Group C7 report

Project title: HEART-DISEASE
Team members: Karl Gustav Gailit, Kertu-Carina Kallaste, Markus Kikkatalo
Repository: https://github.com/kertucarina/IDS2021-project/blob/main/README.md
Datasets:
https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv
https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset

Task 2. Business understanding

1. Background

1.1 Problem

Our business (US healthcare system) wants to address that heart disease can be diagnosed from certain health indicators.

Heart disease is the leading cause of death (1 in every 4 deaths) and it costs the US government hundreds of billions of dollars annually (368Bn in 2016-17). Large investments have gone into curing different heart problems, but as the number of cases have risen to new heights, a new approach must be taken.

Commision members believe a prevention system, which notifies the patient that he's in a risk group, would make the patient start making steps on taking care of his health.

1.2 Business goal

The business goal is to reduce the annual death rate caused by heart diseases by 5 percent.
A prevention system will notify the person and his family doctor, if one's health indicators have increased the likelihood of the person getting a heart disease. For that, regular check-ups must be done to keep the data fresh. Thanks to a decent amount of data from the previous years, we can make a model, which predicts from the input, if the subject is prone to heart disease or not.

1.3 Business success criteria

The business wants to reduce the amount of people having cardiovascular disease by 5 percent year over year.

2. Assessing your situation

2.1 Inventory of resources:

We have anonymous data gathered by the Behavioral Risk Factor Surveillance System (BRFSS) over a span of 5 years of people with different health indicators and their history with heart disease. The dataset is imbalanced (~10% percent of surveyors had a history with heart disease) and some sampling will probably be done to get more accurate results.

Our team consists of three motivated members, who will analyze the data and generate a model for the outcome.

2.2 Risks and contingencies:

Due to our small team size, every member's current health is important and due to our current pandemic situation. If one member should fall out due to health problems, our final product will not be as complete as it could have been. Due to job shortage, we will have to take that risk and take preventive steps to keep ourselves healthy.

Because our team has other projects to tackle at the same time (other courses), then our job will be done in short sprints over the given time. This risk can be alleviated by proper time management and good communication.

Because the data is quite imbalanced (target group is just 1/10 from total), we need to be attentive when training our data to receive correct outcome

2.3 Terminology

Data - values collected through record keeping or by polling.

Learning - training models (estimating their parameters) based on existing data.

Overfitting - a tendency of some modeling techniques to assign importance to random variations in the data by declaring them important patterns.

Oversampling - Duplicating samples from the minority class.

Precision - fraction of relevant instances among the retrieved instances.

Formula - tp / (tp + fp)

Prevalence - the measure of how often the collection of items in an association occur together as a percentage of all the transactions.

Recall - fraction of the relevant documents that are successfully retrieved.

Formula - tp / (tp + fn).

Significance - a probability measure of how strongly the data support a certain result (usually of a statistical test).

Test data - a data set independent of the training data set, used to fine-tune the estimates of the model parameters (i.e., weights).

Training data - a data set used to estimate or train a model.

Undersampling - Deleting samples from the majority class.

Validation - the process of testing the models with a data set different from the training data set.

2.4 Costs and benefits:

Benefits:

- Notify and potentially save thousands of people from having a cardiovascular disease.
- Save billions of dollars annually thanks to risk group persons, who will start to take more care of themselves.

Costs:

- 180 Work hours of three university students - 0 euros.
- Gathering data - 0 euros (already collected by BRFSS)

3. Defining your data-mining goals

3.1 Data-mining goals

Processed datasets - https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset

Deliverables - Model that can predict if a subject is likely to have cardiovascular disease.

3.2 Data-mining success criteria

We want our model to have a 90% recall and precision rate of predicting a subject is in a risk group.

Task 3. Data understanding

1. Gathering data

Our project is based on a previously existing Kaggle dataset and thus this part of this homework task is based more on explaining how the pre-existent data suits our goal of finding out potential causes of heart disease in humans.

1.1. Outline data requirements

In order to analyze potential causes of heart disease there needs to be data about people who have and who have not experienced heart disease. The dataset needs to include a true-false, or boolean, value of whether or not the person has experienced heart disease. Other important values are the person's age, sex, BMI; whether they exercise or not; whether they smoke, drink and/or do drugs; whether they have a genetic risk of heart disease and many other personal details that may or may not cause heart disease.

