Analysis of Heart Disease Data

Project Report

presented by
Club der toten Dichten (Team 12)
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

December 2022

Contents

1 Application Area and Goals **1**

2 Structure and Size of the Dataset **1**

3 Preprocessing **3**
3.1 Cleaning 3
3.2 Hyperparameter Tuning 4

4 Datamining **6**

5 Results **6**

1 Application Area and Goals

Heart disease is currently still one of the highest causes of mortality on earth (Nahar et al., 2013; Kavitha and Kannan, 2016; Statistisches Bundesamt, 2020). Given the successful application of data mining in other sectors e.g. banking and finance or marketing (Keleş, 2017) possible applications in the medical industry are plentiful. Yet the healthcare sector is information rich but knowledge poor (Soni et al., 2011). According to Soni et al. (2011) medical data sets provide great potential for data mining to be used in clinical diagnosis.

The aim of this project was the application of data mining methods, more specifically classification methods, to predict whether or not a patient could suffer from a heart disease. The successful application could help doctors and medical staff with diagnosing patients by automatically analysing historical test result data of the patient and give a prediction when a higher potential of heart problems arise. By doing this analysis patients flagged for potential heart disease could possibly be prioritised. Due to the immense amount of stress and long working hours medical personal are facing, having a standardized scheme looking at the data could be beneficial. In the past such approaches have already been tested and proven to be a good diagnostic option (Usha Rani, 2011). Jabbar, Deekshatulu, and Chandra (2013) state that data mining techniques answer several important and critical questions related to healthcare and that they can improve the provision of quality services to patients.

Denke, dass der konkrete Fall eher der ist, dass wir günstigere/simplere Mehtoden nehmen und dann eine Vorhersage treffen, wenn die positiv ist, dann machen wir komplexere analysen(mrt mit kontrastmittel)

This project report is based on the "Heart Disease Data Set" (Janosi et al., 1988) which, despite its age is still relevant given the fact that it consists of results of medical tests. In addition to that the validity is assumed because it is frequently used in contemporary research (see Nahar et al., 2013; Usha Rani, 2011; Aha and Kibler, 1988).

2 Structure and Size of the Dataset

According to the CRISP-DM reference model, Data Understanding begins with the initial data collection. For our problem, we use a Heart Disease Data from 1988, consisting of 77 attributes and 899 instances resulting in 69.223 observations. To create our customized dataset we combined multiple datasets collected in Budapest, Zurich, Basel and Cleveland as seen in table 1. This dataset was made available by the University of California, Irvine (UCI) Machine Learning Repository under the Creative Commons Attribution 4.0 International License (Janosi et al., 1988). Despite the age of the dataset, it is still suitable for analysis due to the comprehensive and detailed data collection.

Fit table to width

Index	Publisher of data set	# of instances	Distribution o
1	Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.	294	106 / 188
2	University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.	123	115 / 8
3	V.A. Medical Center, Long Beach	200	149 / 51
4	Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.	282	125 / 157

Table 1: Content of the dataset

After the creation of the dataset, we perform initial analysis of the dataset. The attributes are highly diverse and oftentimes describe specific medical information. Common attributes include age, sex or patient ID (id), whereas medically specific attributes focus on information like type of chest pain (cp). An overview of all attributes is provided in our code documentation as well as the UCI Machine Learning Repository.

These medical-specific attributes have presented us with challenges. Although the attributes are listed in the information provided by UCI, the explanations are very brief or non-existent.Thus, it became clear our team lacked specialized knowledge to interpret attributes such as type of chest pain (cp) or resting electrocardiographic results (restecg). Therefore, we researched the different attributes in order to understand their meaning on the one hand and to be able to interpret their values better on the other hand. This research served as a foundation for assembling the data frame for our model in the Data Preparation.

Our target variable is resembled by the attribute num and encoded binary, with the value 1 indicating the diagnosis of a heart disease and the value 0 contradicting that indication. Looking at the distribution of the target variable we observe strongly varying distributions for the different datasets as specified in Table 1. The distribution in the combined dataset is relatively equal with 495 positive and 404 negative measurements, meaning a disease prevalence of roughly 55,1%. When examining distribution of the other attributes we also notice uneven distributions. For example, the attribute sex´s distribution contains 78% male and 22% female patients in our data set, contributing to the Gender Data Gap prevalent in medicine. We also notice further uneven distributions across our datasets. Whilst highlighting them in the report would exceed the frame, we address these deviations in the preprocessing of our dataset and documented them.

It should also be noted that the data set is not sorted or aligned to a time series and therefore not suitable for a time series analysis.

Lastly, verifying data quality also acts as an important part of Data Understanding. When checking for missing data we observed multiple interesting results. We observe a generally high number of cells with missing values, lacking 21397 or 30,8% of values. Some individual attributes in particular have many missing val-

ues like history of diabetes (dm) with roughly 90% missing values. These missing values are sometimes dataset-specific (e.g. painloc is mainly missing in the Cleveland dataset) but also cross-dataset (e.g. dm).

