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

Heart disease is currently still one of the highest causes of mortality on earth (**nahar2013**; **kavitha2016**; **statistischesbundesamt2020**). Given the successful application of data mining in other sectors e.g. banking and finance or marketing (**keles2017**) possible applications in the medical industry are plentiful. Yet the healthcare sector is "information rich but knowledge poor" (**soni2011**). According to **soni2011** medical datasets 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 (**usharani2011**). **jabbar2013** 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 Dataset" (**janosi1988**) 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 **usharani2011**; **aha1988**; **nahar2013**).

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 dataset 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 (**janosi1988**). Despite the age of the dataset, it is still suitable for analysis due to the comprehensive and detailed data collection.

reference to explanation of model

After the creation of the dataset, we perform initial analysis of the dataset. The attributes are highly diverse and oftentimes describe specific medical infor-

add the following sentence??: Furthermore, there won't be systematic change in the medical parameters since the creation of the dataset, as the age of the data is negligible from an evolutionary standpoint.

Fit table to width

Index	Publisher of dataset	# 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

mation. 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 dataset, 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 dataset 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 high number of cells with missing values, lacking 21397 or 30,8% of values. Some individual attributes in particular have many missing values 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

absatz über gruppen von features schreiben; das hatten die extra nochmal betont in der letzten Fragestunde, also gerne auch ein etwas ausführlicherer Absatz

am besten nochmal umschreiben. Ich würde nur beschreiben wie wir vorgegangen sind, nicht dass uns das knowledge fehlt. Finde der Absatz klingt eher nach Nacherzählung

auch noch bissl anders formulieren. Wir haben ja schon Daten drin. Nur gibt es pro patient pro messung nur einen Eintrag; das ist eher das Problem

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

alle wörtlich genannten attribute (z.B. restecg; nicht resting electrocardiographic results) kursiv schreiben: *restecg*

3 Preprocessing

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

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*).

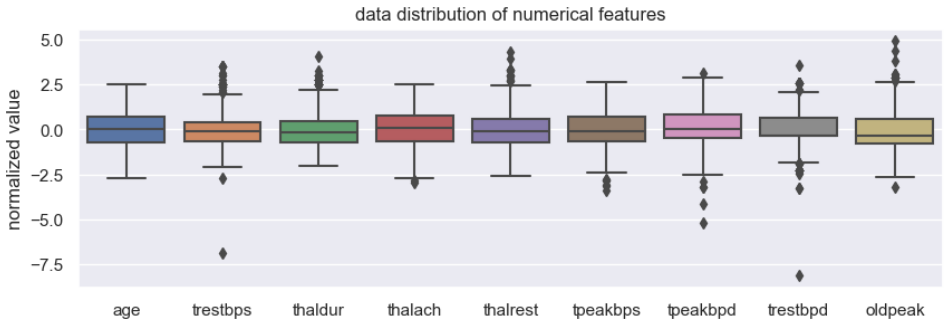

Figure 1: data distribution of all numeric elements

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 (>75%). 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.

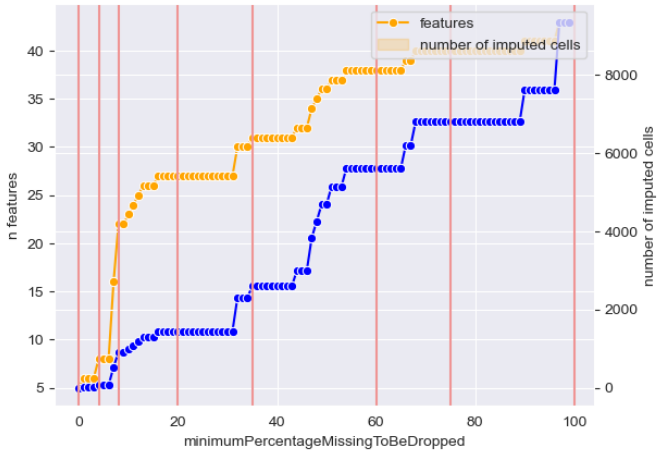

Figure 2: Number of features and values to be imputed by number of `NaNs`

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.

