VŠE

FAKULTA
INFORMATIKY
A STATISTIKY

2. Project

Bc. Danila Smaliak

Bc. Petr Kalina

Table of contents

List of Figures

List of Tables

Introduction

This project focuses on the analysis of the *HeartBit data cz.csv* dataset, which contains medical and fitness-related information about 469 patients diagnosed with heart failure. The dataset is notable for combining standard clinical variables (such as diagnoses, treatments, and laboratory results) with measurements from physical performance tests (such as walking distance and treadmill endurance), offering a multidimensional view of the patient's condition.

To structure the analysis, we apply the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology — one of the most widely used frameworks in the field of data science. CRISP-DM defines six interconnected phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This iterative and flexible process ensures that the analysis remains aligned with its objectives while allowing room for adaptation as new insights emerge. By following this structure, we aim to better manage complexity and produce interpretable, educational outcomes (Shearer, 2000).

The primary goal of this project is exploratory. Rather than aiming to build high-accuracy predictive models, we seek to evaluate the potential of selected data mining techniques — such as classification, clustering, and association rule learning — to extract meaningful patterns from the dataset. The project also reflects on typical challenges in working with healthcare data, including missing values, class imbalance, and the interpretability of results in a clinical context.

“Business” Understanding

Heart failure is a chronic disease that significantly reduces patients’ physical capabilities and quality of life. In clinical practice, functional capacity is commonly assessed using scales like NYHA or exercise-based tests, but a deeper understanding of the connections between clinical factors, comorbidities, and physical performance remains a challenge.

The goal of this project is to explore patterns in patient data and assess whether selected clinical and lifestyle variables can help explain or predict a patient’s physical condition, using standard data mining techniques. The dataset combines various categories of information — including clinical diagnoses, treatments, blood test results, and fitness test outcomes — allowing for initial experimentation with classification, segmentation, and association rule discovery.

From a business perspective, the purpose of this analysis is not to produce deployable tools or precise models, but rather to:

- Experiment with data-driven approaches in the healthcare context,
- Identify possible predictors of physical capacity and their limitations,
- Gain hands-on experience in handling real-world medical data,
- Assess how suitable different modeling techniques are for this type of task.

The findings may help generate hypotheses for future clinical studies or inspire follow-up analyses with larger and more complete datasets. The project also highlights typical challenges in working with healthcare data — such as missing values, variable redundancy, and class imbalance — which are critical to address in any real-world application.

Data Understanding

Overview of the Dataset

The dataset used in this project — *HeartBit data cz.csv* — contains clinical and fitness-related data from 469 patients diagnosed with chronic heart failure. It includes a total of 63 variables, capturing a wide range of information: demographic data, comorbidities, treatments, lab values, echocardiographic measurements, and results from physical performance tests such as treadmill exercises or the 6-minute walk test.

All variables are documented in an official specification, which includes the name, data type, and medical category of each field. A sample of the official documentation is shown below:

Variable	Description	Data Type	Category
DEATH?	Whether the patient is alive or dead	Binary	Clinical
AGE	Age at the time of examination	Numeric	Demographic
BMI	Body mass index	Numeric	Anthropometry
MI	History of myocardial infarction	Binary	Comorbidities
BB	Treated with beta blockers	Binary	Treatment
HB	Hemoglobin level in blood	Numeric	Biochemistry
CPX.TIME	Duration of treadmill exercise	Numeric	Fitness level

Table 1. Sample of the description of the variables

All 63 variables are divided into the following functional groups:

- **Clinical**: Includes key diagnostics, classification systems (e.g. NYHA), heart ultrasound values, blood pressure, and heart rate.
- **Demographic**: Basic patient characteristics like age.
- **Anthropometry**: Body height, weight, and BMI.
- **Comorbidities**: Presence of conditions like diabetes, stroke, atrial fibrillation, etc.
- **Treatment**: Information about medications prescribed, such as diuretics, statins, beta blockers.
- **Biochemistry**: Lab values from blood samples, e.g. sodium, potassium, BNP (heart failure marker).
- **Fitness level**: Performance in physical tasks — walking tests, flexibility, treadmill duration, oxygen consumption.

This logical structure makes it possible to isolate factors that may influence patient outcomes or physical performance, and later use these groupings for targeted analyses. The following is a preview of the actual dataset, taken from the first five rows:

ID	AGE	NYHA	BMI	REST.HR	EXERCISE1	CPX.TIME
HB1	46.99	2	40.1	72	8	14.52
HB2	47.33	2	25.2	72	6	17.27
HB3	59.85	1	25.5	76	1	15.19
HB4	61.19	3	29.6	64	20	13.45
HB5	23.73	1	30.5	92	1	16.38

Table 2. Snapshot of raw data

The data is stored in European number formatting (decimal comma), which must be standardized during preprocessing. Also, missing values are present in multiple variables, a challenge that will be addressed in the data preparation phase.

Variable Types and Distributions

The raw dataset contains 469 rows and a wide variety of attributes that differ not only in content but also in structure and format. Most variables are either numeric (e.g. age, weight, lab values), binary encoded (0/1 for yes/no), or categorical (e.g. NYHA classification). A smaller portion includes date values, such as date of examination or birth.

Type	Example variables	Notes
Numeric	AGE, BMI, HB, BNP, REST.HR, METS	Often stored as strings with commas (e.g. `"13,6"`)
Binary	DEATH?, MI, AF, ANTIPLAT, DIGOX	1 = yes, 0 = no; encoded as numeric (0/1)
Categorical	NYHA, WEBER, MR	Some have unexpected values (e.g. `1.5`, `3.5`)
Date	DOB, DOE, DEATHDATE	Format: DD.MM.YYYY

Table 3. Variables types

A considerable number of variables contain **missing data**. These are represented either as **blank cells** or "?" and are often concentrated in groups (e.g., most CPX and 6MWT values are missing for some patients). These patterns suggest that not all tests were performed on every patient, and some sections (like treadmill-based measurements) were skipped entirely for subsets of the cohort.

We observe high variability in key numerical variables:

- AGE ranges from ~20 to 90 years, with most patients between 50–70.
- BMI values range from underweight (~17) to morbidly obese (>40).
- BNP, an indicator of heart failure, shows extreme variation — from low hundreds to values above 10,000 in some cases.
- CPX,TIME, treadmill exercise duration, ranges from 3 to 17 minutes for patients with available data.

Distributions will be visualized in the following section, but it's already clear that many attributes contain long tails, skewed distributions, and non-normal shape, which must be handled with care in modeling.

