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

The aim of this project is the application of data mining methods, more specifically classification methods, to predict whether a patient could suffer from a heart disease or not. The successful application could help doctors and medical staff diagnose patients by automatically analysing a patient's historical test results and predicting whether there is a heart problem or not. Through this analysis, patients who have been diagnosed with heart disease could receive special treatment. Given the immense stress and long working hours that medical staff are exposed to, a standardized scheme for the evaluation of data could be beneficial. In the past such approaches have already been tested and proven to be a good diagnostic option (Usha Rani, n.d.). Jabbar et al. (n.d.) 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.

This project report is based on the "Heart Disease Dataset" (Janosi et al., n.d.) which 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., n.d.). At the time of writing, the dataset is partially corrupt, so the data is not identical to the original publication. The data is despite its age still relevant given the fact that it consists of results of medical tests. In addition to that it assumed to be valid because it is frequently used in contemporary research (see Nahar et al., n.d.; Usha Rani, n.d.; Aha and Kibler, n.d.).

2 Structure and Size of the Dataset

The data used in this paper comes from four individual datasets collected at different universities (Zurich, Budapest, Long Beach, Cleveland). The distribution of the target variable can be seen in table 1 on the following page.

For the purposes of this paper, the four datasets are considered as one coherent dataset consisting of 77 attributes (33 numeric, 42 categorical, 1 constant) and a total of 899 measurements. The attributes can be divided into the following categories:

Origin of data set	# of instances	Distribution target variable
Hungarian Institute of Cardiology, Budapest	294	106 / 188
University Hospital, Zurich	123	115 / 8
V.A. Medical Center, Long Beach	200	149 / 51
Cleveland Clinic Foundation, Cleveland	282	125 / 157

Distribution = heart disease / no heart disease

Table 1: Content of the dataset

- Patient data
- Electrocardiogram
- Cardiac fluoroscopy
- Coronary angiograms

The category of patient data includes characteristics such as *age*, *sex* or type of chest pain (*cp*), whereas the category electrocardiogram includes various electrocardiographic information obtained during an exercise electrocardiogram like the peak blood pressure (*tpeakbps*). Cardiac fluoroscopy contains all measurements obtained from a cardiac fluoroscopy, a medical measure that allows to see the flow of blood through vessels to evaluate the presence of blockages. An example for an attribute from this category is *ca* which denotes the number of found major vessels. The last category is coronary angiograms, which contains the results of the examination of the same name. The main attribute of interest of this category is *num* because it is the target variable that denotes whether a patient has a heart disease or not. In the raw data, patients without a heart disease are shown with a value of 0 for *num*, while patients with heart disease are shown with a value ≥ 1. An overview of all remaining attributes which were not named here is provided in the code documentation as well as the UCI Machine Learning Repository (Janosi et al., n.d.). In general, it is assumed that the collected values are of high quality, as they are the result of standardised medical measurement procedures, apart from the individual characteristics described by the patient, such as the location or type of pain. For this reason, it is assumed that a combination of the individual data sets is possible. This is also true because the measurements are not sorted in a certain way, neither in the individual data sets nor in the combined dataset.

The dataset contains the dates of the electrocardiogram and coronary angiograms. These dates are unevenly distributed over a period of seven years. These dates are considered irrelevant because the date of an examination does not affect its outcome as the period is to short to show evolutionary or general health changes in the society. Therefore a time series analysis is not possible with this dataset.

Finally, the distribution of the different variables in the combined dataset was analysed. The target variable was found to be reasonably symmetrically distributed

with 495 positive and 404 negative measurements, implying a disease prevalence of about 55.1% of the examined patients. Non-representative distributions were also found, e.g. *sex* is non-representatively distributed with 78% males and 22% females. The same applies to *age*, which is distributed similarly to a normal distribution in the range 28 to 77.

While assuming a high quality of the existing values, it was found that $30,8\%$ of all fields are lacking measurements. This is attributed to the fact that not all universities have carried out all measurements. For example the Cleveland subset lacks the location of chest pain. Some individual attributes in particular have many missing values like history of diabetes (*dm*) with roughly 90% missing values. Whereby only positive cases may have been entered here, since 95% of all filled cells show a diabetes disease which again is not representative. Furthermore, when checking whether meanings of attributes and their contained values fit together, some irregularities were observed. For example, the values 0 for cholesterol (*chol*) and blood pressure *testbps* are not compatible with life. Therefore, it is assumed that these values are misreported non-surveys, which are dealt with in the following chapter Preprocessing.

