Proceedings of the Federated Conference on Computer Science and Information Systems pp. 155–163
DOI: 10.15439/2017F219
ISSN 2300-5963 ACSIS, Vol. 11

Predictive and Descriptive Analysis for Heart Disease Diagnosis

František Babič, Jaroslav Olejár
Department of Cybernetics and Artificial Intelligence,
Faculty of Electrical Engineering and Informatics,
Technical university of Košice, Slovakia
frantisek.babic@tuke.sk, jaroslav.olejar@tuke.sk

Zuzana Vantová, Ján Paralič
Department of Cybernetics and Artificial Intelligence,
Faculty of Electrical Engineering and Informatics,
Technical university of Košice, Slovakia
zuzana.vantova@tuke.sk, jan.paralic@tuke.sk

***Abstract*—The heart disease describes a range of conditions affecting our heart. It can include blood vessel diseases such as coronary artery disease, heart rhythm problems or and heart defects. This term is often used for cardiovascular disease, i.e. narrowed or blocked blood vessels leading to a heart attack, chest pain or stroke. In our work, we analysed three available data sets: Heart Disease Database, South African Heart Disease and Z-Alizadeh Sani Dataset. For this purpose, we focused on two directions: a predictive analysis based on Decision Trees, Naive Bayes, Support Vector Machine and Neural Networks; descriptive analysis based on association and decision rules. Our results are plausible, in some cases comparable or better as in other related works**

I. Introduction

THE availability of various medical data leads to a reflection if we have some effective and powerful methods to process this data and extract potential new and useful knowledge. A diagnostics of different diseases represents one of the most important challenges for data analytics. The researchers focus their activities in several directions, e.g. to generate prediction models with high accuracy, to extract IF-THEN rules or to investigate new cut-off values for relevant input variables [30]. All directions are important and can contribute to improving the effectiveness of the medical diagnostics.

Heart disease (HD) is a general name for a variety of diseases, conditions, and disorders that affect the heart and the blood vessels. The symptoms depend on the specific type of this disease such coronary artery diseases, stroke, heart failure, hypertensive heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, etc. HD is the leading global cause of death based on a statistics of American Heart Association. The World Health Organization estimated that in 2012 more than 17.5 million people died from HD (31% of all global deaths). This growing trend can be reversed by an effective prevention, i.e. early identification of warning symptoms or typical patient's behavior leads to HD. It is a task for data analytics to support the diagnostic process with results in simple understandable form for doctors or general practitioners without deeper knowledge about algorithms and their applications.

The paper consists of three main sections. At first, we introduce the motivation and possible approaches to support the effective diagnostics of HD. The second section describes six phases of the CRISP-DM methodology and related experiments. The last section concludes the paper and proposes some improvements for our future work.

A. Related Work

As HD is a very serious disease, many researchers tried to predict it or to extract crucial risk factors. The relatively known data sets are Cleveland, Hungarian, and Long Beach VA freely available on UCI machine learning repository.

El-Bialy et al. used all these datasets in their study [18]. At first, authors selected five common variables for each dataset (Cleveland, Hungarian, Long Beach VA and Statlog project) and applied two data mining techniques: decision tree C4.5 and Fast Decision Tree (improved C4.5 by Jiang SU and Harry Zhang). These operations resulted in accuracy from 69.5% (FDT, Long Beach VA data) to 78.54% (C4.5, Cleveland). Next, authors merged all datasets into one containing variables like cp, age, ca, thal, or thalach. The merging resulted in 77.5% accuracy by the C4.5 algorithm and 78.06% by FDT.

Verma, Srivastava, and Negi in their work [19] designed a hybrid model for diagnosing of coronary artery disease (one type of HD). For this purpose, authors used data from the Department of Cardiology at Indira Gandhi Medical College in Shimla in India, which contained 335 records describing by 26 attributes. The authors pre-processed data via correlation and features selection by particle swarm optimization (PSO) method. For modelling authors used four analytical methods: Multi-layer perceptron (MLP), Multinomial logistic regression model (MLR), Fuzzy unordered rule induction algorithm (FURIA) and Decision Tree C4.5. As first, they applied the MLP to the whole dataset. The obtained accuracy 77% was not satisfactory, so they tried to improve it by MLR. This step resulted in 83.5% accuracy. The methods FURIA and C4.5 did not offer higher value. In the next step, authors tried to optimize data

pre-processing and tried to identify the prime risk factors using correlation based feature subset (CFS), selection with PSO, k-means clustering and classification or a combination all of these. Finally, they achieved 88.4% accuracy using MLR and they applied proposed hybrid model on Cleveland dataset. This model improved the accuracy of classification algorithms from 8.3 % to 11.4 %.