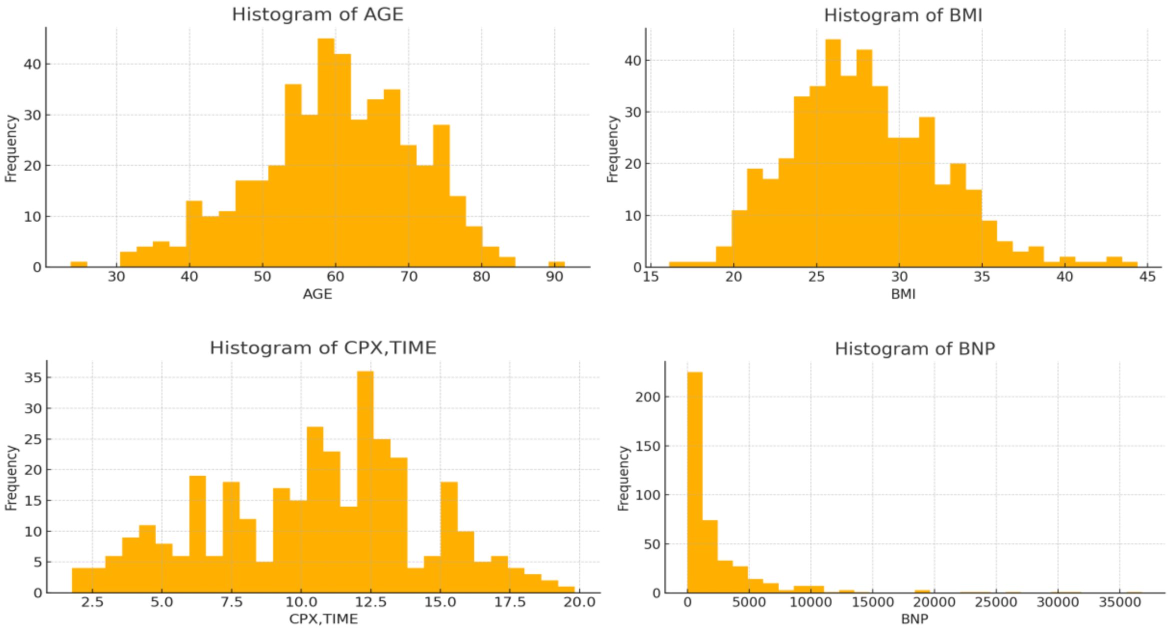


Figure 1. Histograms of chosen atributes

A significant portion of the HeartBit dataset contains missing values, with their presence varying considerably across both columns and individual records. These missing entries appear as blank cells or, in some cases, the "?" character, and are often clustered in specific variable groups.

Column-level patterns

Missingness is particularly prevalent among variables related to physical performance testing, such as those connected to the 6-minute walk test (6MWT) and treadmill-based CPX testing. These tests may not have been performed for all patients due to health restrictions or clinical decisions, leading to substantial data gaps.

For example:

Variables like CPX,TIME, CPX,PEAKVO2, and 6MWT,HR2 show over 60% missing values.

Blood-based markers such as BNP and CRP also contain around 25–30% missing entries.

On the other hand, basic demographic and diagnostic variables — such as AGE, BMI, and NYHA — are complete or nearly complete.

Row-level patterns

Missing values are not randomly distributed. Some records are nearly complete, while others have missing values in large contiguous blocks, particularly for patients who did not undergo physical exercise testing. This reflects real-world heterogeneity in medical data collection, where not every test is applicable or feasible for every patient.

Correlated Variables and Redundancy

In a medical dataset like HeartBit, it is common for certain variables to reflect overlapping or derived information. Recognizing this redundancy is a crucial step in understanding the dataset, as it can lead to biased models or overrepresentation of specific clinical signals if not addressed properly.

Rather than conducting a full statistical correlation analysis at this stage, we relied on logical analysis based on variable definitions provided in the dataset documentation. For example:

- BMI is mathematically derived from weight and height.
- AGE is based on date of birth and examination date.
- WEBER classification is based on peak VO_2.
- SLOPE>35 and PEAK>18 are binary versions of continuous indicators (SLOPE, CPX,PEAKVO2FORBM).

During data preparation, several of these redundant variables were explicitly removed to reduce collinearity. This included BMI, CPX,PEAKVO2, PEAK>18, SLOPE>35, and others. The decision to keep or exclude certain variables was based on clinical relevance and the need to avoid duplicating the same signal. To further illustrate the issue of redundancy, we generated a correlation heatmap of selected numerical variables known to be related. The following plot includes variables such as BMI, WEIGHT,KG, HEIGHT,CM, CPX,PEAKVO2, CPX,PEAKVO2FORBM, SLOPE, and AGE.

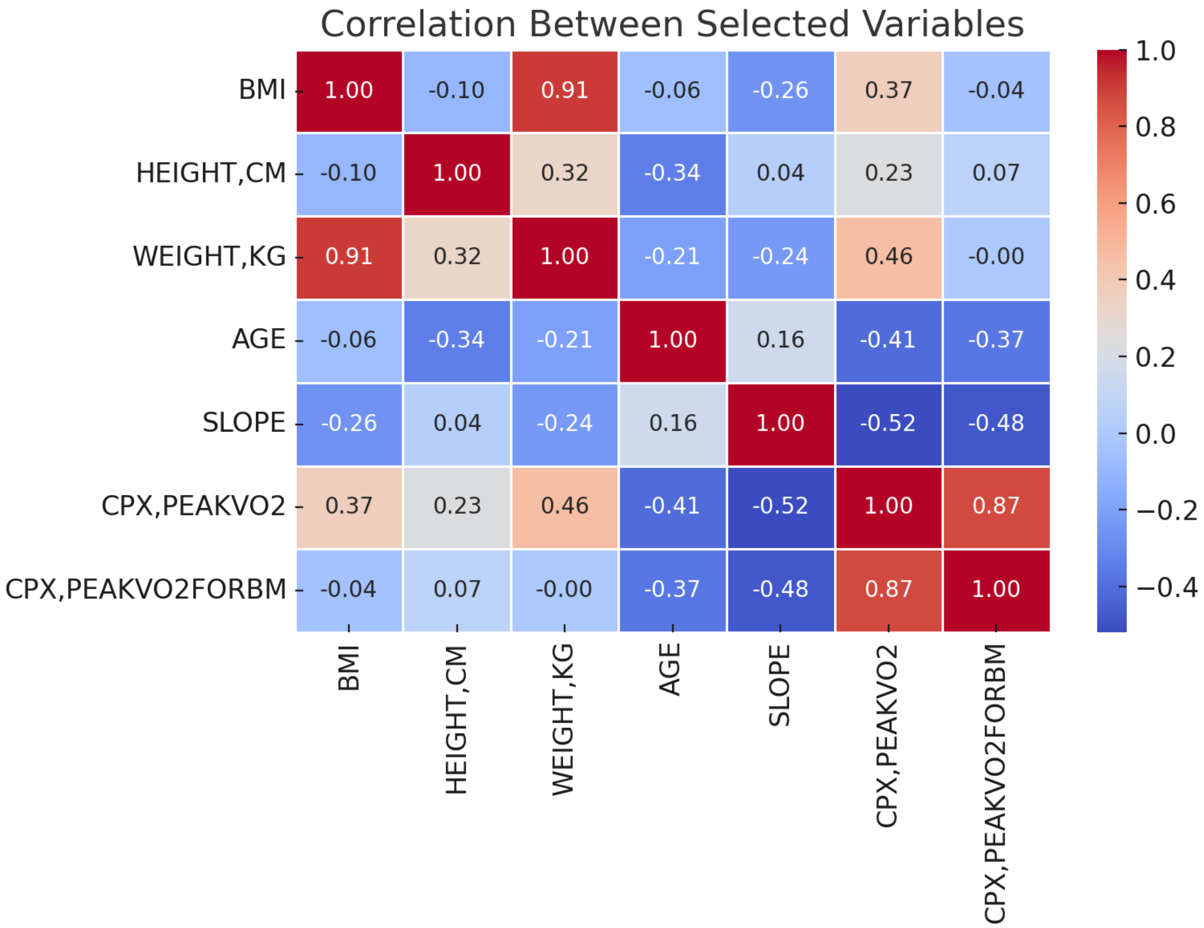


Figure 2. Correlation Between Selected Variables

- BMI has near-perfect correlation with WEIGHT,KG, as expected.
- CPX,PEAKVO2 and CPX,PEAKVO2FORBM are highly correlated.
- SLOPE also correlates with CPX-based performance metrics.
- AGE shows only weak correlation with other variables in this subset.

This visualization confirms that many performance-related variables are not independent and that a thoughtful feature selection strategy is needed.

Redundant and derived variables can distort modeling results if left unchecked. While some of them were removed prior to modeling, the observed correlations reinforce the importance of understanding variable relationships early in the process — both logically and visually.

Initial Insights from Fitness Measures

One of the most distinctive aspects of the HeartBit dataset is the inclusion of fitness-related variables. These variables reflect the physical performance of patients with heart failure and include both objective measurements (e.g., walking distance, exercise repetitions) and subjective assessments (e.g., fatigue, dyspnea after exertion).

