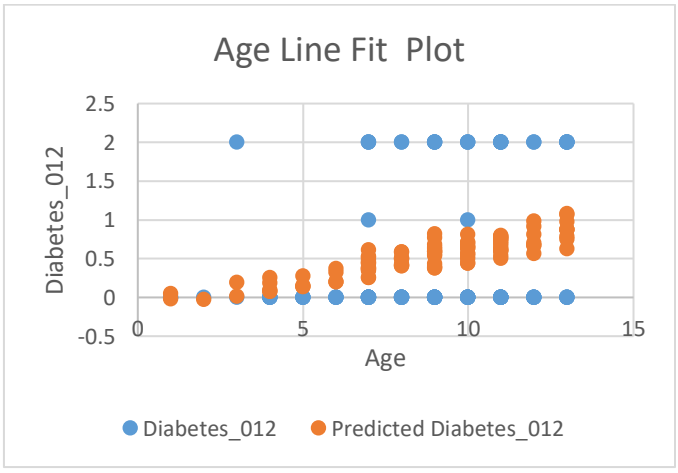


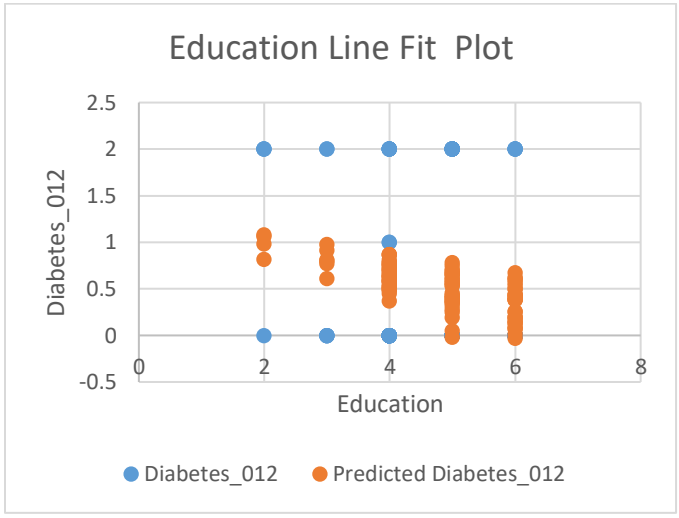
IV. Proposed Solution

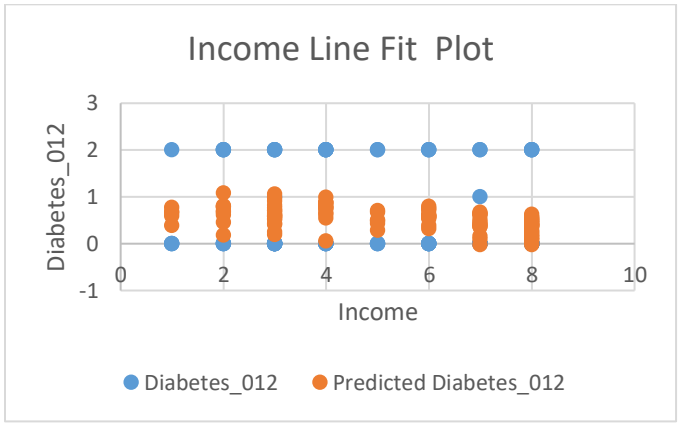
Now that I have completed the preprocessing of the data, I will now process the data to come to a solution. I will use the machine learning technique linear regression. This is viewing the relationship between an independent variable and dependent variable [5].

I made a linear regression model to find the connection between diabetes and age, income and education. From all of the line plot graphs it is clear to see that the higher age the more likely to contract diabetes. However, the more educated you are on diabetes the less likely you are to get diabetes. This is perhaps because you are understanding the severity of the illness making you more cautious of your health.

On the other hand, if you look at the Income line plot, it doesn't swing either way and it is difficult see whether depending on income you are more likely to become diabetic. From this it is clear to see that income has no effect on the illness and depends more perhaps on lifestyle.







References

[1] S. Bernell and S. W. Howard, "Use Your Words Carefully: What Is a Chronic disease?," *Frontiers in Public Health*, vol. 4, no. 159, Aug. 2016, doi: https://doi.org/10.3389/fpubh.2016.00159.

[2] U. Kose, D. Gupta, V. Hugo, and A. Khanna, *Data Science for COVID-19 Volume 1*. Academic Press, 2021.

[3] A. Fitzgerald, "Data Cleansing: What It Is, Why It Matters & How to Do It," *blog.hubspot.com*, Feb. 03, 2022. https://blog.hubspot.com/marketing/data-cleansing

[4] "Data Reduction in Data Mining," *GeeksforGeeks*, Jan. 27, 2020. https://www.geeksforgeeks.org/data-reduction-in-data-mining/

[5] V. Kanade, "What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022," *Spiceworks*, Dec. 14, 2022. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/