B12 REPORT

Heart Failure Prediction

TASK 1

Setting up

Link to the repository - https://github.com/hardopost/KHR-Heart

Group members - Hardo Post, Karl-Magnus Laikoja, Risto Voor

TASK 2

Business understanding

Background

Our project belongs into medical domain with the task of predicting heart disease of a patient. Heart disease refers to any condition affecting the cardiovascular system. There are several different types of heart disease, and they affect the heart and blood vessels in different ways. Heart disease generally, remains the leading cause of death globally, according to the annual heart disease and strokes statistics update from the American Heart Association. It is necessary to detect early, if a person has already developed a heart disease or not, so treatment or life-style change can be implemented to avoid possible heart disease caused death.

Business goals

A predictive machine learning algorithm can be trained by existing data and diagnosis already made by doctors with a purpose of predicting a categorical label, if a person with specific features has a heart disease or not. Such algorithm could reduce the work for doctors reviewing all cases and maybe doctors will have to look only cases that algorithm has predicted to have heart disease. If that kind of assistant to doctor works well, it may be possible to call more people for testing in government programs and reduce potential heart disease caused deaths.

Our project has a data set of 918 rows where each row is info about one patient. Each row contains 12 features about the patient. First two are age and sex. Following 9 features that are more directly related to heart condition. The last feature shows if a patient has heart disease or not.

More specifically what we hope to achieve with current data:

1. Train a model that predicts based on dataset features and instances if a patient has heart disease of not.
2. Find out which features have the highest correlation to heart disease.
3. Find out separately among male and female patients which features have the highest correlation to heart disease and compare if they are the same or different.
4. Find out which features have the highest correlation to heart disease in different age groups (e.g., 20-29, 30-39, …), but not take age itself into account. Find out if there are features that patients can directly control.

Business success criteria

At current stage success would be if we get a working predictive model on current data with high accuracy, with the accuracy score of 0.90 or more. It is especially important to keep low false negative rate on the predicted results, so the predictions wouldn't label patients with actual heart disease with no disease. If a model is working on a certain data set, doesn't mean it work on all data sets, but a working model creates trust and basis for similar ideas to be used in future data sets.

Assessing your situation

Inventory of resource

For this project, we got the data set from Kaggle and it is already in quite acceptable state for analysis. If additional info is required, our potential sources would be health care institutions all around the world.

Requirements, assumptions, and constraints

At this stage there are no legal or security requirements that would impede our project to continue. Project is well in schedule to be completed by 16th of December 2021.

Risks and contingencies

Currently there are no known causes that can cause delay in project completion. Our project has 3 members, in current situation there is a possibility that some team member can catch a virus and may not be able to work a few days, but then it is possible to divide work among other team members.

Terminology

As we obtained the data from Kaggle in a single csv file, we have not used any data mining in our project, therefore have no terms to declare that are related to data mining.

Costs and benefits

The costs of the project are minimal. Only costs are the work hours of 3 project members put into this project and it can be viewed as an alternative cost to some other work that they could have done during the time spent on this project and earned money, but it is difficult to evaluate in euros. Benefit is for the team members to get experience from this project and realize the experience in later stages of school and work life and possibly earn some currency. The risk-benefit analysis favors continuing with the project.

Defining your data-mining goals

Data-mining goals and Data-mining success criteria

We didn’t use any proper data mining for getting info for our project. Google search was used, and different websites www.stat.ee, www.who.int among others, but our goal was to find data that suited for training a machine learning algorithm that could predict a categorical label or perform a regression analysis. It was difficult to find suitable data on our own, therefore we chose Kaggle data set.

TASK 3

Gathering data

Data requirements

In order to accomplish our main goal of creating a working model which predicts whether a patient has heart disease or not based on a number of medical features, we will need a dataset of patients with their information regarding these features and whether or not they have heart disease. Preferably this dataset will be in CSV format for easier use. For this, we have decided to use a dataset from Kaggle which is in accordance with all of our requirements.

Verify data availability

Kaggle datasets are publically available and easily downloadable, so there are no problems with data availability.

Define selection criteria

The specific data source that we will use is the heart.csv file found within the Kaggle dataset. Within the file, every row and column is relevant to our project.

Describing data

Within the heart.csv file there are 918 rows, each row containing a unique patient's data. There are 12 columns, each column containing the value for a medical feature regarding the patient on that row. These features are:

- Age: the patient's age.
- Sex: the patient's sex.
- ChestPainType: the type of chest pain that the patient is experiencing. ATA (Atypical angina), NAP (Non-anginal pain), ASY (Asymptomatic - Silent (asymptomatic) myocardial ischemia (SMI)) or TA (Typical angina).
- RestingBP: the patient's resting blood pressure in mm/Hg.
- Cholesterol: the patient's total serum cholesterol in mg/dl.
- FastingBS: the patient's fasting blood sugar level, 1 if over 120 mg/dl (considered high), otherwise 0.
- RestingECG: the patient's resting electrocardiogram results, Normal (normal), ST(having ST-T wave abnormality) or LVH (showing probable or definite left ventricular hypertrophy by Estes' criteria).
- MaxHR: the maximum heart rate achieved for the patient.
- ExerciseAngina: whether the patient has exercise-induced angina, Y(yes) or N(no).
- Oldpeak: a line in EKG doesn't come back to zero state, where it should come. ST [Numeric value measured in depression] ST depression induced by exercise relative to rest.
- ST_Slope: the slope of the peak exercise ST segment (the shape of the line in a certain place in EKG). Up (upsloping), Flat (flat) or Down (downsloping).
- HeartDisease: whether the patient has heart disease or not, 1 for yes, 0 for no.

These features are enough for us to create our prediction model.

Exploring data

The most obvious data quality problem is the fact that in this dataset there aren't a lot of patients. 918 is a pretty low number of instances to create a good model, and considering that we also want to be able to find the features that are the most correlated to heart disease separately among male and female patients and within different age groups, then the number of instances in each of these groups can get really small. Also, the number of patients with heart disease is a lot smaller than the number of patients without heart disease. A model created with such skewed data would not be accurate, which is why we have to prepare the data by balancing it.

Verifying data quality.

The data we need exists, we can have it and there aren’t any fatal data quality issues. With proper data preparation, all the quality issues can be fixed and creating a good model will be possible. This single dataset is good enough to accomplish our goals.

TASK 4

Planning your project

List of tasks

- Preparing data - balancing data, transforming features. This will give us a dataset that we can use for all our other tasks. It will be the first task that has to be completed, before we can move on to other tasks. All of our group members will be participating in this task and this should take about 5 hours.
- Training a model that predicts whether someone has heart disease or not. Estimated time for this task is 15 hours and done by Risto Voor.
- Finding out which feature has the highest correlation to heart disease. Estimated time for this task is 5 hours and done by Karl-Magnus Laikoja.
- Finding out which feature has the highest correlation to heart disease in men and women. Estimated time for this task is 5 hours and done by Karl-Magnus Laikoja.
- Which feature has the highest correlation to heart disease in different age groups. Whether the patient can control this feature. Estimated time for this task is 15 hours and is done by Hardo Post.
- Making a poster. All of our group members will be participating in this task and this should take about 5 hours.

Methods and tools used

Tools we will be using:

- Python
- Jupyter Notebook
- Pandas
- Matplotlib
- Numpy
- Sklearn

Methods we will be using (during the course of the project):

- DecisionTreeClassifier
- KNeighborsClassifier
- RandomForestClassifier
- OneHotEncoder
- Cluster