**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 cvs 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 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](http://www.stat.ee), [www.who.int](http://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.