Data Mining Team Project

Proposal

presented by
Marie-Christin Häge, 1913888
Finn Hülsbuch, 1913864
Thilo Dieing, 1692328
Lasse Lemke, 1914420
Eric Jacquomé, 1903834

submitted to the
Data and Web Science Group
Prof. Dr. Right Name Here
University of Mannheim

October 2022

Chapter 1

Proposal

1.1 What is the problem you are solving?

Diseases of the circulatory system are the most common cause of death in Germany. Most of these are heart diseases. A faster and more accurate identification of these will result in immediate and more precise treatment, and could thereby save many lives.

Our goal is to develop an algorithm that will support medical staff in making more data-based decisions related to heart diseases. This will be achieved by identifying the key factors playing a role in determining whether a patient has a heart disease.

Consequently, through the gained insights the quality and accuracy of choosing the correct treatment and deciding on the fitting measures will improve. Furthermore, analyzing the key factors individually can also lead to recommendations for the identification process of heart diseases.

1.2 What data will you use?

We will use the Heart Disease Data Set. The dataset was created in 1988 by combining datasets collected in Budapest, Zurich, Basel and Cleveland. The dataset consists of 75 attributes and has 899 instances. The condition of the patient is labeled as either 0 (no heart disease) or 1 (heart disease).

The dataset is publicly available in the UCI Machine Learning Repository divided by the origin of the data. We will obtain the data by downloading the dataset and combining the different files of the folder.

1.3 How will you solve the problem?

1.3.1 Preprocessing

To solve the problem, the first step is to preprocess the data. Therefore, multiple different aspects need to be checked and adjusted if necessary. Depending on the gained insights of the data, the steps might vary while executing the preprocessing.

The first step will be to transform the raw data into a usable format.

Then, the data will be analyzed and the usability of the data set will be increased through different actions. First, the missing values will be replaced with NaN to facilitate the analysis. In addition to this, a validity test, a test for outliers and a test for duplicate entries will be conducted. Based on the provided result, rows or columns will be dropped that do not pass these tests. Furthermore, features with constant values will be dropped. Next, a check for inconsistencies between features and in the format of their values will be done. If there are inconsistencies between features, a decision needs to be made on what feature will be used and which will be dropped. If there are differences in the format, a standard format will be chosen and consequently adopted in the data set. Moreover, feature extraction and combination will be addressed. The next step of data preprocessing will be data normalization. For that, different transformers and scalers from the sci-kit learn library will be used. Also, feature binning will be used where needed.

Before the evaluation a feature reduction will be done to decrease the number of features and to increase the significance of the features. Therefore a correlation analysis is required to select the most significant features. Furthermore, this could also decrease the amount of required computing power and evaluation time.

1.3.2 Algorithms

To find the best results, multiple algorithms will be compared. Therefore, it is planned to use algorithms presented in the lecture (K-nearest neighbor, Naive-Bayes, Decision Tree). Additionally the following algorithms will be used for comparison:

- Support vector classification (linear, poly, rbf, sigmoid, precomputed)
- XGBoost
- CatBoost
- AdaBoost
- RandomForest

- IsolationForest
- QuadraticDiscriminantAnalysis

To evaluate the different algorithms a simple cross validation will be applied. The best algorithms are then tuned with different hyperparameters using nested cross validation and compared regarding their scores, costs and explainability.

1.4 How will you measure success?

Die Methode der Klassifizierung wird entsprechend der Datengewichtung nach der Datenaufbereitung ausgewählt. In den Rohdaten liegt ein Verhältnis von 404 Gesunden zu 495 Erkrankten vor. Ziel allgemein ist es, möglichst akkurate Vorhersagen zu treffen. Allerdings wird das Nichterkennen einer Krankheit als schwererer Fehler im Verhältnis zu einer falsch positiven Diagnose angesehen.

1.5 What do you expect your results to look like?

It is assumed that it is possible to identify diseases on the basis of other symptomatology, as this can already be achieved in medicine today through targeted interrogation.

Based on our models, we want to identify a minimal set of features that can be used to make a diagnosis about heart diseases with the highest possible accuracy. This could lead to a more efficient diagnosis of heart diseases and would also decrease the amount of medical tests and examinations needed beforehand. Not only would this benefit medical facilities as time and money can be saved but the results should also provide helpful insights to the patients. With the help of our results, patients should be able to self-diagnose the chance of a heart disease based on personal information and their health data. Therefore, patients have the opportunity to react to this and maybe set precautionary measures to decrease the chance of a heart disease.

Bibliography