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

For our project we have decided to test ten different classification algorithms. K nearest Neighbors, Random Forest, Decision Tree, Support vector machine, logistic regression, four naïve bayes classifiers based on Bernoulli complement gaussian multinomial and XGBoost.

Satzende für mich nicht ganz klar, da fehlen kommas oder so; zur besseren Lesbarkeit würde ich die 4 NBs ans Ende packen

These ten classification algorithms will be evaluated to determine which one fulfills our criterions best. One criterion is that the model should be as simple as possible like stated by the Occam's razor law. The reason for this is that we want our model to be used universally by doctors but also by patients who may not have great expertise in medical practice. An additional criterion is precision, in order to avoid possible Type 2 errors (a patient with a heart disease is diagnosed as negative) in the diagnosis.

sind das alle kriterien? wenn ja würd ich etwas anders schreiben; so wirkt das als hätten wir hier wahllos 2 unserer Kriterien vorgestellt. Wenn nein dann auch noch die anderen Kriterien kurz auflisten

Besides the before mentioned use of scalers, imputers and samplers we additionally apply a variety of classifier specific hyperparameters. We combine both in a grid search approach while training and testing the models. For the KNN we added two hyperparameters for calculating new instances based on their nearest neighbor. The options were simply taking the majority class or weighting the values by distance. Since KNN relies on some kind of a distance measure we included the options for using the Manhattan distance and the Euclidian distance. Moreover, we tested nearest neighbor values from 2 to 97 (moving in 5-unit steps). Random Forest was tested with numbers of estimators ranging from 10 to 90 (10-unit steps), a maximum depth of null, 2, 6 and 10 and a minimum number of instances per leaf of 2, 6, and 10. Decisions Trees were tested with both the gini index and entropy as impurity measures while the values for maximum tree depth and minimum instances per leaf were the same as for Random Forest. In the case of logistic regression, we added the same distance options as for KNN. For SVC we tested for C values ranging from 120 to 160 (20-unit steps), as well as the gamma values 0.0001, 0.001, 0.01 and the kernel options linear, polynomial, sigmoidal and radial basis function. For the four naïve bayes estimators we applied alpha values from 0 to 19.5 (0.5-unit steps).

For all models we choose a majority class baseline approach. In our case this means that having a heart disease will be the models‘ baseline. Due to the small number of observations in our dataset, we chose to train and test our models by using a nested stratified cross validation with 10-folds, to use as much data as possible for both the training and testing of our models.

Applying all these methods result in a large number of trained models (approx. 45.360.000). In order to be able to still discuss different models we have decided to evaluate the one model for each classification algorithm that comes closest to our criterions. Therefore, for each classification algorithm we chose the one model that

Classifier	Scaler	Sampler	Type 2	Acc.	Conf. interval	Pre.	Rec.	F1	Conf. interval	MPD
KNN	none	none	70	76.29	[76.25, 76.32]	0.78	0.77	0.77	[0.73, 0.80]	0
XGB	Normalizer	RUS	109	76.54	[76.49, 76.59]	0.78	0.77	0.76	[0.70, 0.82]	75
Random Forest	StandardScaler	none	79	78.33	[78.27, 78.39]	0.82	0.79	0.78	[0.70, 0.85]	100
Decision Tree	none	none	73	76.70	[76.67, 76.74]	0.79	0.78	0.77	[0.73, 0.81]	0
SVC	PowerTransformer	none	92	77.87	[77.82, 77.92]	0.80	0.78	0.77	[0.71, 0.83]	20
NB (bernoulli)	StandardScaler	none	111	79.07	[79.02, 79.11]	0.80	0.79	0.79	[0.74, 0.83]	8
NB (complement)	MinMaxScaler	none	78	77.48	[77.44, 77.53]	0.78	0.78	0.78	[0.74, 0.82]	100
NB (gaussian)	Normalizer	none	233	70.46	[70.42, 70.49]	0.76	0.69	0.67	[0.63, 0.71]	20
NB (multinomial)	MinMaxScaler	none	99	76.46	[76.43, 76.50]	0.77	0.77	0.77	[0.73, 0.80]	100
logistic regression	Normalizer	none	75	76.03	[75.99, 76.06]	0.78	0.77	0.76	[0.73, 0.80]	0