Group	Example variables	Description
Performance metrics	CPX,TIME, 6MWT,DIST, EXERCISE1–3	Duration, distance, or number of repetitions during physical activity
Subjective assessments	6MWT,FATIGUE, 6MWT,DYSPN	Patient-rated fatigue or breathlessness on a 0–10 scale
Flexibility/mobility	EXERCISE4, EXERCISE5	Range of motion tests in cm, including negative values if not completed

Table 4. Types of fitness measures

Preliminary inspection of these variables reveals considerable variability among patients:

- 6MWT,DIST (6-minute walk test) values range from under 100 meters to over 600 meters, showing a wide spread in physical capacity.
- CPX,TIME (treadmill duration) ranges between 3 and 17 minutes, with a skew toward shorter times, indicating early exhaustion in many patients.
- Repetition-based exercises (e.g., EXERCISE2, EXERCISE3) vary from 1 to over 25 repetitions, but are missing in a substantial number of cases.
- EXERCISE4 and EXERCISE5 contain negative values, indicating how far a patient fell short of completing the movement — a useful feature, but one that needs careful interpretation.
- Subjective fatigue (6MWT,FATIGUE) and dyspnea (6MWT,DYSPN) ratings cluster in the middle-to-high range (4–8 out of 10) for most cases.

Unfortunately, many of these variables suffer from high levels of missing data, especially for the CPX treadmill-based tests. This limits their usability in modeling but still allows for exploratory group comparisons and descriptive analysis.

Despite the gaps, these fitness measures are central to understanding the functional state of heart failure patients and were used as targets or explanatory variables in later parts of the project.

Target Variable Considerations

Among the various clinical variables in the dataset, the NYHA classification stood out as the most meaningful and suitable target for supervised learning. It reflects the functional severity of heart failure based on patients' symptoms during physical activity, following a well-established ordinal scale from 1 (mild) to 4 (severe).

NYHA is widely used in both clinical practice and research. Its interpretability and availability across nearly all records made it the logical foundation for modeling.

However, to reduce noise and improve consistency, we later constructed a refined version called NYHA_UPDATED, which grouped intermediate and ambiguous values (such as 1.5, 2.5, and 3.5) into broader categories. This step helped stabilize class definitions while preserving clinical meaning. The final models in this project used NYHA_UPDATED as the main target variable.

As shown, most patients fall into Classes II and III. Classes I and IV are relatively rare, which reflects the typical outpatient population profile.

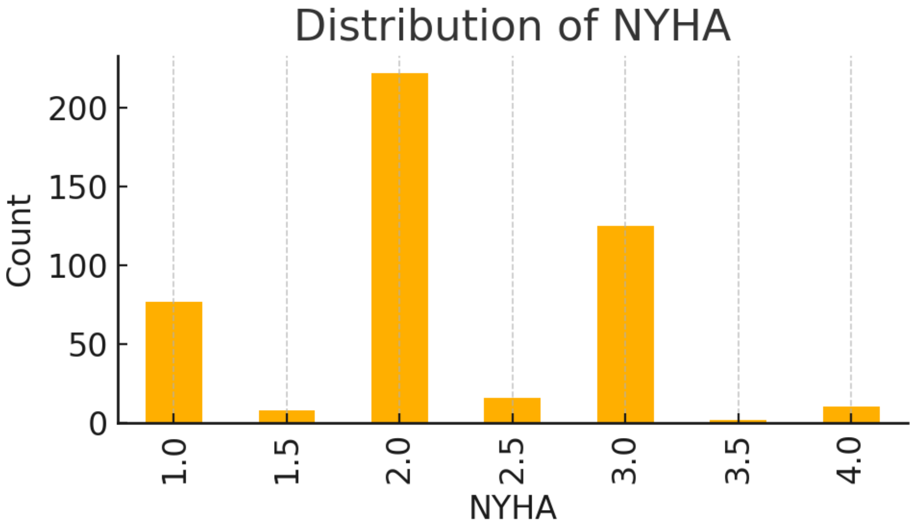


Figure 3. NYHA

Data Preparation

Missing Values

As previously said, we chose the file *HeartBit data cz.csv* as our dataset. First, we have to deal with missing data in the dataset. BigML can deal with missing data automatically,[1] but we still want to delete instances where at least 1/3 of data are missing.

To do that, we opened the file *HeartBit data cz.csv* in Excel. Once there, we created an additional column called *Number of Missing Data Cells.* Next, we put the following formula into the first row of the new column:

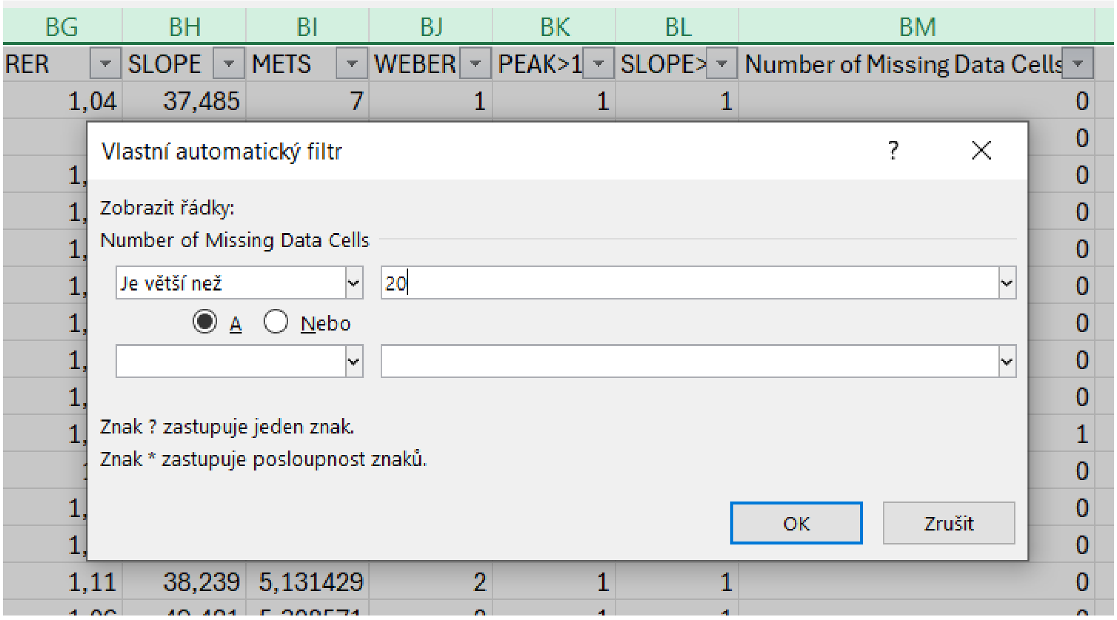
=COUNTIF(A2:BL2; "?") + COUNTBLANK(A2:BL2)

This formula counts the number of the value "?" and the number of blank cells in the first row. The formula was then adjusted to fill the whole column and to count the missing values for all the rows.

Next, we searched the dataset[2] for instances with more than 20 attribute values missing. With the dataset having 63 variables, this translates to rows with approximately 1/3 of values missing.

[1] *BigML can handle missing data as input to obtain predictions and also in your training data to build models. Generally, missing values are treated as a unique value in itself, not mapped to any other value (either 0, the mean, or whatever).* (BigML, 2024)

[2] To do that, we used the Excel function *Filter*.



BG	BH	BI	BJ	BK	BL	BM
RER	SLOPE	METS	WEBER	PEAK>1	SLOPE>	Number of Missing Data Cells
1,04	37,485	7	1	1	1	0
1,11	38,239	5,131429	2	1	1	0

Vlastní automatický filtr

Zobrazit řádky:

Number of Missing Data Cells

Je větší než: 20

● A ○ Nebo

Znak ? zastupuje jeden znak.

