Capstone Project –
Connections between Lifestyle & Disease

Question to answer:

How do various lifestyle choices and chronic diseases correlate with each other?

My initial hypothesis is that there is a correlation between lifestyle choices and diseases.

Background on the subject matter area:

There have been many academic studies analysing the question about whether a certain life style choice is correlated with a certain disease. Lifestyle choices are e.g.: the type of diet, the use of antibiotics, the amount of sun exposure, BMI (body mass index), smoking etc. Usually, as it is common practice in academic research, the correlation between only one lifestyle choice and one disease are being examined.

In my capstone project, I am interested in analysing whether there are correlations between SEVERAL lifestyle choices and a specific disease. Also, I am interested in analysing whether there are correlations between the type of diet and SEVERAL chronic diseases. Data science and machine learning specifically lend themselves well for this purpose, since they allow for a flexible high-dimensional analysis of data; i.e.:

- an analysis of several independent variables simultaneously,

- using a certain parameter (e.g. diet type) either as an independent or a dependent variable, depending on the question at hand.

Details on the source of the data and the dataset itself.

The data set that I am using is from the so-called "American Gut Project" (https://github.com/biocore/American-Gut). They have conducted a questionnaire with people mostly from the US, UK and Australia, in which they asked about their lifestyle choices, incidence of some chronic diseases and other personal data like age, gender and place of residence. The data format is a .cvs file. The original data set contains 4816 rows and 202 columns. One data point corresponds to the information about one study participant. Additionally, participant samples from the gut (in some cases also from the skin and saliva) where examined for the presence of genetic material of bacteria. This was done to establish correlations between the microbiome and diseases studied.

A summary of the preprocessing, feature engineering and any other data cleaning/transformation, and exploratory data analysis (EDA) performed and the motivation and reasoning behind it.

1) Check which columns are relevant to the data analysis. The data set contains information about the participants themselves and about details regarding the chemicals and machine settings used for analysing the microbiome genetic material. Using visual inspection, I have deleted the latter type of columns, since they are not immediately relevant to the analysis, since the occurrence of artefacts due to technical mistakes is low. By doing so I removed about 45% of the columns.

2) Check which rows are relevant to the data analysis. There are about 400 rows that are blanks which I deleted too. That way only columns that correspond to participants remained.

3) After step 2, there where 9 rows with NaN values, I deleted them too, since relative to the data set size that is just a small loss.

4) In order to build a model, I choose to start with a relatively small amount of features that seem relevant to the incidence of a certain disease and then potentially add more features that may improve the accuracy of the model. As the first two disease to study I have chosen diabetes and IBS.

Therefore, I have started using the features of:

- percentage of fat in the diet,
- percentage of carbohydrates in the diet,
- age,
- latitude.

The latitude of the location at which the participant lives is relevant since it correlates with the average sun light exposure throughout the year and hence the Vitamin D3 levels. Vitamin D3 levels have been shown to correlate strongly with the incidence of various diseases.

5) The original data are of the object type. For modelling I have changed the object type into float for all four features used.

6) The features contained several cells saying ‘unknown’ or ‘no_data’. For the modelling, I have imputed numerical values into those cells.

- For the percentage of fat and carbohydrates in the diet, I have entered 33.3%, asuuming that the person has an average diet in which the percentages of the three macronutrients (fat, carbohydrates and protein) are split equally.

- For the missing values of the latitude, I checked if other features give me information about the state or country of that person. Once I got that, I calculated the average value for people from that location and used it to impute the missing latitude value. After imputing, the data type was changed to float for all features.

A summary of all the modelling completed including the process of model evaluation, selection, and results.

Logistic regression models and a decision tree classifier were used to evaluate the predictive power of each feature in terms of the incidence of diabetes or IBS. Both model types result in test accuracies above 90% and the features used indicate that they indeed have predictive power in terms of disease incidence for diabetes and IBS.

Findings and conclusions based on all analysis and modelling of the data - how do your results compare against your initial goals & hypotheses?

My initial hypothesis is that there is a correlation between lifestyle choices and diseases, which has been confirmed through the analysis.

A final summary of the business applications of the project as well as potential next steps and future directions.

Future work will be to add more features to the models, improve the model recall and precision and use the diet types as the targets, not just diseases. Potentially, one could add the information about the bacteria found in the gut samples, to analyse correlations between the microbiome and disease incidence.