RUS = random under sampler, MPD = minimum percentage to be dropped

Table 2: Best models for every classification algorithm

accomplished the highest average level of accuracy, while minimizing the average deviation (size of accuracy confidence intervals) and the number of Type 2 errors. As measure of simplicity, we selected those models that also showed the lowest level of columns to be dropped.

gegenvorschlag letzter Satz: "According to Occams Razor [reference] we favoured models with a low number of columns. Therefore a models with a low number in percentageToBeDropped were chosen.

Table 2 includes these best models and displays those models with their used algorithms, scaler, imputer, sampler, in addition to some prediction metrics.

Before taking a closer look at the model evaluation we first want to view what features are relevant to our models to predict a heart disease. For this purpose, we have visualized the best decision tree model in figure 3. The decision tree uses no scaler nor sampler and uses the simple imputer and as an additional hyperparameters the gini index was used as impurity measure, while the maximum tree depth was set to none, with a minimum instances per leaf value of 2. The total depth of the tree is 5. However, we only display the tree with depth 3 since we aim at focusing on the main attributes. Overall, the decision tree model has an average accuracy of 76.7% and 73 (14.75% of people with heart disease) Type 2 errors. The root node of the tree is whether the participant had asymptomatic chest pain. The node is then split into people who have asymptomatic chest pain and those who do not. The resulting follow up nodes both use gender as the next split. After that the tree splits according to age. Interestingly one can see from this visualization that older men with asymptomatic chest pain have the highest change to be predicted to have a heart disease. Overall, the model predicts men of all ages and with or without asymptomatic chest pain to have a higher probability of having a heart disease compared to women. The group with the lowest chance of having a heart disease are young women with no asymptomatic chest pain.

After having viewed these first preliminary results we now evaluate how well the different classifiers performed. While all models have accuracies between 70%

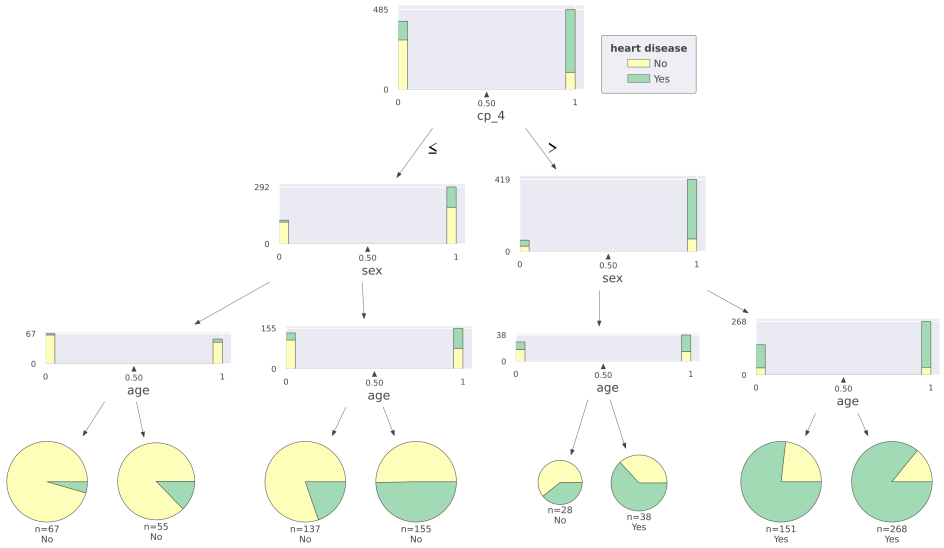

Figure 3: Decision tree visualized

and 79%, with small confidence intervals respectively, they show great discrepancies in both the total Type 2 errors and the minimum percentage to be dropped. The model with the lowest Type 2 error is the one with KNN (Type 2 = 70), the one with the highest uses naïve bayes gaussian (Type 2 = 233), which is also the model that shows the lowest average accuracy. A total of three models have a minimum percentage to be dropped of 0% (only features that do not have missing values are maintained), KNN, Decision tree and logistic regression. The models with the highest minimum percentage to be dropped which is 100% are the ones using the Random Forest, and the naïve bayes models using multinomial and complement.