Znak * zastupuje posloupnost znaků.

Figure 4. Data Preparation in Excel (source: authors)

Out of the 469 rows (patients), 31 rows had more than 20 missing attribute values. These instances were then deleted. The dataset was then left with 438 remaining instances with tolerable levels of missing data.

Another problem to be solved was the format of remaining empty values – some were represented as “?” and some as blank cells. BigML prefers the format “?” for empty values, so we replaced all the blank cells in the dataset with the value “?”. The number of blank cells thus replaced was 1195.[3]

BigML Data Transformation

Next, we uploaded the edited dataset to BigML. The first step was to **set the data types** of the columns in the module *Configure source.* The type of the following attributes was changed from *numeric* to *categorical:*

NYHA

DIGOX

WEBER

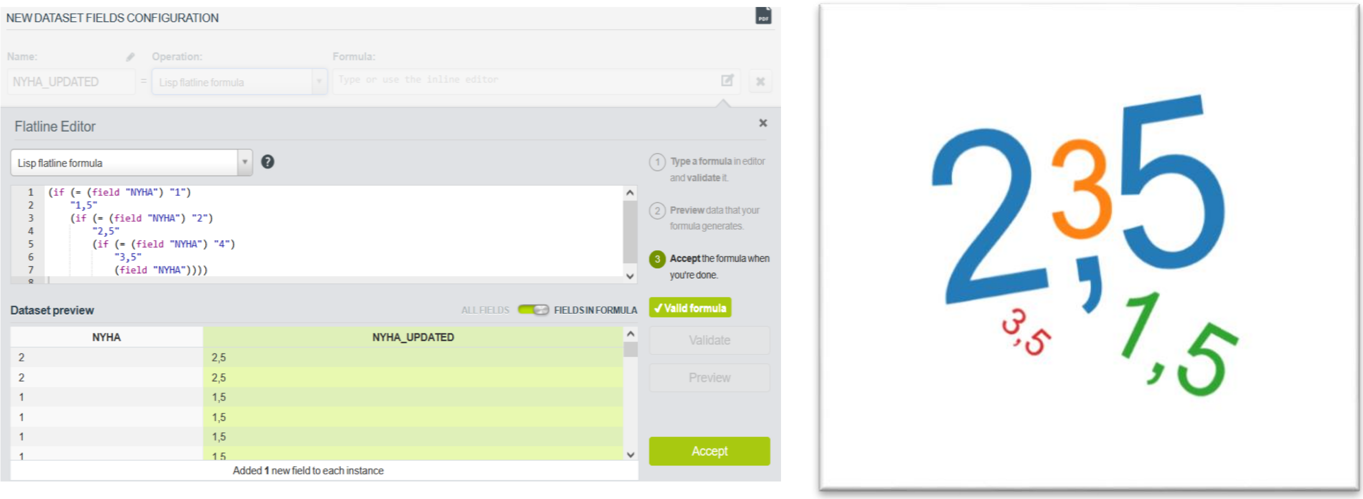
[3] We could have replaced the blank cells already in step 1, but we did not realize this problem until later.

Before we created the dataset, we also deselected unsuitable attributes. These “deleted” attributes can be found in Appendix A. We left the attribute HEIGHT.CM in the dataset, even though it is highly correlated with the attribute BMI.[4] We decided to keep it in because bigger bodily proportions are more taxing for the heart, and height could therefore be a relevant attribute.

After the creation of the dataset, we created a new attribute – *NYHA_UPDATED*. In this new attribute, we merged the values 1, 1.5, 2, 2.5, 3, 3.5, and 4 in the following way: we merged 1 with 1.5, 2 with 2.5, and 4 with 3.5. We did this by going to *Transform Dataset* and *Add Fields.* There we named the attribute NYHA_UPDATED and gave it the following Lisp definition:

```
(if (= (field "NYHA") "1")
    "1,5"
    (if (= (field "NYHA") "2")
        "2,5"
        (if (= (field "NYHA") "4")
            "3,5"
            (field "NYHA"))))
```

Below is the output of this Lisp code and the new value distribution of NYHA_UPDATED:



NEW DATASET FIELDS CONFIGURATION

Name: NYHA_UPDATED = Operation: Lisp flatline formula Formula: Type or use the inline editor

Flatline Editor

Lisp flatline formula

```
(if (= (field "NYHA") "1")
    "1,5"
    (if (= (field "NYHA") "2")
        "2,5"
        (if (= (field "NYHA") "4")
            "3,5"
            (field "NYHA"))))
```

1. Type a formula in editor and validate it.
2. Preview data that your formula generates.
3. Accept the formula when you're done.

Dataset preview

ALL FIELDS FIELDS IN FORMULA

NYHA	NYHA_UPDATED
2	2,5
2	2,5
1	1,5
1	1,5
1	1,5
1	1,5

Added 1 new field to each instance

Figure 5. Lisp Code and NYHA_UPDATED Value Distribution (source: authors)

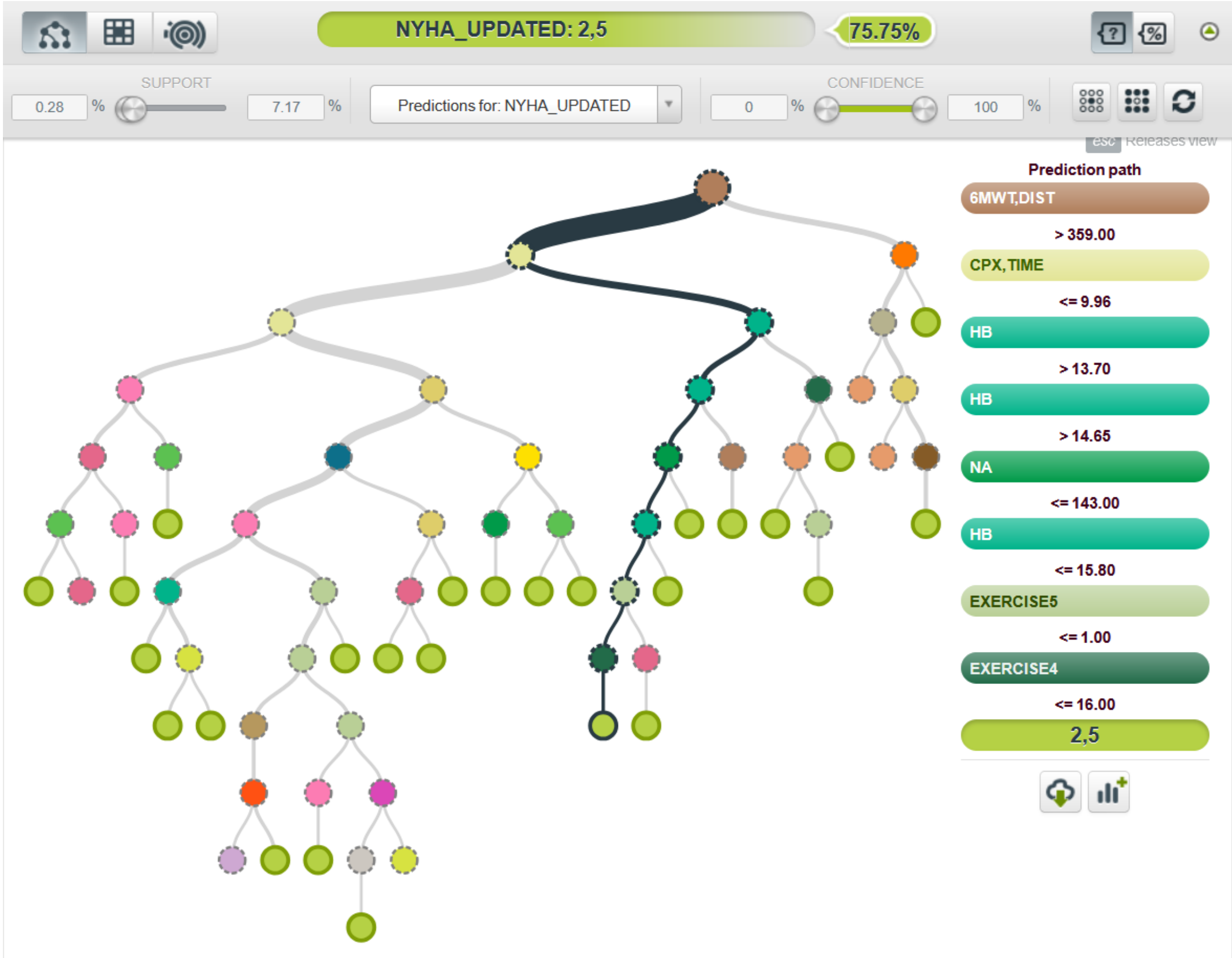
[4] The attribute WEIGHT.KG was deleted because of that correlation.