Cleveland dataset is only one of several datasets typically used by researchers for analytical support of HD diagnostics. The second is Z-Alizadeh Sani dataset collected at Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre. This dataset contains 303 records with 54 features. Alizadehsani et al. aimed to classify patients into two target classes: suffered by coronary artery disease or normal [20]. They divided the original set of variables into four groups: demographic, symptoms and examination, ECG, laboratory, and echo. They focused on pre-processing phase, mainly on features selection by Ginni index or Support Vector Machine. They also created several new variables like LAD (Left Anterior Descending), LCX (Left Circumflex) and RCA (Right Coronary Artery). For classification, authors used four methods: Naive Bayes, Sequential Minimal Optimization (SMO), SMO with bagging and Neural network. They applied these algorithms to different pre-processed data sets: all original variables without three new, all original variables with three new, only selected variables from the original set without three new and the selected variables from the original set with three new. The best-obtained results were 93.4% for SMO with bagging, 75.51% for Naive Bayes, 94.08 for SMO and 88.11 for the Neural network. All results were verified by 10-cross validation. The authors extracted Typical Chest Pain, Region RWMA2, age, Q-Wave and ST Elevation as the most important variables.

The same dataset was used by Yadav´s team [21] focusing on an optimization of Apriori algorithm by Transaction Reduction Method (TRM). The new algorithm decreased a size of the candidate´s set and a number of transactional records in the database. The authors compared it with some traditional methods and obtained accuracy 93.75%; SMO (92.09%), SVM (89.11%), C4.5 (83.85%), Naïve Bayes (80.15%).

All mentioned works focused mainly on predictive analysis within traditional methods like Decision Trees, Naive Bayes, Support Vector Machine or Neural networks. Next, authors tried to improve these results by suitable operations in pre-processing phase or by other analytical methods like SMO. We selected these works to set a baseline for our research activities.

B. Methods

The CRISP-DM represents the most popular methodology for data mining and data science. This methodology defines six main phases from business understanding to the deployment [1], [2]. The first phase deals with a specification of business goal and its transformation to the data mining context. The second phase focus on detailed data understanding through various graphical and statistical methods. Data preparation is usually the most complex and most time-consuming phase including data aggregation, cleaning, reduction or transformation. In modeling, different machine learning algorithms are applied to the preprocessed datasets. Traditionally, this dataset is divided into training and testing sample, or analysts use 10-cross validation. The obtained results are evaluated by traditional metrics like accuracy, ROC, precision, or recall. Next, the analysts verify accomplish of the specified business goals. The last phase is devoted to the deployment of the best results in real usage and identification of the best or worst practices for the next analytical processes.

The decision tree is a flowchart-like tree structure, where each non-leaf node represents a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent target classes or class distributions [3]. We decided to use this method because we were able to visualize the tree or to extract the decision rules. The C4.5 algorithm used normalized information gain for splitting [4]. The C5.0 algorithm represents an improved version of the C4.5 that offers a faster generation of the model, less memory usage, smaller trees with similar information value, weighting, and support for the boosting [5]. CTree is a non-parametric class of regression trees embedding tree-structured regression models into a well-defined theory of conditional inference procedures. This algorithm does not use the traditional variable selection based on information gain or Ginni coefficient but selects the variables with many possible splits or many missing values [6]. It uses a significance test procedure. The CART (Classification and Regression Trees) algorithm builds a model by recursively partitioning the data space and fitting a simple prediction model within each partition [7]. The result is a binary tree using a greedy algorithm to select a variable and related cut-off value for splitting with the aim to minimize a given cost function.

Naive Bayes is a simple technique for constructing classifiers that require a small number of training data to estimate the parameters necessary for classification [8]. It uses the probabilities of each attribute belonging to each class to make a prediction. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. To make a prediction, it calculates the probabilities of the instance belonging to each class and selects the class value with the highest probability.

Support Vector Machine (SVM) is a supervised machine-learning algorithm, mostly used for classification. SVM plots each record as a point in n-dimensional space (where n is a number of input variables). The value of each variable represents a particular ordinate [9]. Then, SVM performs classification by finding a hyperplane distinguishing the two target classes very well. New examples are mapped into

same space and predicted to a category based on which side of the gap they fall.

Neural networks are a computational model, which is based on a large collection of connected simple units called artificial neurons. We will use this method if we don't want to know the decision mechanism. They work like a black box, i.e. we know the model is some non-linear combination of some neurons, each of which is some non-linear combination of some other neurons, but it is near impossible to say what each neuron is doing. This approach is opposite to the Decision trees or SVM. We used a feedforward neural network, in which the information moves in only one direction – forward [10].

Association rules are a popular and well-researched method for discovery of interesting relations between variables in (large) databases [11], [12]. The most often used algorithm to mine association rules is Apriori [13]. Association rules analysis is a technique to uncover how items are associated with each other. The Apriori principle can reduce the number of item sets we need to examine.

We used some statistical methods to investigate possible relations between input variables themselves or between them and target diagnostics [14]. For this purpose, we applied Shapiro-Wilks normality test (null hypothesis: sample x_i came from a normally distributed population [22]); 2-sample Welch's t-test (the population means from the two unrelated groups are equal [23]); Mann-Whitney-Wilcoxon Test [29] (the two populations are equal); Pearson chi-square independence test (two categorical variables are independent [24]); Fisher's Exact test (the relative proportions of one variable are independent of the second variable [25]) and logistic regression [28]. Next, we used k-Nearest Neighbour and multiple linear regression to replace the missing values in a dataset by some plausible values [26], [27].

For experiments, we used a language and environment for statistical computing and graphics called R. It provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible[1].