3 Preprocessing

Preprocessing was approached in two steps. In the first step the dataset was prepared based on knowledge obtained in section 2 and further analysis. Secondly, the data was transformed with the help of various algorithms that are combined in a pipeline.

3.1 Data Preparation

In the first step a custom loading function was implemented to put the four data sets into a usable format and harmonise known differences in encodings. In this step, the missing values marked -9 were replaced by `NaN` and the target variable was binarised by replacing all values greater than 1 by 1, as the UCI does not describe how the different positive cases ($num \geq 1$) differ.

After that irrelevant columns such as IDs, dates, constants, undescribed columns as well as irrelevant columns according to the UCI were removed. All of these columns are either not causally related to the presence of heart disease or cannot be examined for new patients because it is unknown what they measure. The feature *pncaden* was also dropped because it is the sum of the binary variables pain location, pain during exercise and relief on rest and therefore contains no additional information.

After the removal 55 attributes remained and were used to train the first model. XGBoost was chosen for that because it natively can handle `NaN` values. The accuracy of a 10-fold cross validation without any tuning was 0.98. This result was unexpectedly good, which is why the importance of the different features was analysed. It was found that the features belonging to the category of the coronary angiograms were serving as false predictors and therefore removed, leaving us with 45 remaining features. To validate that the removal hat the desired effect the cross validation was run again this time with an accuracy of 0.76 and a more reasonable distribution in the weights of the features.

After irrelevant columns were dropped the remaining features were analysed for inconsistencies. Here it was found that *thaltime* which describes the moment a measurement was taken during the exercise is sometimes larger than *thaldur* which denotes the duration of the exercise. In this case *thaltime* was replaced by `NaN` in all 17 instances. Another inconsistency was found between the maximum heart rate (*thalach*) and the heart rate at rest (*thalrest*). If the maximum heart rate was below the resting heart rate the maximum heart rate was replaced by `NaN` since the values were unusually small.

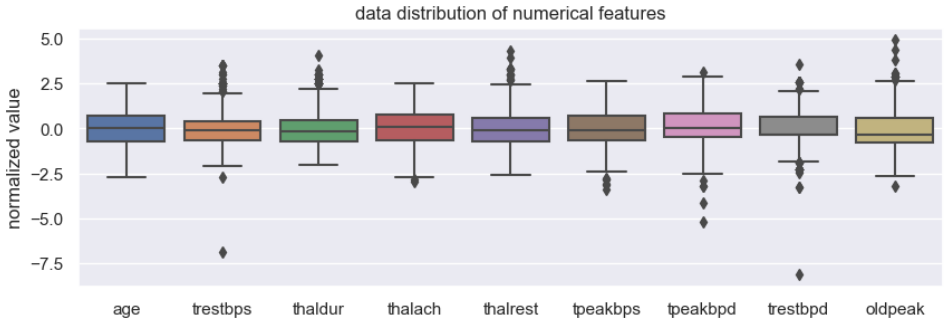

Figure 1: data distribution of all numeric elements

As shown in figure 1 normalised box plot of all numeric features were created 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 as stated in section 2 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 it was decided to keep these outliers as they might be a strong indication of a heart disease.

After handling all errors in the dataset the categorical features *cp, restecg, slope, ca, thal, restwm* were OneHotEncoded. All features were analysed regarding their pearson correlation, where only two pairs of features with a substantial

amount of data ($< 75\%$ NaNs) have a strong correlation ($> 75\%$). These are asymptomatic chest pain $\leftrightarrow$ pain during exercise and *rldv5* $\leftrightarrow$ *rldv5e*. The first correlation seems plausible, as pain that only occurs under exercise is classified as asymptomatic. The features are still relevant, however, as the cases where there is no correlation are of particular interest. For *rldv5* which denotes the height of the peak in the ECG during rest while *rldv5e* is the same under exercise also the cases where there is no correlation are of special interest because this might be typical for people with a heart disease. In order to better understand the interrelationships two new features were computed. Both represent the difference between a peak and a resting measurement. Once for the ECG (*rldv5_diff*) and once for the blood pressure (*thal_diff*). Furthermore, the feature indicating whether a patient smokes was enhanced by using the number of cigarettes per day and the length of time a patient has been smoking to infer the boolean variable. Hereby, the number of NaNs was reduced from 74% to 43%.