Next, we created a **Training | Test Split** – 80 % of the data was determined as training data and 20 % as testing data. Of course, we deselected the attribute NYHA prior to that, as it was now superseded by NYHA_UPDATED.

Modelling

Decision tree

We chose the *decision tree* as our first data mining algorithm. To create the tree, **we selected NYHA_UPDATED as our target attribute** and selected the option *Model* in *1-Click Supervised.*



NYHA_UPDATED: 2,5

75.75%

SUPPORT 0.28 % 7.17 %

Predictions for: NYHA_UPDATED

CONFIDENCE 0 % 100 %

Prediction path

6MWT,DIST

> 359.00

CPX,TIME

<= 9.96

HB

> 13.70

HB

> 14.65

NA

<= 143.00

HB

<= 15.80

EXERCISE5

<= 1.00

EXERCISE4

<= 16.00

2,5

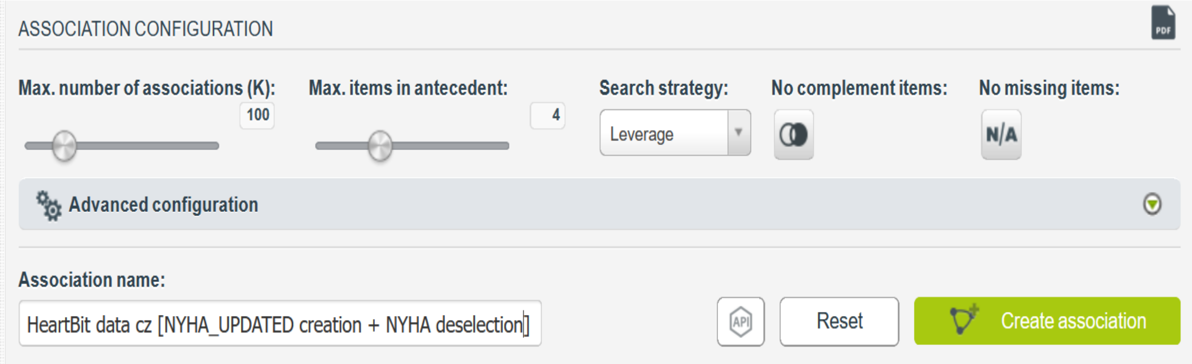
Figure 6. Decision Tree Model (source: authors)

Segmentation

We also experimented with clustering, choosing the algorithm k-means and number of clusters 5. The result was not really valuable, and the model did not meet our sufficiency levels. We therefore provide only its summary report in Appendix D.

Association Rules

We also chose association rules as our data mining method. Below we can see the parameters of our association:



ASSOCIATION CONFIGURATION

Max. number of associations (K): 100

Max. items in antecedent: 4

Search strategy: Leverage

No complement items:

No missing items: N/A

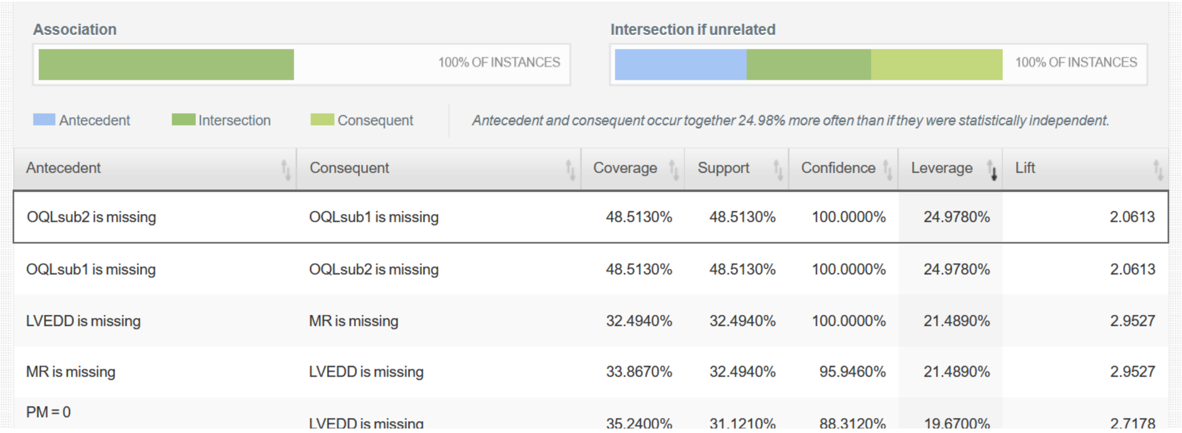
Advanced configuration

Association name: HeartBit data cz [NYHA_UPDATED creation + NYHA deselection]

Figure 7. Association Parameters (source: authors)

We chose the *Leverage* strategy to find interesting associations that are statistically significant – items that occur together more than might be expected. Of course, we had to deselect the attribute NYHA again, as this was replaced.

In the first round, we selected the option to include missing items. This was a mistake, as the most common association rules then were, for instance, "if OQLsub2 is missing, then OQLsub1 is missing":



Antecedent and consequent occur together 24.98% more often than if they were statistically independent.

Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
OQLsub2 is missing	OQLsub1 is missing	48.5130%	48.5130%	100.0000%	24.9780%	2.0613
OQLsub1 is missing	OQLsub2 is missing	48.5130%	48.5130%	100.0000%	24.9780%	2.0613
LVEDD is missing	MR is missing	32.4940%	32.4940%	100.0000%	21.4890%	2.9527
MR is missing	LVEDD is missing	33.8670%	32.4940%	95.9460%	21.4890%	2.9527
PM = 0	LVEDD is missing	35.2400%	31.1210%	88.3120%	19.6700%	2.7178