II. CRISP-DM

A. Business Understanding

As we mentioned before, the HD diagnostics is complicated process containing many different input variables that need to be considered. In this case, data mining can help to process and analyze available data from a different point of views. The business goal is to provide an application decision support for doctors and general practitioners. From data mining point of view, we can transform this goal into two possible directions: predictive and descriptive analysis. The first one was represented by binary classification and the second one by extraction of association or decision rules. In data preparation phase we aimed to investigate possible relations between different combinations of variables within some statistical tests. We determined a minimal 85% accuracy based on performed state of the art. For this evaluation, we used a traditional confusion matrix (Table I.)

TABLE I.
CONFUSION MATRIX

	True values	
Predicted value	TP	FP
	FN	TN

TP (true positive) – healthy people were classified correctly as healthy.

FP (false positive) – healthy people were classified incorrectly as patients with the positive diagnosis.

FN (false negative) – patients with positive diagnosis were classified incorrectly as healthy people.

TN (true negative) – patients with positive diagnosis were classified correctly as sick.

The overall accuracy was calculated within following formula:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (1)$$

For evaluating the association rules, we used two traditional metrics: support as the proportion of health records in the dataset, which contains a combination of relevant antecedents; and confidence representing the proportion of health records containing antecedents and consequent two. If we define association rules as X=>Y, then antecedents represent the left part of the rule (X) and the consequent the right part (Y).

B. Data Understanding and Preparation

We selected three available data sets devoted to the heart diseases. The first dataset dates from 1988 and consists of four databases: Cleveland (303 records), Hungary (294), Switzerland (123), and Long Beach VA (200). Each record is described by 14 variables (Table II.).

The distribution of target attribute in each dataset is following: 139 patients with positive diagnostics/164 healthy people, 106/188, 115/8, 149/51.

This dataset contained nearly 2 thousand missing values, mainly in variable *ca*. For nominal attributes, we used the k-NN method, for numerical multiple linear regression to solve this problem. In the case of k-NN, we normalized all variables to interval < -1, 1> and set up the k-value as 5. This cleaning operation changed the original distributions very slightly.

Fig. 1 visualizes a relation between *resting blood pressure* and target attribute. We can say that many patients with ideal

[1] https://www.r-project.org/about.html

blood pressure (around 120) have a positive diagnosis of heart disease.

TABLE II.
VARIABLES – DATASET 1

Name	Description
age	age in years (28 - 77)
sex	0-female 1-male
cp	chest pain type 1-typical angina 2-atypical angina 3-non-anginal pain 4-asymptomatic
tresbps	resting blood pressure (mm/Hg) (0 - 200)
chol	serum cholesterol (mg/dl) (0 - 603)
fbs	fasting blood sugar 0-false(< 120 mg/dl) 1-true (> 120 mg/dl)
restecg	resting electrocardiographic results 0-normal 1-having ST-T wave abnormality 2-showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved (60 - 202)
exang	exercise induced angina 0-no 1-yes
oldpeak	ST depression induced by exercise relative to rest (mm) (-2.6 – 6.2)
slope	the slope of the peak exercise ST segment 1-upsloping 2-flat 3-downsloping
ca	number of major vessels colored by fluoroscopy (0-3)
thal	3-normal 6-fixed defect 7-reversable defect
num	diagnosis of heart disease (angiographic disease status) – target attribute 0 - negative diagnosis (absence) 1 - 4 (from least serious most serious - presence)

Fig.2 visualizes a relation between *maximal achieved heart rate* and target attribute. We expect that the higher value covers people with regular exercise. Fig. 3 visualizes a relation between *sex* and target attribute. The first finding was that in the integrated dataset we had significantly more male than female. The second was a ratio between healthy people and patients with positive diagnosis in both groups.

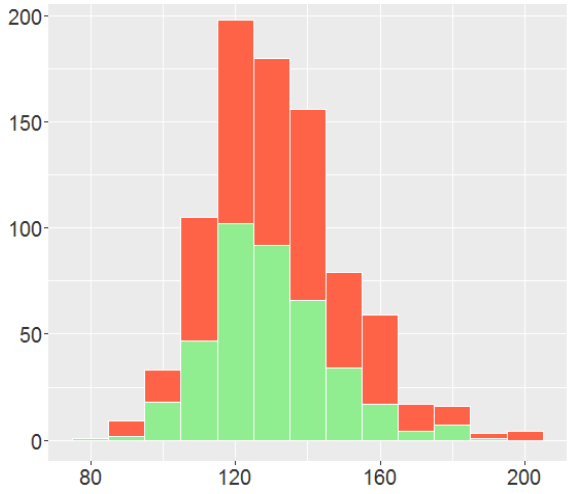

Fig. 1 Histogram (x- trestbps, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

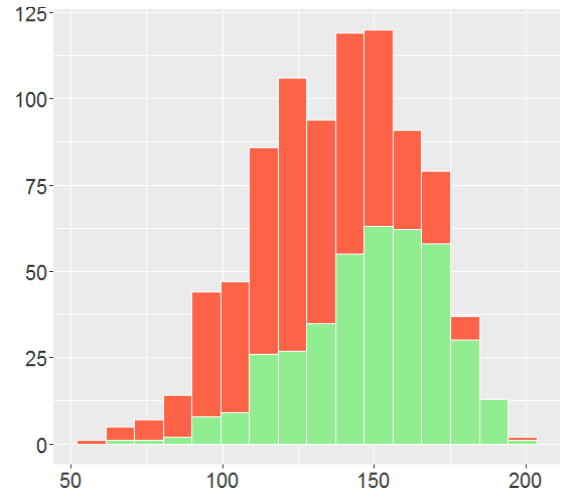

Fig. 2 Histogram (x- thalach, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

