Portfolio for Data Science COM618: Diabetes Data Analysis

> **Student Name:** Matthew Wilcox

> **Student Number:** Q16048563

Introduction

Diabetes is a large, growing problem within the Medical community, with Diabetes being the direct cause of 1.6 million deaths in 2021 [^1], with the mortality rates increasing since 2000. Furthermore, 59% of adults (30 and over) were living without medication with Diabetes in 2022. These figures represent a challenge in the medical area with effectively screening for people who have the disease and also a problem in distributing care, leading to these increased mortality rates. This represents that the industry could benefit from some analysis of data on the problem in order to help them make informed decisions quicker to help take preventative measures to stop the disease/ distribute care to people with the disease to stop mortality rates rising further.

[^1]: WHO Diabetes Fact Sheet

- ** DELETE Write about the background of the topic area and a brief literature review, some challenges in the sector and/or the problem statement.**

Aims/Objectives of the coursework:

- This Coursework aims to provide an analytic insight into some of the variables that can lead to diabetes. This would be helpful in the Medical Sector as having access to data analysis of this problem could lead to better diagnoses of future patients, which could help set up preventative treatments beforehand to stop strain on medical services, but also allow patients to have access to quality healthcare leading to better quality of life, patient satisfaction and survival rate.

Methods

Dataset

This Diabetes Dataset was sourced from Kaggle [^2] and contains data from the National Institute of Diabetes and Digestive and Kidney Diseases. This Dataset was used to predict whether a patient had diabetes. This Classes patients in the Outcome variable as either 0 (Non Diabetic) or 1 (Diabetic). This dataset provides insight into many key features (BMI, Glucose, Blood Pressure etc) that can be a large indicator in the cause of diabetes. By leveraging this dataset using different data analysis techniques and data modelling, we can help identify indicators of the disease that can be used for prevention and management of the disease. This dataset has missing values , so will need to be preprocessed to ensure valuable insight can be obtained.

[^2]: Diabetes Dataset (Kaggle)

Analysis and Results

Dataset Preparation

This Dataset had to be prepared for Usage in Data Analysis by cleaning up any null / not included ie included as 0 data points. We first did this by checking for null values by using the `isnull()` function on the DataFrame, which returned 0. Then by checking for 0 values it was found that there are lots of missing values. These were turned to Nan using `diabetes_data_cleaned[columns_to_clean] = diabetes_data_cleaned[columns_to_clean].replace(0, np.nan)`. The proportion of values that were now null in each column of the dataset was then calculated and displayed , giving this output:

```
Proportion of missing values in the cleaned dataset: Id 0.000000
Pregnancies 0.000000
Glucose 0.650289
BloodPressure 4.515896
SkinThickness 28.901734
Insulin 48.049133
BMI 1.408960
DiabetesPedigreeFunction 0.000000
Age 0.000000
Outcome 0.000000
dtype: float64
```

Due to this Output , it was decided that 2 rows: **SkinThickness** and **Insulin** would be dropped. This was because imputing thousands of lines to data to something ie the mean would lead to a big skew in data, making the dataset overall unfit for purpose of data analysis as any output would be massive different to real life. **Id** was also dropped as it serves no analytical value for us.

The other Columns with missing data: **Glucose**, **BloodPressure** and **BMI** had the missing columns filled in as they had a lower proportion of missing values, therefore not effecting the data as much. This was done using the Mean(Average) values using the code below, with the example showing how it was done on the **BMI** column:

```
imputer = SimpleImputer(strategy='mean')
diabetes_data_cleaned['BMI'] =
imputer.fit_transform(diabetes_data_cleaned[['BMI']])
```

This was done with the averages as this gives the biggest representation of the overall trend of the column from the data, which will help the replaced values being as accurate as possible.

Overall, this step (Data Preprocessing) has helped the data be more fit for its purpose as it has made the dataset more trimmed and has removed any null/unrepresentative(0) values that may skew any findings.

Exploratory Data Analysis

DELETE Primarily, EDA is for seeing what the data can tell us. Utilize suitable data mining tools and analysis techniques to find significant patterns and trends (SPSS, Excel, Tableau, WEKA, Python libraries, etc.). You do not need to use all the tools, but using Python or a mix of tools will provide you with higher grades. Add histograms, bar charts, etc., to see if you can get any insights from the data, look at the distribution, etc. Look for null values, drop unnecessary columns or rows, find outliers, correlation among variables, clusters, etc. This is not limited to what is in this document; the further you dive, the better it is, and you should document each step. Here you can add another table explaining the new dataset if you wish after making some changes.

Data Modelling and Visualisation

DELETE Identify your independent and dependent variables. Perform analysis such as time series, multivariate, bivariate analysis, linear regression, etc. You can try to explain what you are trying to predict, following the code examples from the class activities. Use appropriate tools to perform some visualization on the chosen dataset. The choice is yours, based on your future intentions of work and also your familiarity with the tool. Your report should document and justify the techniques you have used to mine and analyse the data based on examples from the weekly class activities as well as from your own research.

Evaluation

DELETE After you have performed some modelling and obtained some results, identified any patterns in the data or provided some insights, you can compare your work with existing work from other researchers in a few paragraphs to see what the effects of one variable on another are or among other variables. Evaluate your work critically, explaining what could have been done to support your analysis. What approaches could have been used? Look for models/data cleaning/data pre-processing techniques from the literature to compare your approach and evaluate critically.

Limitations and Challenges

Some of the limitations with this dataset include: Missing values - having missing values in any dataset isn't ideal as it creates a lack of real life, accurate data. We got around this in our data Preprocessing/ Preparation Phase, however 2 variables were dropped from the table during this and other variables having missing values imputed from the average (mean). Due to this, we may lack understanding of how those other variables contribute, which can affect healthcare and also the imputed values may skew any models trained as they are not true organic values gained from study.

Conclusion

DeleteA few lines to conclude on the work and what you have learned. The conclusion can be more like a reflective summary. A reflection paper is meant to illustrate your understanding of the material and how it affects your ideas and possible practices in the future.

Reference List (Harvard Style)

> [1] World Health Organization (2021) Diabetes Fact Sheet. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed: 27 December 2024). [2] Mehmet Akturk (Mathchi) (n.d.) Diabetes Data Set. Available at: https://www.kaggle.com/datasets/mathchi/diabetes-data-set (Accessed: 27 December 2024).