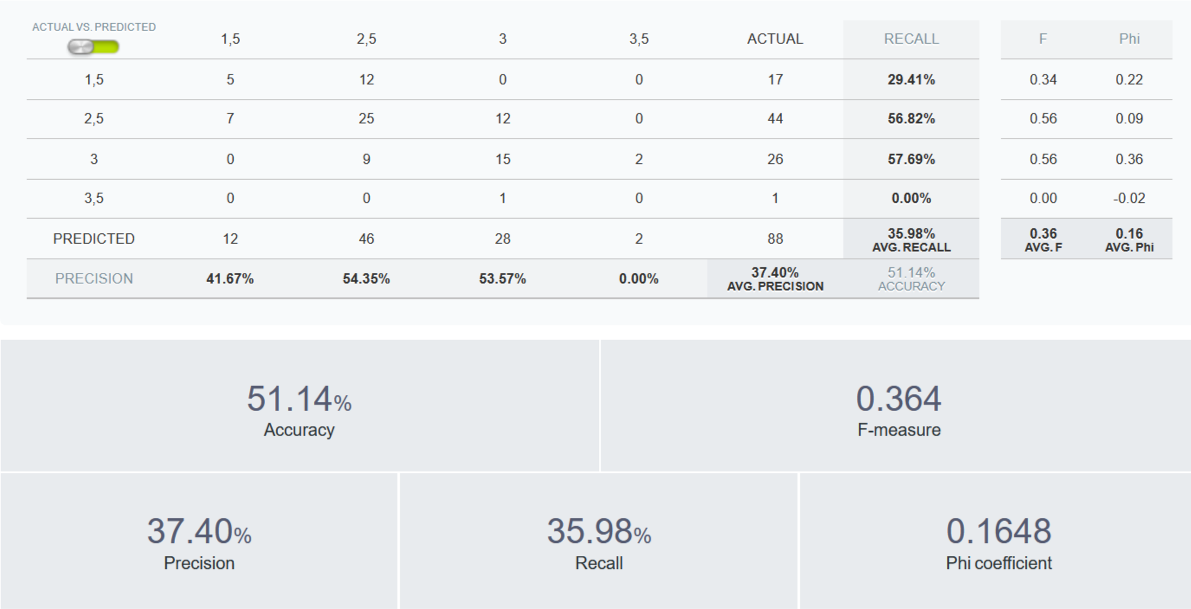
Figure 8. Association: First Try (source: authors)

We therefore created the association again, this time without allowing missing item

Evaluation

Decision Tree

After its creation, the decision tree was evaluated using the testing data subset. The result of this evaluation can be seen below.



ACTUAL VS. PREDICTED	1,5	2,5	3	3,5	ACTUAL	RECALL	F	Phi
1,5	5	12	0	0	17	**29.41%**	0.34	0.22
2,5	7	25	12	0	44	**56.82%**	0.56	0.09
3	0	9	15	2	26	**57.69%**	0.56	0.36
3,5	0	0	1	0	1	**0.00%**	0.00	-0.02
PREDICTED	12	46	28	2	88	**35.98%** **AVG. RECALL**	**0.36** **AVG. F**	**0.16** **AVG. Phi**
PRECISION	**41.67%**	**54.35%**	**53.57%**	**0.00%**	**37.40%** **AVG. PRECISION**	51.14% ACCURACY		

51.14% Accuracy

0.364 F-measure

37.40% Precision

35.98% Recall

0.1648 Phi coefficient

Figure 9. Decision Tree Characteristics (source: authors)

Here are the values of *Lift* for each category of *NYHA_UPDATED:*

1,5:	215.69 %
2,5:	108.70 %
3:	181.32 %
3,5:	0.00 %

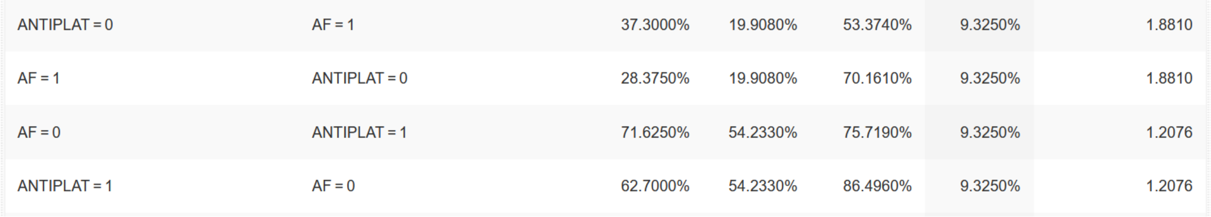
As we can see, the results of the decision tree are relatively poor. Accuracy of 51.14 % means that the model is only slightly better than random guessing, which is not the result we have hoped for. Precision and Recall are also very low – 37.4 % and 35.98 % respectively. The *Lift* values, at least, are looking better – the model is approximately 2.16 times better than random guessing for the value **1.5**.

The decision tree is probably **underfitting** – it is too simple to find complex patterns in our data. We evaluate this model as **not really sufficient.**

Association rules

The first several rules with the highest leverage do not provide us with much helpful information. They link attributes together, that are highly correlated *apriori*, e.g. “if 6MWT,HR2 > 80, then 6MWT,HR1 > 98”, or “if 6MWT,SBP1 <= 111, then 6MWT,SBP2 <= 105”.[5]

One interesting rule is the following:

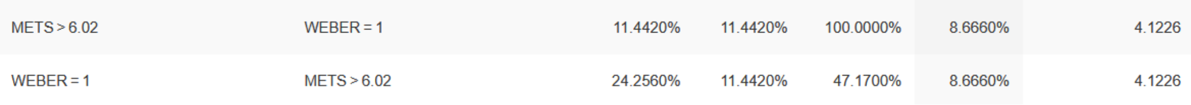


ANTIPLAT = 0	AF = 1	37.3000%	19.9080%	53.3740%	9.3250%	1.8810
AF = 1	ANTIPLAT = 0	28.3750%	19.9080%	70.1610%	9.3250%	1.8810
AF = 0	ANTIPLAT = 1	71.6250%	54.2330%	75.7190%	9.3250%	1.2076
ANTIPLAT = 1	AF = 0	62.7000%	54.2330%	86.4960%	9.3250%	1.2076

Figure 10. AF, ANTIPLAT rule (source: authors)

It means, for instance, that if the patient has atrial fibrillation, he has a lower chance of taking antiplatelet drugs, if he does not suffer from atrial fibrillation, he has higher chance of taking antiplatelet drugs etc. All of these four rules have a high confidence (the exception having only 53.37 % confidence) and relatively high levels of lift and leverage – all of the instances have leverage around 9.33 % which is statistically significant.

Another pattern is the following. The rule “if METS[6] > 6.02, then WEBER = 1” has the confidence of 100 %, lift is 4.1226 and leverage is 8.6660 %:



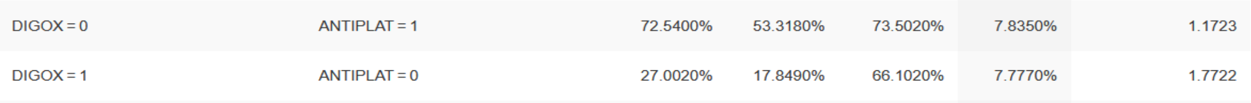
METS > 6.02	WEBER = 1	11.4420%	11.4420%	100.0000%	8.6660%	4.1226
WEBER = 1	METS > 6.02	24.2560%	11.4420%	47.1700%	8.6660%	4.1226

Figure 11. METS, WEBER rule (source: authors)

Another rule can be found bellow. It says that the treatment with digoxin and the treatment with antiplatelet drugs tend to exclude each other with high confidence (both around 70 %) and relatively high leverage (both around 7.8 %) and lift (both around 1.172):

[5] Thus connecting „systolic blood pressure before the walking test” with “systolic blood pressure after the walking test” and “heart rate before the walking test” with “heart rate after the walking test”.

[6] “Number of metabolic equivalents (level of work performed by the patient during exercise using a treadmill)”.



DIGOX = 0	ANTIPLAT = 1	72.5400%	53.3180%	73.5020%	7.8350%	1.1723
DIGOX = 1	ANTIPLAT = 0	27.0020%	17.8490%	66.1020%	7.7770%	1.7722

Figure 12. DIGOX, ANTIPLAT rule (source: authors)

Generally speaking, rules describe expected patterns linking attributes indicating physical fitness, linking a health issue and its treatment etc. The summary of the association can be found in Appendix C.

Deployment

The results of our data mining will require domain knowledge experts who could evaluate whether the association rules, for instance, are relevant or not. Is it relevant, that the treatment with digoxin and the treatment with antiplatelet drugs tend to exclude each other? This requires medical expertise to properly evaluate.