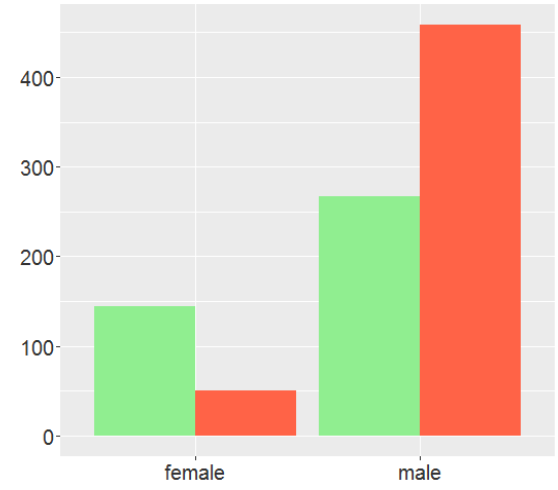

Fig. 3 Histogram (x- sex, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

For a better understanding of variables describing the EKG results, we present an example of ST segments (Fig. 4). The normal ST segment has a slight upward concavity. Flat, downsloping or depressed ST segments may indicate coronary ischemia. ST elevation may indicate transmural myocardial infarction.

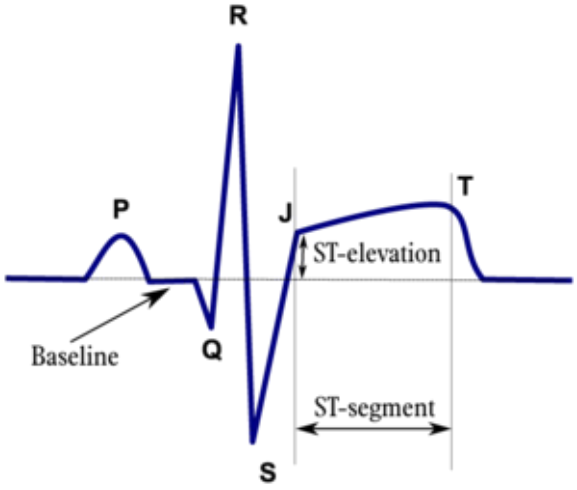

Fig. 4 An example of ST segment from electrocardiography [16]

Next, we investigated a possibility to reduce the input set of variables. Three variables described the ECG: *Restecg*, *Oldpeak*, and *Slope*. If we visualized their distributions, the first two had a low distinguish ability for target classes. We have omitted these two attributes from next experiments.

Next, we investigated a relationship between nominal variables and target attribute. For this purpose, we tried to use Pearson's Chi-squared Test and we obtained following p-values: *sex* (< 2.2e-16), *exang* (< 2.2e-16), *cp* (< 2.2e-16), *fbs* (5.972e-05), *slope* (< 2.2e-16), *ca* (8.806e-16) and *thal* (< 2.2e-16). These results rejected the null hypothesis and confirmed the expected relationship.

In the case of numerical variables, we started with Shapiro-Wilks normality test: *age* (p-value = 2.268e-05), *trestbps* (3.052e-15), *chol* (< 2.2e-16), *thalach* (1.906e-05). Based on this result, we choose a non-parametric Mann-Whitney-Wilcoxon Test: *age* (p-value = 2.2e-16), *tresbps* (0.001596), *chol* (4.941e-05), *thalach* (2.2e-16). If we set up the 0.05 as significance level, we rejected the null hypothesis for all numeric variables, i.e. existing dependency between them and target attribute.

Finally, we applied a logistic regression on this data. If we set a level for statistical significance to 0.005, the most significant variables were *sex = male* (p-value = 0.000602), *slope = 2* (0.000355), *ca = 1* (1.04e-05), *ca = 2* (3.01e-06), *ca = 3* (0.008745), *cp = 4* (0.001302) and *tresbps* (0.007671). Based on relevant z-values, all these variables increase the chance for classification into positive target class.

Fig.5 visualizes a distribution of target attribute after its transformation to binary one (1-4 were aggregated in 1).

Fig. 5 Histogram (x- target attribute (absence, presence), y- multiplicity)

Since in previous dataset we investigated the higher occurrence of HD in male sample, as second dataset we selected a sample of males in a heart-disease high-risk region of the Western Cape, South Africa. It contains 462 records described by 10 variables without missing values (Table III.). This data sample is part of a larger dataset, described in [15]. The target attribute contained 302 records of healthy persons and 160 positive diagnoses of HD.

TABLE III.
VARIABLES – DATASET 2

Name	Description
age	age in years (15 - 54)
sbp	systolic blood pressure (101 - 218)
chd	coronary heart disease – target attribute 0 – negative diagnosis 1 – positive diagnosis
tobacco	cumulative tobacco (kg) (0 - 31.2)
LDL	low density lipoprotein cholesterol (0.98 - 15.33)
adiposity	measure of % body fat (6.74 - 42.49)
obesity	measure weight-to-height rations (BMI). (14.70 – 46.58)
famhist	family history of heart disease (present, absent)
typea	type-A behavior is characterized by an excessive competitive drive, impatience and anger/hostility
alcohol	current alcohol consumption (0.00 – 147.19)

We performed similar operations to understand the data. Figure 5 visualizes a relation between family history of heart disease (*famhist*) and target attribute. We can say that a

chance to be positive diagnosed is around 50% if some heart disease occurred in the family history.