1.2. Verify data availability

It is impossible to get accurate information about every single person. The best method to gain information about people's health is by doing interviews. That way you will get data from both, people who have and people who have not experienced heart disease. Using this method there are however multiple issues. People may have lied during their interviews, for example, they are afraid of legal repercussions of using illegal drugs and thus claim they do not use them. Another issue is that people who experience or have experienced heart disease will either describe their current state, which could be much different from their lifestyle and physical than when they experienced heart disease, or they would have to describe themselves in the past, leading to inaccuracies caused by imperfect memory.

An American health-related survey conducted annually via telephone interviews by the Centers for Disease Control and Prevention (CDC) named The Behavioral Risk Factor Surveillance System (BRFSS) has collected such information since 1986. Every year around 400 000 Americans participate in this survey. Thus, the data required for this project does exist.

There also exists a dataset based on the BRFSS responses from 2015 for specifically analyzing potential heart disease causes. While the full BRFSS dataset has 330 columns, the Heart Disease Health Indicators (HDHI) dataset contains only 22 columns, one column being whether or not the person has experienced heart disease. The other 21 columns are informative such as the person's sex, whether they been told they have high cholesterol or that they have had a stroke, whether they are smoker and their BMI. The full BRFSS dataset is more granular: for example, while in the HDHI dataset all different kinds of heart disease are put in one column, the full dataset has a column for each disease. The large amount of columns would thus need to be trimmed down, which is the topic of the following section.

1.3. Define selection criteria

The smaller HDHI dataset contains 22 columns of fully relevant information, the columns are described in the relevant Kaggle entry: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset.

The full BRFSS dataset contains 330 columns of data that are described in the following PDF document: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.

Many of the BRFSS dataset columns are unnecessary for this project. Such columns are, for example, the columns regarding outpatient rehabilitation, as they are not preventative and are relevant for about 1000 of the surveyed people out of 440 000 (based on the 2015 survey results). In general, columns where the majority has not responded can not be used as the amount of potentially useful rows will not end up being large enough for analysis. We will also remove columns regarding information about the interview itself as they certainly do not contain relevant information.

As the goal of the project is to find potential causes of heart disease, we do want to keep columns that at first glance might not contain relevant information. For example, income might not directly influence a person's health, but it does give information whether or not they could struggle financially and thus have more stress or whether they have enough money to use for medicine, which is especially important in the United States of America.

2. Describing and exploring data

The full description of the BRFSS dataset is available on the CDC website: https://www.cdc.gov/brfss/annual_data/annual_data.htm. For each year, the description is in a file called the codebook. The URL for the 2015 survey codebook is the following: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf. Due to the length of the list, all of the usable variables will not be listed here.

The description of the HDHI dataset, which we would use in its entirety, is available on Kaggle: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset.

3. Verifying data quality

There is a lot of data available for us to use. At a first go-through of the BRFSS dataset, there seem to be no major quality issues. There were several columns where the majority of answers were "Not asked or Missing" and thus could not be used for this project, but they are removed from the final dataset. If the answer to a question is among the lines of "Refused", we will keep this answer as a refusal to answer could contain as much information as any other answer.

Task 4. Planning the project

A list of tasks and methods we will be using:

1. Getting the data and setting up - can be done by one person (approx. 30 minutes)
    - Downloading the dataset.
    - Creating a IPYNB notebook.
    - Reading in the dataset.
2. Analyzing the data - 10 hours each (330+ features in the dataset!)
    - Gaining a deeper understanding of the features (columns).
    - Selecting only the features which could be relevant and of interest to us.
3. Cleaning up the data - 4 hours each
    - Renaming the columns so that feature names would be more intuitive to understand.
    - Modifying the values to be more suitable for ML algorithms (e.g using one-hot encoding).
    - Removing outliers.
4. Predictive models - 6 hours each
    - Choosing the appropriate models from the ones covered throughout the course.
    - Finding new ML models which could be of use.
    - Splitting the dataset into training and testing sets.
    - Training the chosen models.
    - Assessing the performance of the models.
5. Finding features which have the strongest correlation with heart disease - 5 hours each
    - For this task we will be using computational statistics.
    - Statistical hypothesis testing (Welch’s t-test, permutation test, …)
6. Presentation - 5 hours each
    - Gathering the conclusions.
    - Writing a report on our findings.
    - Designing a visually appealing poster.