Based on the different model results and our predefined criterion we argue that the best model for classifying whether somebody has a heart disease or not is the model which uses the KNN classifier, followed closely by the before discussed decision tree model. An argument against this decision would be that KNN are lazy classifiers taking much time for testing. However, this argument renders unimportant because of the size of our dataset which leads the KNN classifier to be comparably fast at labeling new records. The KNN model uses no scaler nor sampler and as an imputer it uses the simple imputer. The hyperparameters were set for one to the Manhattan distance as proximity measure as well as calculating the weighted average instead of only choosing the majority class of the nearest neighbor. Furthermore, the model uses a total of 82 nearest neighbors to predict a new record. Overall, the model does so with an accuracy of 76.28% and like previously

der Satz macht keinen Sinn für mich

discussed with a type 2 error of 70. Due to the good results in the Type 2 errors, the minimum percentage to be dropped and the small size of the accuracy confidence intervals we find the KNN model despite not having the highest average accuracy to be the one that fulfills our criteria the best.

ich glaubve hier fehlen einige Kommas, evl lassen wir den ganzen report nochmal durch einen Grammar checker laufen

5 Results

In order to be able to critically assess our result with KNN being the best model, a comparison with existing evaluations is advisable. Therefore, we conducted a search for papers, articles and competitions working with the dataset which describe their approach well. We observe most of the literature findings obtain much higher accuracy (**alotaibi2019**; **garate-escamila2020**; **uyar2017**) and overall perform well better not only than our preferred model regarding our criteria, the KNN, but also than our other models as seen in table 1. For example, **garate-escamila2020** was able to achieve accuracies of up to 99,4% with Random Forest, aforehand selecting features by chi-square and using PCA.

However, two differences in our analysis make comparison with other work difficult. On the one hand, due to the fact that plenty of research using the 13 preselected features provided by UCI already exists, we used the entire dataset and didn´t limit our features to the recommendations of the UCI, which were created using in-depth domain knowledge. On the other hand, our predictions are based on the combination of all four datasets provided by the UCI. Many other papers only use one of the four datasets, oftentimes 'Cleveland', as it has the best data quality. Combining these factors, it explains that other research surpasses our findings, as they followed other approaches and different criterions. For our approach, we could not find any comparable papers in our literature search.

Analyzing our result against the majority class baseline (55,1% accuracy), we achieve significantly increased accuracy not only for the KNN but also all our other models. Combined with only 70 Type 2 errors this indicates that our model performs well over the baseline.

To conclude whether or not our project helps doctors on diagnosing possible heart diseases more easily, we need to take certain limitations into account. Type 2 errors in disease prediction are particularly problematic because a sick patient is mistakenly found to be healthy. Contrary to the Type 1 error, in which a healthy patient is found to be ill, the diagnosis of the model is especially serious because no further treatment of the patient takes place.

However, if one takes more account of the actual application of the model in practice, the decision tree comes into sharper focus. To overcome the "black box" of machine learning for users, explainable AI (XAI) models like the decision tree

"to take more account" hab ich noch nie gehört; vielleicht nochmal checken

help to be interpretable and trustworthy even for laymen figure 3. Due to only marginal differences to the KNN (more type 2 errors, larger confidence intervals), the Decision Tree could serve as a valuable addition for application in real-world problems, esp. in interaction with humans, since it allows them to understand the decision of the model. In addition, the accuracy of our model, at just under 80%, is ultimately too low to make a reliable diagnosis. Considering all these limitations, we conclude that our model cannot replace a physician, but it can complement the physician's diagnosis in a quality-assuring way.

Eric hat im business understanding einen schönen Satz geschrieben: "By doing this analysis patients flagged for potential heart disease could possibly be prioritised" den würde ich so ähnlich oder darauf bezug nehmend hier auch nochmal schreiben