Furthermore, when checking whether meanings of attributes and their contained values fit together, we observe some irregularities. For example, looking at the attribute cholesterin (chol), the values listed seem to be unrealistic (e.g. many instances with 0 serum cholesterol in mg/dl).

After obtaining an understanding of properties and meaning of our data, we perform initial analysis to explore additional inisights of our dataset

3 Preprocessing

We approach preprocessing in two steps. In the first step we clean the data set based on knowledge we obtained in section 2 and further analysis. Secondly, we transform the data using hyperparameter tuning.

data set vereinheitlichen

3.1 Cleaning

Firstly, we implement a custom loading function to transform the four datasets into *CSV* format, so we can use it further on.

We remove the false predictors *lmt, ladprox, laddist, diag, cxmain, ramus, om1, om2, rcaprox* and *rcadist* as our target variable *num* is a combination of these according to the UCI.

Willst du ggf noch die anderen erwähnten features so machen wie die hier, dann ist das alles etwas übersichtlicher.

Furthermore, the features *thalsev, thalpul, lvx1, lvx2, lvx3, lvx4, lvf, dummy* and *junk* are considered irrelevant or are not described by the UCI, so we also drop them. Other irrelevant attributes we remove are IDs (*id*), constants (*ccf, name, earlobe, restckm, exerckm*) and dates of medical examinations (*ekgmo, ekgday, ekgyr, cmo, cday, cyr*). We consider these dates irrelevant because we assume that the date of an examination does not affect its outcome.

We drop the feature *pncaden* because it is the sum of *painlox, painexer* and *relrest* and therefore contains no additional information.

The features *cp, restecg, slope, ca* and *restwm* were oneHotEncoded as they represent categorical values.

When checking for inconsistencies between features, we detected that *thaltime* is sometimes lower than *thaldur*. As *thaltime* describes the moment a measurement is taken within the exercise, it has to be lower than the duration of the exercise *thaldur*. We replace *thaltime* by `NaN` in all 17 instances that do not satisfy this criterion.

Also, the maximum heart rate (*thalach*) was replaced with `NaN` if it was lower than the heart rate at rest (*thalrest*).

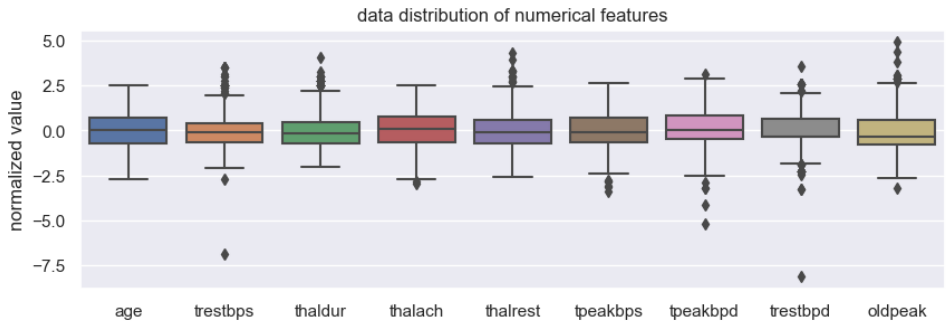

Figure 1: data distribution of all numeric elements

As shown in figure 1 we created a normalized box plot of all numeric features to check for outliers. The features *trestbps* and *trestbpd* show extreme outliers with a value of 0. These are assumed to be incorrectly specified `NaN`s and are therefore replaced by `NaN`. All other outliers are not as extreme and come in groups. As the data contains sick persons, values diverging from the norm are expected. For these reasons we decided to keep these outliers as they can be a strong indication of a heart disease.

The remaining features were analysed regarding their pearson correlation. Only two pairs of features with substantive amount of data (<75% `NaN`s) have a very strong correlation (>80%). These are *cp_4* $\leftrightarrow$ *painexer* and *rldv5* $\leftrightarrow$ *rldv5e*. The highest correlation is between *cp_4* and *painexer*. The feature *cp_4*, which was oneHotEncoded from the categorical variable *cp*, describes whether the patient has no chest pain at rest. Painexer describes whether the patient only has pain when exercising.

Concluding from the high correlation between the EKG amplitude when resting (*rldv5*) and the EKG amplitude when exercising (*rldv5e*), we decided to create a new feature (*rldv5_diff*) by using the difference between these. We did the same with resting heart rate and maximum heart rate (*thal_diff* = *thalach* - *thalrest*).

Furthermore, we enrich the feature *smoke* using *years* and *cigs*. Hereby, we reduce the number of `NaN`s from 74% to 43%.

3.2 Hyperparameter Tuning

Additionally to the hyperparamter tuning of the estimators, which is described in section 4, we also optimize which method is used with which hyperparameters in

the preprocessing steps.

Firstly, we try binning the feature *age*. We choose either 2 or 5 bins or no binning at all. We decided to use equal width binning so that the age groups are simpler and more intuitive to a doctor.