3.2 Hyperparameter Tuning

Additionally to the hyperparamter tuning of the estimators, which is described in section 4, the preprocessing steps were also optimised by different hyperparameters.

Firstly, binning for the feature *age* was added. Either 2, 5 bins or no binning at all are set as hyperparameters. The bins were encoded in an ordinal variable.

To impute the missing data a simple imputer was used. Missing values are replaced by the mean, median or mode of the feature. The KNN imputer was not used, because it is computationally much more expensive, while the iterative imputer is not used, as it is still experimental and therefore subject to change. Due to the high number of missing values and their uneven distribution across the features it was analysed how the number of imputed cells and number of features behave when columns with a certain amount of missing values are dropped. This is visualized in figure 2.

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

It becomes apparent that there are certain steps where the number of features increases significantly. To decide when a feature is included based on the number of missing values, the steps 0, 4, 8, 20, 35, 60, 75 and 100% were set as hyperparameter. They are shown as vertical lines in the graph.

To account for the different ranges of the features the MaxAbsScaler, MinMaxScaler, PowerTransformer, RobustScaler, StandardScaler and Normalizer are used. As the parameters of most scalers turn on or off core functionalities of the scaler, only tune the norm of the Normalizer was changed by hyperparameters with the norms l1, l2 and max.

To account for the slight imbalance of healthy and unhealthy patients oversampling and undersampling in the training data in comparison to no sampling was used. After this step the preprocessing is done.

4 Datamining

The first step in the process of data mining is to decide on a fitting algorithm. Because the best algorithm is not known in advance multiple are tested. For this project these were: KNNeighbors, Random Forest, Decision Tree, SVC, logistic regression, XGBoost, and four naïve bayes classifiers based on bernoulli, complement, gaussian and multinomial.

These ten classification algorithms will be evaluated to determine which one yields the best results. The quality of an algorithm is determined by two properties. The first criterion is a high recall, in order to avoid possible Type 2 errors (a patient with a heart disease is diagnosed as negative) in the diagnosis. The second criterion is the number of examinations required. The reason for this is that the model should be universally used by doctors. It is assumed that this is especially the case when simple examinations yield good diagnosis.

Besides the before mentioned use of scalers, imputers and samplers a variety of classifier specific hyperparameters were applied. For the KNN the number of used neighbors from 2 to 97 moving in 5-unit steps as well as the distance metric L1 and L2 as hyperparameters were added. The Random Forest was tested with 10 to 90 (10-unit steps) estimators and a maximum depth of None, 2, 6 and 10 and a minimum split of 2, 6, and 10. Decision Trees were tested with both the gini index and entropy as impurity measures while the values for maximum tree depth and minimum split were the same as for Random Forest. In the case of logistic regression, the same distance options as for KNN were added. For SVC C values ranging from 120 to 160 (20-unit steps), in combination with the gamma values 0.0001, 0.001, 0.01 and the kernel options linear, polynomial, sigmoidal and radial basis function were tested. For the four naïve bayes estimators alpha values from

0 to 19.5 (0.5-unit steps) were applied.

In order to find the best model a stratified nested cross validation was conducted using 10-folds each. To later decide which model was the best a classification report for every outer loop of the CV was saved. To run all CVs the work was split into several small parts where every unique combination of estimator, scaler and imputer represents its own part that was run on its own. In the following evaluation only the best model according to the previously defined criterions (high recall and simplicity) for each estimator is reviewed in greater detail. For measuring the simplicity the minimum percentage to be dropped was used as a metric. If models performed equally Occam's Razor was applied and favoured models with a low number of columns and basic preprocessing. The results without any ordering can be seen in table 2 in addition to some further prediction metrics.