For more precise models and more valuable results we would need a dataset that is 1) larger and 2) has better data quality.[7]

[7] As it stands now, our dataset contained a lot of empty values.

Sources

Shearer, C. (2000). *The CRISP-DM model: The new blueprint for data mining*. *Journal of Data Warehousing,* *5*(4), 13–22. [Online] Retrieved from https://mineracaodedados.wordpress.com/wp-content/uploads/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf [accessed 2025-05-25]

BigML (2024). *How does BigML handle missing values to predict with your models and ensembles?*. [Online]. 2024. In: BigMl.com. Retrieved from: https://support.bigml.com/hc/en-us/articles/206616349-How-does-BigML-handle-missing-values-to-predict-with-your-models-and-ensembles [accessed 2025-05-30].

Appendix A – Deselected Attributes Before the Creation of the Dataset

1. **ID**
2. **DEATHDATE** date of death (if death=1) or date of the confirmation that the patient is still alive
3. **TIMEFU** number of days between examination and date death or date of the confirmation that the patient is still alive
4. **QOL** result of the survey measuring the quality of life (QoL, total score range 0–105, from best to worst)
5. **DOB** date of birth
6. **DOE** date of the examination
7. **WEIGHT.KG** body mass in kg
8. **AETH.HF** information about the clinical cause of heart failure (1=ischemic disease or 2 = other)
9. **CPX.PEAKVO2** peak oxygen consumption during exercise testing on a treadmill
10. **CPX.PEAKVO2FORBM** peak oxygen consumption during exercise testing on a treadmill per body mass
11. **PEAK>18** Dividing patients based on a cutoff value of peak oxygen consumption (used for Weber)
12. **SLOPE>35** Dividing patients based on a cutoff value of slope (used for Weber)
13. **Number of Missing Data Cells** The newly created column

Appendix B – Decision Tree: Model Summary Report

Data distribution:

1,5: 18.05% (63 instances)

2,5: 51.58% (180 instances)

3: 27.22% (95 instances)

3,5: 3.15% (11 instances)

Predicted distribution:

1,5: 12.03% (42 instances)

2,5: 60.46% (211 instances)

3: 25.79% (90 instances)

3,5: 1.72% (6 instances)

Field importance:

1. CPX,TIME: 27.85%
2. 6MWT,DIST: 25.21%
3. EXERCISE1: 7.63%
4. HB: 6.83%
5. BMI: 6.52%
6. EXERCISE5: 5.77%
7. DIUR: 4.67%
8. OQLsub2: 2.68%
9. METS: 2.23%
10. BNP: 1.34%
11. NA: 1.31%
12. REST,HR: 1.29%
13. EXERCISE4: 1.16%
14. CRP: 1.13%
15. DEATH?: 1.05%
16. 6MWT,DYSPN: 1.00%

17. MI: 0.65%
18. EXERCISE3: 0.64%
19. AGE: 0.40%
20. EXERCISE2: 0.27%
21. REST,SBP: 0.22%
22. HT: 0.18%

Rules summary:

1,5: (data 18.05% / prediction 12.03%) 6MWT,DIST > 359 and CPX,TIME > 9.96289

· 28.57%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 and METS > 5.63429 and BNP <= 2562.5 [Confidence: 75.75%]

· 16.67%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 <= 4.53889 and DEATH? = 1 and NA > 137 [Confidence: 64.57%]

· 11.90%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 [Confidence: 54.28%; impurity: 0.21%]

· 7.14%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 <= 4.53889 and DEATH? = 0 and BNP <= 138.5 [Confidence: 43.85%]

· 7.14%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 5.35 and DIUR = 0 and METS > 5.22286 [Confidence: 43.85%]

· 7.14%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and -11 < EXERCISE5 <= -7 and MI = 0 and REST,SBP <= 130 [Confidence: 43.85%]

· 7.14%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 and METS > 5.63429 [Confidence: 67.20%; impurity: 0.10%]

· 4.76%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI <= 24.75 and BNP > 1713 [Confidence: 34.24%]

· 2.38%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 5.35 and DIUR = 0 [Confidence: 25.05%; impurity: 0.24%]

· 2.38%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and 26.25 < BMI <= 26.52917 and -17 < EXERCISE5 <= -11 [Confidence: 20.65%]

· 2.38%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and 24.75 < BMI <= 26.75 and 4.94286 < METS <= 5.63429 [Confidence: 20.65%]

· 2.38%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 and 6.25143 < METS <= 6.81714 and BNP > 2562.5 [Confidence: 20.65%]

2,5: (data 51.58% / prediction 60.46%)

· 32.23%: 6MWT,DIST > 359 [Confidence: 53.06%; impurity: 0.28%]

· 9.48%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI > 26.52917 and HB > 15 [Confidence: 83.89%]

· 7.11%: 6MWT,DIST > 405 and CPX,TIME <= 9.96289 and 13.7 < HB <= 14.65 [Confidence: 79.61%]

· 6.64%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and EXERCISE5 > -7 and EXERCISE3 <= 19 and AGE <= 67.475 [Confidence: 78.47%]

· 6.64%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI > 26.52917 and HB <= 15 [Confidence: 48.41%; impurity: 0.22%]

· 5.69%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and 14.65 < HB <= 15.8 and NA <= 143 and EXERCISE5 <= 1 and EXERCISE4 <= 16 [Confidence: 75.75%]

· 5.69%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 <= 4.53889 and DEATH? = 0 and BNP > 138.5 [Confidence: 75.75%]

· 4.27%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI <= 24.75 and BNP <= 1713 [Confidence: 70.08%]

· 3.79%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI > 26.52917 and HB <= 15 and CRP <= 4.69 [Confidence: 67.56%]

· 2.84%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and 4.53889 < EXERCISE1 <= 5.35 and DIUR = 0 [Confidence: 60.97%]

· 2.84%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.25 and -17 < EXERCISE5 <= -11 [Confidence: 60.97%]

· 1.90%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 26.75 and METS <= 5.63429 [Confidence: 51.01%]

· 1.42%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB <= 13.7 and EXERCISE4 > -2 and 6MWT,DYSPN <= 2 [Confidence: 43.85%]

· 1.42%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 5.35 and DIUR = 0 and METS <= 5.22286 [Confidence: 43.85%]

· 0.95%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR <= 58 and 6MWT,DYSPN <= 8 [Confidence: 34.24%]

· 0.95%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 [Confidence: 41.92%; impurity: 0.25%]

· 0.95%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and -11 < EXERCISE5 <= -7 and MI = 1 and CRP <= 3.55 [Confidence: 34.24%]

· 0.95%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and EXERCISE5 > -7 and EXERCISE3 <= 19 and AGE > 67.475 and HT = 1 [Confidence: 34.24%]

· 0.95%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI > 26.52917 [Confidence: 69.86%; impurity: 0.14%]

· 0.47%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR > 58 and EXERCISE1 <= 9.35 and EXERCISE2 > 15 and MI = 1 [Confidence: 20.65%]

· 0.47%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB <= 13.7 and EXERCISE4 > -2 and 6MWT,DYSPN > 2 and EXERCISE5 > 7 [Confidence: 20.65%]

· 0.47%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and 14.65 < HB <= 15.8 and NA <= 143 and EXERCISE5 > 1 and METS <= 2.52571 [Confidence: 20.65%]

· 0.47%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 <= 4.53889 and DEATH? = 1 and NA <= 137 [Confidence: 20.65%]

· 0.47%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 <= 4.53889 and DEATH? = 0 [Confidence: 56.99%; impurity: 0.15%]

· 0.47%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI <= 24.75 [Confidence: 55.20%; impurity: 0.14%]

· 0.47%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 and METS > 6.81714 and BNP > 2562.5 [Confidence: 20.65%]

3: (data 27.22% / prediction 25.79%)