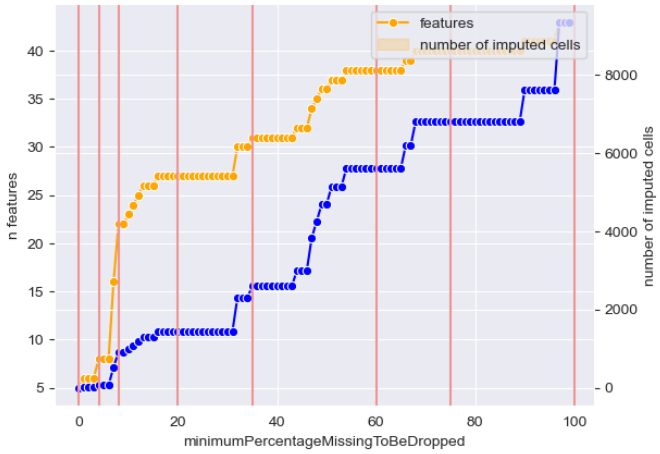

Figure 2: Number of features and values to be imputed by number of `NaN`s

Figure 2 shows the number of features, that have less than a certain number of missing data and how many cells we would need to impute if we included these features. It becomes apparent that there are certain steps where the number of features goes up a lot. To decide when a feature is included based on the number of missing values, we try the steps 0, 4, 8, 20, 35, 60, 75 and 100 % in the model. They are shown as vertical lines in the graph. Additionally, we decided to drop features based on their correlation. For this we decided to use the steps [X,X,X,X]. Insert steps

To impute the missing data we use a simple imputer. Missing values are replaced by the mean, median or mode of the feature. We decided against using a KNN imputer, because it is computationally much more expensive. The iterative imputer is not used, as it is still experimental and therefore subject to change.

To account for the different ranges of the features different scalers are tried out. We only use scalers that are applicable for all floats as some features contain negative values. We compare the MaxAbsScaler, MinMaxScaler, PowerTransformer, RobustScaler, Standardscaler and Normalizer. As the hyperparameters of most scalers turn on or off core functionalities of the scaler, we decided to only tune the hyperparameter norm of the Normalizer with the norms l1, l2 and max.

To account for the different amounts of healthy and unhealthy patients we try oversampling and undersampling in the training data in comparison to passing the

values through.

4 Datamining

5 Results

References

Nahar, Jesmin et al. (2013). "Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females". In: *Expert Systems with Applications* 40.4, pp. 1086–1093. ISSN: 09574174. DOI: 10.1016/j.eswa.2012.08.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S095741741200989X (visited on 11/07/2022).

Kavitha, R. and E. Kannan (2016). "An Efficient Framework for Heart Disease Classification Using Feature Extraction and Feature Selection Technique in Data Mining". In: *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*. 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS). Pudukkottai, India: IEEE, pp. 1–5. ISBN: 978-1-4673-6725-7. DOI: 10.1109/ICETETS.2016.7603000. URL: http://ieeexplore.ieee.org/document/7603000/ (visited on 11/06/2022).

Statistisches Bundesamt (2020). *Gestorbene: Deutschland, Jahre, Todesursachen, Geschlecht*. URL: https://www-genesis.destatis.de/genesis/online?operation=previous&levelindex=0&step=0&titel=Tabellenaufbau&levelid=1665392693797&acceptscookies=false#abreadcrumb (visited on 10/10/2022).

Keleş, Mümine KAYA (2017). "An Overview: The Impact of Data Mining Applications on Various Sectors". In: p. 5.

Soni, Jyoti et al. (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". In: *International Journal of Computer Applications* 17.8, pp. 43–48. ISSN: 09758887. DOI: 10.5120/2237-2860. URL: http://www.ijcaonline.org/volume17/number8/pxc3872860.pdf (visited on 11/07/2022).

Usha Rani (2011). "Analysis Of Heart Diseases Dataset Using Neural Network Approach". In: *International Journal of Data Mining & Knowledge Management Process* 1.5, pp. 1–8. ISSN: 2231007X. DOI: 10.5121/ijdkp.2011.1501. URL: http://www.aircconline.com/ijdkp/V1N5/0911ijdkp01.pdf (visited on 11/06/2022).

Jabbar, M. Akhil, B.L. Deekshatulu, and Priti Chandra (2013). "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm". In: *Procedia Technology* 10, pp. 85–94. ISSN: 22120173. DOI: 10.1016/j.protcy.2013.12.340. URL: https://linkinghub.elsevier.com/retrieve/pii/S2212017313004945 (visited on 11/06/2022).

Janosi, Andras et al. (1988). *UCI Machine Learning Repository: Heart Disease Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (visited on 10/10/2022).

Aha, D and Dennis Kibler (1988). "Instance-Based Prediction of Heart-Disease Presence with the Cleveland Database". In: *University of California* 3.1, pp. 3–2.