Classifier	**Scaler**	**Sampler**	**Rec.**	**Rec. Std.**	**AUC.**	**F1**	**MPD**
Baseline	none	none	1	0	0.50	0.71	0
KNN	none	none	0.85	0.04	0.76	0.80	0
XGB	Normalizer	RUS	0.77	0.09	0.76	0.78	75
Random Forest	StandardScaler	none	0.84	0.09	0.78	0.81	100
Decision Tree	none	none	0.85	0.06	0.76	0.80	0
SVC	PowerTransformer	none	0.81	0.11	0.77	0.80	20
NB (bernoulli)	StandardScaler	none	0.79	0.08	0.76	0.79	8
NB (complement)	MinMaxScaler	none	0.84	0.03	0.77	0.81	100
NB (gaussian)	Normalizer	none	0.52	0.12	0.70	0.64	20
NB (multinomial)	MinMaxScaler	none	0.79	0.06	0.76	0.79	100
logistic regression	Normalizer	none	0.84	0.07	0.76	0.80	0

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

Table 2: Best models for every classification algorithm

The table includes all estimators as well as the baseline (majority vote).

As the majority vote predicts a disease in all cases there are no false negatives and therefore the recall of the baseline is always 1. However, the model would not be usable in the real world because the doctors would not benefit from the prediction as the model would predict the same for every patient. This is why the table also includes AUC and F1 as comparison, where the trained models outperform the baseline. This makes all models better to use than the baseline, as they can provide real added value.

In order to find the best model first the recall is analysed. Here larger differences can be observed. But the best performing models all have a recall of 0.84 to 0.85 (KNN, Random Forest, Descision Tree, NB (complete), logistic regression). To further narrow down the selection, the simplicity of the models is considered on the basis of the MPD. Here the Random Forest and the NB model are excluded

as they use all columns compared to the other models which only use completely filled columns. Since the remaining models do not differ significantly, a LeaveOneGroupOut cross validation is performed. Here, three data sets are used as the basis for training in order to classify the content of the fourth. The results for this procedure are also all within 0.01 for recall (0.78), precision (0.72), F1 (0.76), and accuracy (0.72). So they are assumed to be equal and Occam's Razor is conducted once more. KNN and the Decision Tree are simpler in preprocessing compared to the Logistic regression as they both do not use a scaler. If KNN and the Decision Tree are compared in their simplicity the Decision Tree outperforms the KNN because medical staff can just follow the Decision Tree whereas a KNN computation can not be done easily. So it is argued that the best model for classifying whether somebody has a heart disease or not is the model which uses the Decision Tree, followed closely by the before discussed KNN and logistic regression models.

The Decision Tree uses no scaler nor sampler. Also, no imputer was used because MPD is 0 and therefore no features with missing values remain. The best configuration was gini index as purity measure, a maximum tree depth of `None` and a minimum split of 2. This is identical to the defaults set by scikit-learn. In order to view what the model assumes to be important indicators it is visualised in figure 3. Only the tree with depth 3 (total depth 5) is displayed since focus is on the main attributes.

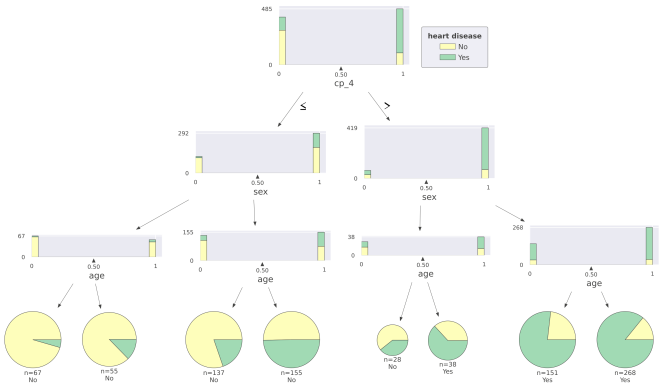

Figure 3: Decision Tree visualized

The root node of the tree decides whether a participant had asymptomatic chest pain or not. The resulting follow up nodes both use gender as the next split criterion. After that the tree splits according to age. The visualisation shows 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. The left side of the cp split conducts two more splits based on cp that are not displayed here.