We solved an unbalanced distribution of the target attribute with an oversampling method, i.e. we re-sampled the minority class in the training set. The result was 302 records for absent and 256 for the present.

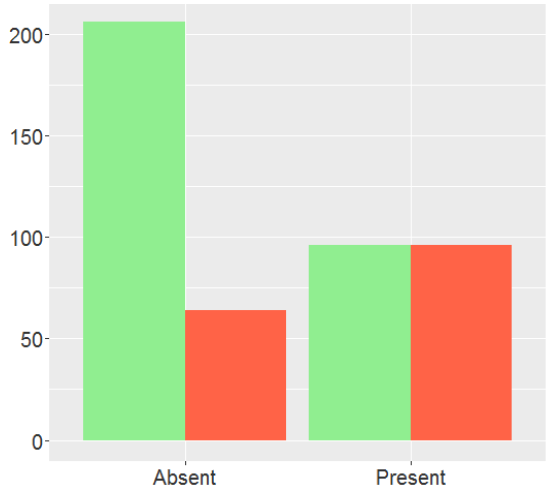

Fig. 6 Histogram (x- famhist, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

Next, we investigated potentially existing relationships between input variables and target attributes. Since we had mainly numeric variables, at first we tested if they had a normal distribution. For this purpose, we used Shapiro-Wilks test with following results: *age* (p-value = 4.595e-13), *sbp* (1.253e-14), *tobacco* (< 2.2e-16), *LDL* (7.148e-15), *adiposity* (4.245e-05), *obesity* (9.228e-10), *typea* (0.008604) and *alcohol* (< 2.2e-16). In all cases, the distribution deviated from normality based on cut-off p-value 0.05. Based on this result, we choose a non-parametric Mann-Whitney-Wilcoxon Test. We obtained following p-values: *age* (3.364e-15), *sbp* (0.000214), t*obacco* (4.31e-12), *LDL* (6.058e-09), a*diposity* (1.385e-07), *obesity* (0.02073), *typea* (0.05219) and *alcohol* (0.1708). If we set up the 0.05 as significance level again, we confirmed the null hypothesis for the two last variables, i.e. they were independent and we omitted them.

Finally, we generated a logistic regression model based on data with a reduced set of input variables. This model has confirmed the importance of *famhist* = present (p-value = 2.68e-05), *age* (0.000572), *tobacco* (0.001847) and *LDL* (0.002109). All four variables increase a chance for the positive diagnosis of the HD.

The last dataset was Z-Alizadeh Sani Dataset containing 303 records about patients from Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre [17]. Each patient was characterized by 54 variables. These variables were arranged in four groups: demographic, symptom and examination, ECG, laboratory and echo features. The target classes were a positive diagnosis of coronary artery disease (CAD) and normal health status. The patient is categorized as CAD, if his/her diameter narrowing is greater than or equal to 50%, and otherwise as Normal. The dataset did not contain any missing values. In Table IV, we present only selected set of input variables as an example.

TABLE IV.
VARIABLES – DATASET 3

Name	Description
age	age in years (30 - 86)
Diabetes Mellitus	0 - No 1 - Yes
Fasting Blood Sugar (mg/dl)	62 - 400
Pulse Rate	50 - 100
ST Elevation	0 - No 1 - Yes
ST Depression	0 - No 1 - Yes
Low-Density Lipoprotein (mg/dl)	18- 232
White Blood Cell	3 700 – 18 000
Obesity	0 - No (BMI<25) 1 – Yes (BMI >25)
Creatine (mg/dl)	0.5 - 2.2
Ex-Smoker	0 - No 1 - Yes
Hemoglobin (g/dL)	8.9 - 17.6
Non-anginal CP	0 - No 1 - Yes
Heart disease (target attribute)	0 - Normal 1 - CAD

We performed similar data understanding, Fig. 7 visualizes a relation between typical chest pain and target attribute. If the patient felt chest pain, then change for heart disease is very high. If the patient did not feel the chest pain, the chance for HD is still 50/50.

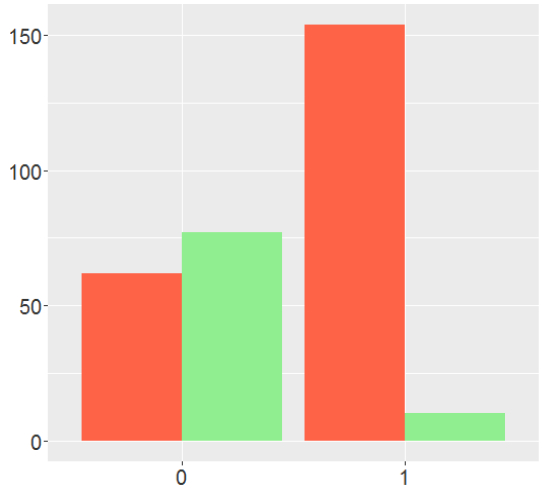

Fig. 7 Histogram (x- typical chest pain, y-multiplicity, green color – healthy person, orange color – positive diagnosis (CAD))

We have proceeded in the same way as in the two previous cases. We omitted the variables with only one values as e.g. Exertional CP. We started with an investigation of the possible normal distribution of numerical variables. We confirmed it for three variables through histograms (Fig. 8) and Shapiro-Wilks normality test: HB (p-value = 0.1301), Lymph (0.08769) and Neut (0.2627).

