Data Mining Team Project

Proposal

presented by
Club der toten Dichten (Team 12)
Marie-Christin Häge, 1913888
Finn Hülsbuch, 1913864
Thilo Dieing, 1692328
Lasse Lemke, 1914420
Eric Jacquomé, 1903834
Timotheus Gumpp, 1913876

submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim
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 (Statistisches Bundesamt, 2020). 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 lives.

Our goal is to develop a ML-Model that will support medical staff in making more data-based decisions related to heart diseases. This will be achieved by identifying the key features playing a role in determining whether a patient has a heart disease or not.

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 a Heart Disease Data Set that 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 University of California, Irvine (UCI) Machine Learning Repository (Janosi et al., 1988) divided by the origin of the data. We will obtain the data by downloading the dataset and combining and cleaning the different files of the folder. It is planned to not use the 14 commonly used features that are chosen by the UCI but instead gain insights into all available features and

combine the data into a new feature set that then is used to train models. The success of this approach can be measured by comparing existing models trained on the 14 features against our models.

1.3 How will you solve the problem?

It is planned to solve the problem in multiple steps. These steps are described in more detail in the following.

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 and replace missing values with NaN to facilitate the analysis. Then, the data will be analyzed and the usability of the data set will be determined by looking for outliers and duplicates. In addition, a validity test is conducted. Based on the gained result, rows or columns will be dropped that do not seem usable. Next, an analysis for inconsistencies and correlations between features will be conducted to decide what features should be used or 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 last step of the data preprocessing will be data normalization. For that, different transformers and scalers from the sci-kit learn library will be used.

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?

The evaluation method to select the best algorithm is selected according to the weighted data after the data preparation. In the raw data, there is a ratio of 404 healthy individuals to 495 diseased individuals. This would favor the use of the f1 score. However, it is not possible to predict whether this will still be the case after the data processing. Furthermore, the model metrics will be used for comparison between the trained models. In general, the goal is to predict as accurately as possible. Nevertheless, a non-recognition of a disease is considered as more severe than a false positive diagnosis. Consequently, a penalty weighting against false negatives would be justified.

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 questions.

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 reliability. This could lead to a more efficient diagnosis of heart diseases and would also decrease the amount of medical tests and examinations needed beforehand. Consequently, the human error rate could be decreased as well. 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.

Furthermore, the goal is to compare our results to the preprocessed data set with the 14 features and analyze how the two results differ. Therefore, in the end the usability of our models can be assessed based on the accuracy and other measures.

Bibliography

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. UCI Machine Learning Repository: Heart Disease Data Set, 1988. URL https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Statistisches Bundesamt. Gestorbene: Deutschland, jahre, todesursachen, geschlecht, 2020. URL https://www-genesis.destatis.de/genesis/online?operation=previous&levelindex=0&step=0&titel=Tabellenaufbau&levelid=1665392693797.