5 Results

In order to be able to critically assess the result with the Decision Tree as the best model, a comparison is made with existing evaluations. Therefore, a search for papers, articles and competitions working with the dataset which describe their approach well is conducted. In doing so, it became clear that the used approach is not widespread. The dataset is frequently used, but not as a whole. Often only one sub-dataset, mostly Cleveland, is used. In comparison to the works that work exclusively on the Cleveland dataset, our best model is surpassed in every aspect (Ayatollahi et al., n.d.; Alotaibi, n.d.; Uyar and İlhan, n.d.). It should be noted that other models might not be generalisable as the models in this paper, as they may represent noise from the respective data set. Furthermore, it should be noted that the Cleveland dataset is a very pure dataset compared to the other datasets and hardly contains any missing values. A comparison of the most important features is also not possible as the only work that used every feature also included the false predictors in their models Gárate-Escamila et al., n.d.

When comparing the best model against the majority class baseline with a recall of 1, the best model is surpassed. Though this is manly due to the poor selection of recall as key metric as it does not reflect the usefulness of the model because it ignores the performance on negative values. If F1 is used as a metric the Decision Tree is able to outperform the baseline.

To conclude whether the project helps doctors on diagnosing possible heart diseases more easily, certain limitations need to be taken into account. Type 2 errors in disease prediction are particularly problematic because a sick patient is mistakenly found to be healthy and therefore might not receive the correct treatment. Contrary the Type 1 error, might lead to healthy patients getting medication they do not need. When considering the actual application of the model in practice, it is important to overcome the ”black box” of machine learning for users. For this, explainable AI models help to be interpretable and trustworthy even for laymen (see figure 3). As our chosen model is a Decision Tree, this property is fulfilled. Our model describes that old man with asymptomatic chest pain are most likely to have a heart disease. Comparing this to the knowledge of the medical field (Rodgers et al., n.d.) we see that we did not generate any new insights but confirmed what is well known.

References

Nahar, Jesmin et al. (n.d.). "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.

Kavitha, R. and E. Kannan (n.d.). "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/.

Statistisches Bundesamt (n.d.). *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.

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

Soni, Jyoti et al. (n.d.). "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.

Usha Rani (n.d.). "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.

Jabbar, M. Akhil, B.L. Deekshatulu, and Priti Chandra (n.d.). "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.

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

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

Ayatollahi, Haleh, Leila Gholamhosseini, and Masoud Salehi (n.d.). "Predicting coronary artery disease: a comparison between two data mining algorithms". In: *BMC Public Health* 19.1 (), p. 448. ISSN: 1471-2458. DOI: 10.1186/s12889-019-6721-5. URL: https://doi.org/10.1186/s12889-019-6721-5.

Alotaibi, Fahd Saleh (n.d.). "Implementation of Machine Learning Model to Predict Heart Failure Disease". In: *International Journal of Advanced Computer Science and Applications* 10.6 (). ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2019.0100637. URL: http://thesai.org/Publications/ViewPaper?Volume=10&Issue=6&Code=IJACSA&SerialNo=37.

Uyar, Kaan and Ahmet İlhan (n.d.). "Diagnosis of Heart Disease Using Genetic Algorithm Based Trained Recurrent Fuzzy Neural Networks". In: *Procedia Computer Science*. 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary 120 (), pp. 588–593. ISSN: 1877-0509. DOI: 10.1016/j.procs.2017.11.283. URL: https://www.sciencedirect.com/science/article/pii/S187705091732495X.

Gárate-Escamila, Anna Karen, Amir Hajjam El Hassani, and Emmanuel Andrès (n.d.). "Classification Models for Heart Disease Prediction Using Feature Selection and PCA". In: *Informatics in Medicine Unlocked* 19 (), pp. 100330–. DOI: 10.1016/j.imu.2020.100330. URL: https://hal.archives-ouvertes.fr/hal-03491100.

Rodgers, Jennifer L. et al. (n.d.). "Cardiovascular Risks Associated with Gender and Aging". In: *Journal of Cardiovascular Development and Disease* 6.2 (), p. 19. ISSN: 2308-3425. DOI: 10.3390/jcdd6020019. pmid: 31035613. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616540/.