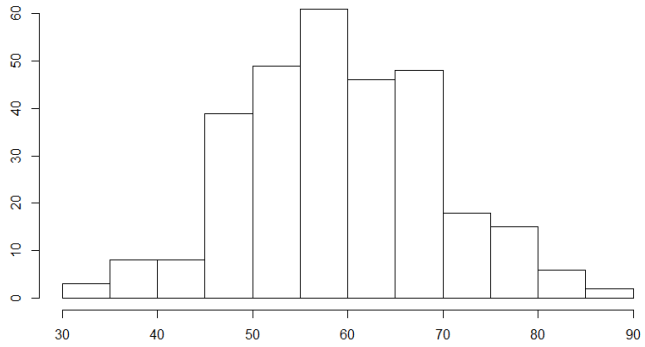

Fig. 8 Histogram (x- number of Neutrophil , y-multiplicity)

For these variables, we performed two-sample Welch's t-test with following results: *HB* (p-value < 2.2e-16), *Lymph* (< 2.2e-16) and *Neut* (< 2.2e-16). We rejected the null hypothesis and accepted the alternative one (the population means are not equal), i.e. variables and target attribute were dependent.

For other numerical variables, we performed the non-parametric Mann-Whitney-Wilcoxon Test. Based on significance level 0.05 we were able to reject the null hypothesis for following variables: *Weight* (0.2137), *Length* (0.9428), *BMI* (0.2537), *Edema* (0.3485), *CR* (0.3251), *LDL* (0.6088), *HDL* (0.5036), *BUN* (0.1293), *Na* (0.09188), *WBC* (0.3672) and *PLT* (0.2292). It means that these variables and target attribute were independent and we excluded them from our experiments.

For nominal variables, we had to use two methods: at first Pearson chi-square independence test and if more than 20% of the contingency table's cells are less than five, we used Fisher's Exact test. We omitted following variables with significance level 0.05: *sex* (p-value = 0.2991), *Current Smoker* (0.2614), *FH* (0.6557), *Obesity* (0.8003), *DLP* (0.9284), *Systolic Murmur* (1.0), *LVH* (0.5251); *Ex-Smoker* (0.7297), *CRF* (0.1875), *CVA* (1.0), *Airway disease* (0.1875), *Thyroid Disease* (0.4138), *Weak Peripheral Pulse* (0.3263), *Lung Rales* (0.7349), *Function Class* (0.1405), *LowTH Ang* (1.0), BBB (0.307) and *Poor R Progression* (0.064).

The mentioned operations reduced an original set of input variables from 53 to 27. Finally, we generated a logistic regression model that confirmed the importance of variables *Typical Chest Pain = 1* (3.63e-05), *Age* (4.06e-05), *Diabetes Mellitus* (0.00321), *T inversion* (0.00140), *Valvular heart disease = normal* (0.00605) and *Regional wall motion abnormality* (0.0057).

This dataset was also unbalanced from the target attribute point of view. Therefore, we used the oversampling method again, i.e. we re-sampled the minority class in the training set. The result was 216 records for CAD and 174 for normal.

C. Modelling and Evaluation

We applied selected data mining methods to predict the HD and to extract relevant rules. For the prediction models, we verified experimentally three data division to training and testing sets (80/20, 70/30, 60/40). For each division, we repeated the experiment for 10 times with different sampling and in respect to the original ratio of target classes. We also performed a stratified 10-cross validation, i.e. each fold contained roughly the same proportions of the two types of class labels. Finally, we applied these methods on original datasets and with reduced sets of variables mentioned above.

Table V. presents the best-achieved results for all three datasets.

TABLE V.
THE BEST ACHIEVED PREDICTIONS

Name	Method	Accuracy (%)
Cleveland, Hungary, Switzerland and Long Beach VA	Decision trees	88.09
	Naive Bayes	86.76
	SVM	88.53
	Neural networks	89.93
South Africa Heart Disease	Decision trees	73.87
	Naïve Bayes	71.17
	SVM	73.70
	Neural networks	68.48
Z-Alizadeh Sani Dataset	Decision trees	85.38
	Naïve Bayes	83.33
	SVM	86.67
	Neural networks	86.32

The first group of experiments resulted in the best model generated by neural network visualized in Fig. 9. This model contained only reduced set of input variables. The best decision tree model was generated by CART, 10-stratified cross-validation, and all original variables. The reduced set of variables provided the same accuracy.

For the second dataset, we obtained less accurate models. The proposed reduction of the variables did not bring any significant improvement. Based on relevant confusion matrices, we found out a relatively high number of records predicted as negative, but in fact, they were positive. This is an important finding for future improvement.

In the last case, we obtained the best model within SVM, which for all datasets provided one of the most accurate predictions. Since it is complicated to visualize the whole SVM model; we present an only 2-dimensional example based on attributes *age* and *tresbps* (Fig. 10).

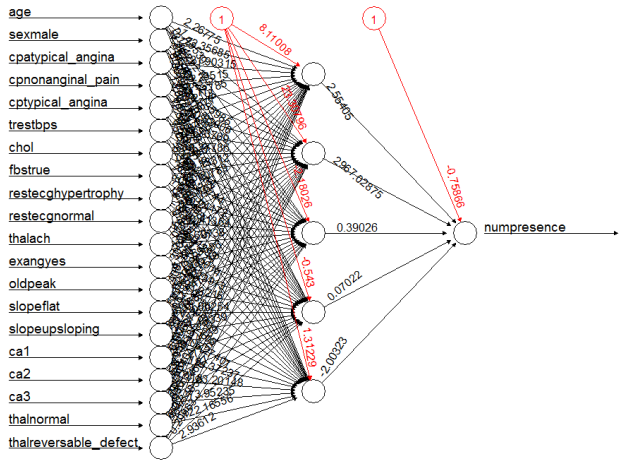

Fig. 9 The structure of the created neural network with the best prediction ability (input layer = 20 neurons, hidden layer = 5 neurons, red numbers - bias)

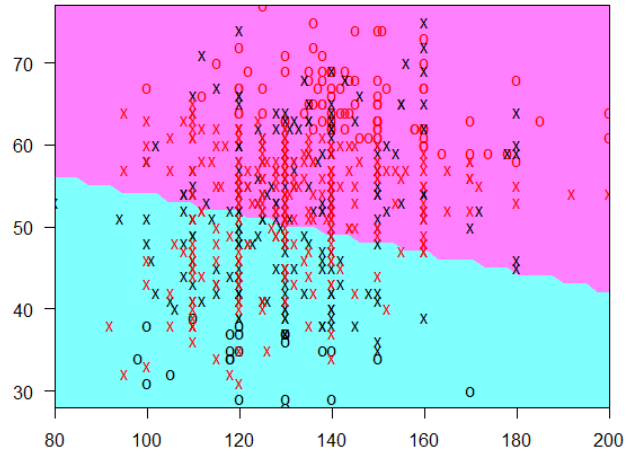

Fig. 10 The example of SVM prediction model (x – tresbps,y – age, purple – positive diagnosis, cyan – negative diagnosis)

Finally, we discretized all numerical attributes in accordance with the typical categories defined by existing medical literature. Next, we applied the Apriori algorithm to generate association rules. These rules were very similar to those we extracted from the decision trees models:

IF *thal* = reversable defect/ fixed defect AND *ca* = 1/2/3 THEN *HD* = positive diagnosis (dataset1 – decision rules)

IF *sex* = male AND *exang* = yes AND *oldpeak* > 0.8 THEN HD = positive (dataset1 – association rules)

IF *age* < 50.05 AND *tobacco* > 0.46 AND *typea* > 68.5 THEN HD = positive (dataset2)

IF famhist = 1 and *LDL* = high THEN HD = positive (dataset2)

IF *Typical Chest Pain* = 0 AND *Age* < 61 and *Region RWMA* = 0/1 THEN HD = positive (dataset3)

IF *Typical Chest Pain* = 1 AND *VHD* = mild THEN HD = positive (dataset3)

We can conclude that the content of the generated prediction models is in accordance with the results of the relevant statistical tests about attributes dependency.

D. Deployment

The last phase is devoted to the deployment of the evaluated and verified models into practice. We focused on simple understandable and interpretable application to support the diagnostic process of the heart diseases. We used an open source R package Shiny that provides an elegant and powerful web framework for building web applications using R . Fig. 11 visualizes our prototype offering all analytical features described in this paper, e.g. data understanding, statistical tests, classification models generation and diagnosis of a new patient.

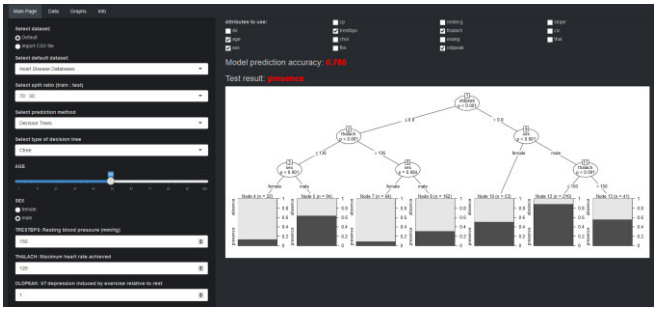

Fig. 11 An example of supporting application

III. Conclusion

This paper presents the application of various statistical and data mining methods to understand three different medical data sets, to generate some prediction models or to extract rules suitable for decision support during the diagnostic process. We used some statistical tests to find out possible existing relationships between input variables and target attribute. Based on relevant results, we prepared the datasets for modeling phase, in which we applied some selected methods as decision trees, Naive Bayes, Support Vector Machine or Apriori algorithm. In comparison with existing studies, our results are plausible, in some cases comparable or better. In our future work, we will focus on several directions: transformation and creation of the new derived variables to improve the data information value; investigation of the new cut-off values for selected variables, boosting for prediction models and e.g. cost matrix for unbalanced distribution

Acknowledgment

The work presented in this paper was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant no. 1/0493/16, by the Cultural and Educational Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grants no. 025TUKE-4/2015 and no. 05TUKE-4/2017.

The authors would like to thank the principal investigators responsible for data collection: Andras Janosi, M.D. (Hungarian Institute of Cardiology, Budapest); William Steinbrunn, M.D. (University Hospital, Zurich); Matthias

Pfisterer, M.D. (University Hospital, Basel); Robert Detrano, M.D., Ph.D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation; J. Rousseauw et al. and Z-Alizadeh Sani et. al.

REFERENCES

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth: "CRISP-DM 1.0 Step-by-Step Data Mining Guide", 2000.

[2] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Ware-housing, vol. 5, no. 4, 2000, pp. 13–22.

[3] K.S. Murthy, "Automatic construction of decision tress from data: A multidisciplinary survey", Data Mining and Knowledge Discovery, 1997, pp. 345–389, doi: 10.1007/s10618-016-0460-3.

[4] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993, doi: 10.1007/BF00993309.

[5] N. Patil, R. Lathi, and V. Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique", International Journal of Engineering Research & Technology, vol. 1, no. 4, 2012, pp. 1–5.

[6] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework", Journal of Computational and Graphical Statistics, vol. 15, no. 3, 2006, pp. 651–674, doi: 10.1198/106186006X133933.

[7] L. Breiman, J.H. Friedman, R.A. Olshen, Ch.J. Stone, "Classification and Regression Trees", 1999, CRC Press, doi: 10.1002/cyto.990080516.

[8] D. J Hand, K. Yu, "Idiot's Bayes-not so stupid after all?", International Statistical Review, vol. 69, no. 3, 2001, pp. 385–399. doi:10.2307/1403452

[9] C. Cortes, V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, 1995, pp. 273–297, doi:10.1007/BF00994018.

[10] K. Hornik, "Approximation capabilities of multilayer feedforward networks," Neural Networks, vol. 4, 1991, pp. 251–257, doi: 10.1016/0893-6080(91)90009-T.

[11] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Data-bases", Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp 487-499.

[12] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for Association Rule Mining &Mdash; a General Survey and Comparison", SIGKDD Explor Newsl 2, 2000, pp. 58–64, doi:10.1145/360402.360421.

[13] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, 1993, pp. 207–216, doi: 10.1145/170035.170072.

[14] B. Shahbaba, "Biostatistics with R: An Introduction to Statistics through Biological Data", 2012, Springer, doi: 10.1007/978-1-4614-1302-8.

[15] J.E. Rossouw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, and P. Jooste, "Coronary risk factor screening in three rural communities", South African Medical Journal, vol. 64, 1983, pp. 430–436.

[16] R. Kreuger, "ST Segment", ECGpedia.

[17] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Gandeharioun, and R. Boghrati, "Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms", IEEE 12th International Conference on Data Mining Workshop, 2012, pp. 9–16, doi: 10.1109/ICDMW.2012.29.

[18] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, pp. 459–468, doi: 10.1016/j.procs.2015.09.132.

[19] L. Verma, S. Srivastaa, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 178, 2016, doi: 10.1007/s10916-016-0536-z.

[20] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease", Computer Methods and Programs in Biomedicine, vol. 111, no. 1, 2013, pp. 52-61, doi: 10.1016/j.cmpb.2013.03.004.

[21] Ch. Yadav, S. Lade, and M. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining", International Journal of Computer Applications, vol. 87, no. 4, 2014, pp. 9-13.

[22] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples)", Biometrika, vol. 52, no. 3–4, 1965, pp. 591–611, doi: 10.1093/biomet/52.3-4.591.

[23] B. L. Welch, "On the Comparison of Several Mean Values: An Alternative Approach", Biometrika, vol. 38, 1951, pp. 330–336, doi: 10.2307/2332579.

[24] K. Pearson, Karl, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", Philosophical Magazine Series 5, vol. 50, no. 302, 1900, pp. 157–175, doi: 10.1080/14786440009463897.

[25] R. A. Fisher, "On the interpretation of $\chi 2$ from contingency tables, and the calculation of P", Journal of the Royal Statistical Society, vol. 85, no. 1,1922, pp. 87–94, doi: 10.2307/2340521.

[26] G. E. Batista, M.C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method", In Proceedings of Soft Computing Systems: Design, Management and Applications, IOS Press, 2002, pp. 251-260, doi=10.1.1.14.3558.

[27] Y. Dong, Ch-Y. J. Peng, "Principled missing data methods for researchers", Springerplus, vol. 2, vol. 222, 2013, doi: 10.1186/2193-1801-2-222.

[28] D. Freedman, "Statistical Models: Theory and Practice. Cambridge", New York: Cambridge University Press, 2009, doi: 10.1017/CBO9780511815867.

[29] H. B. Mann, D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other", Annals of Mathematical Statistics, vol. 18, no. 1, 1947, pp. 50–60, doi: 10.1214/aoms/1177730491.

[30] P. Drotár, Z. Smékal, "Comparative Study of Machine Learning Techniques for Supervised Classification of Biomedical Data", Acta Electrotechnica et Informatica, vol. 14, no. 3, 2014, pp. 5-10, doi: 10.15546/aeei-2014-0021