· 27.78%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR > 58 and EXERCISE1 <= 9.35 and EXERCISE2 <= 15 [Confidence: 86.68%]

· 15.56%: 6MWT,DIST <= 359 [Confidence: 54.97%; impurity: 0.24%]

· 8.89%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB <= 13.7 and EXERCISE4 <= -2 [Confidence: 67.56%]

· 7.78%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB <= 13.7 and EXERCISE4 > -2 and 6MWT,DYSPN > 2 and EXERCISE5 <= 7 [Confidence: 64.57%]

· 5.56%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB <= 13.7 and EXERCISE4 > -2 [Confidence: 33.18%; impurity: 0.25%]

· 5.56%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB > 14.65 and NA > 143 [Confidence: 56.55%]

· 3.33%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and 14.65 < HB <= 15.8 and NA <= 143 and EXERCISE5 > 1 and METS > 2.52571 [Confidence: 43.85%]

· 3.33%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and HB > 15.8 and NA <= 143 [Confidence: 43.85%]

· 3.33%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and EXERCISE5 <= -17 [Confidence: 43.85%]

· 3.33%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI > 26.52917 and HB <= 15 and CRP > 4.69 [Confidence: 43.85%]

· 2.22%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR > 58 and EXERCISE1 > 9.35 and 6MWT,DYSPN > 5 [Confidence: 34.24%]

· 2.22%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and -11 < EXERCISE5 <= -7 and MI = 0 and REST,SBP > 130 [Confidence: 34.24%]

· 2.22%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and EXERCISE5 > -7 and EXERCISE3 <= 19 and AGE > 67.475 and HT = 0 [Confidence: 34.24%]

· 2.22%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and EXERCISE5 > -7 and EXERCISE3 > 19 [Confidence: 34.24%]

· 1.11%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR > 58 and EXERCISE1 <= 9.35 and EXERCISE2 > 15 and MI = 0 [Confidence: 20.65%]

· 1.11%: 359 < 6MWT,DIST <= 405 and CPX,TIME <= 9.96289 and 13.7 < HB <= 14.65 [Confidence: 20.65%]

· 1.11%: 6MWT,DIST > 359 and CPX,TIME <= 9.96289 and 14.65 < HB <= 15.8 and NA <= 143 and EXERCISE5 <= 1 and EXERCISE4 > 16 [Confidence: 20.65%]

· 1.11%: 6MWT,DIST > 359 and 9.96289 < CPX,TIME <= 13.77385 and EXERCISE1 > 4.53889 and DIUR = 1 and BMI <= 26.52917 and -11 < EXERCISE5 <= -7 and MI = 1 and CRP > 3.55 [Confidence: 20.65%]

· 1.11%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and 24.75 < BMI <= 26.75 and METS <= 4.94286 [Confidence: 20.65%]

· 1.11%: 6MWT,DIST > 359 and CPX,TIME > 13.77385 and BMI > 24.75 and 5.63429 < METS <= 6.25143 and BNP > 2562.5 [Confidence: 20.65%]

3,5: (data 3.15% / prediction 1.72%) 6MWT,DIST <= 359

· 50.00%: 6MWT,DIST <= 359 and OQLsub2 > 18 [Confidence: 43.85%]

· 33.33%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR > 58 and EXERCISE1 > 9.35 and 6MWT,DYSPN <= 5 [Confidence: 34.24%]

· 16.67%: 6MWT,DIST <= 359 and OQLsub2 <= 18 and REST,HR <= 58 and 6MWT,DYSPN > 8 [Confidence: 20.65%]

Appendix C – Association Summary Report

Total number of rules: 100

Top 10 by Coverage:

Rule 000035: COPD = 0 -> STROKE = 0 [Coverage=88.33% (386); Support=83.29% (364); Confidence=94.30%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 000034: STROKE = 0 -> COPD = 0 [Coverage=86.04% (376); Support=83.29% (364); Confidence=96.81%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 00003b: PM = 0 -> DEATH? = 0 [Coverage=80.55% (352); Support=47.14% (206); Confidence=58.52%; Leverage=0.07141; Lift=1.17855; p-value=4.11522e-15]

Rule 000037: STATIN = 1 -> MI = 1 [Coverage=76.43% (334); Support=46.91% (205); Confidence=61.38%; Leverage=0.07209; Lift=1.18158; p-value=4.77768e-13]

Rule 000029: DIGOX = 0 -> ANTIPLAT = 1 [Coverage=72.54% (317); Support=53.32% (233); Confidence=73.50%; Leverage=0.07835; Lift=1.17227; p-value=6.13762e-14]

Rule 000047: DIGOX = 0 -> AF = 0 [Coverage=72.54% (317); Support=58.81% (257); Confidence=81.07%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 000054: DIGOX = 0 -> MI = 1 [Coverage=72.54% (317); Support=44.39% (194); Confidence=61.20%; Leverage=0.06713; Lift=1.17814; p-value=1.97638e-10]

Rule 00000c: AF = 0 -> ANTIPLAT = 1 [Coverage=71.63% (313); Support=54.23% (237); Confidence=75.72%; Leverage=0.09325; Lift=1.20763; p-value=7.76026e-19]

Rule 000046: AF = 0 -> DIGOX = 0 [Coverage=71.63% (313); Support=58.81% (257); Confidence=82.11%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 00000d: ANTIPLAT = 1 -> AF = 0 [Coverage=62.70% (274); Support=54.23% (237); Confidence=86.50%; Leverage=0.09325; Lift=1.20763; p-value=7.76026e-19]

Rule 000028: ANTIPLAT = 1 -> DIGOX = 0 [Coverage=62.70% (274); Support=53.32% (233); Confidence=85.04%; Leverage=0.07835; Lift=1.17227; p-value=6.13762e-14]

Top 10 by Support:

Rule 000034: STROKE = 0 -> COPD = 0 [Coverage=86.04% (376); Support=83.29% (364); Confidence=96.81%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 000035: COPD = 0 -> STROKE = 0 [Coverage=88.33% (386); Support=83.29% (364); Confidence=94.30%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 000046: AF = 0 -> DIGOX = 0 [Coverage=71.63% (313); Support=58.81% (257); Confidence=82.11%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 000047: DIGOX = 0 -> AF = 0 [Coverage=72.54% (317); Support=58.81% (257); Confidence=81.07%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 00000c: AF = 0 -> ANTIPLAT = 1 [Coverage=71.63% (313); Support=54.23% (237); Confidence=75.72%; Leverage=0.09325; Lift=1.20763; p-value=7.76026e-19]

Rule 00000d: ANTIPLAT = 1 -> AF = 0 [Coverage=62.70% (274); Support=54.23% (237); Confidence=86.50%; Leverage=0.09325; Lift=1.20763; p-value=7.76026e-19]

Rule 000028: ANTIPLAT = 1 -> DIGOX = 0 [Coverage=62.70% (274); Support=53.32% (233); Confidence=85.04%; Leverage=0.07835; Lift=1.17227; p-value=6.13762e-14]

Rule 000029: DIGOX = 0 -> ANTIPLAT = 1 [Coverage=72.54% (317); Support=53.32% (233); Confidence=73.50%; Leverage=0.07835; Lift=1.17227; p-value=6.13762e-14]

Rule 00003a: DEATH? = 0 -> PM = 0 [Coverage=49.66% (217); Support=47.14% (206); Confidence=94.93%; Leverage=0.07141; Lift=1.17855; p-value=4.11522e-15]

Rule 00003b: PM = 0 -> DEATH? = 0 [Coverage=80.55% (352); Support=47.14% (206); Confidence=58.52%; Leverage=0.07141; Lift=1.17855; p-value=4.11522e-15]

Rule 000036: MI = 1 -> STATIN = 1 [Coverage=51.94% (227); Support=46.91% (205); Confidence=90.31%; Leverage=0.07209; Lift=1.18158; p-value=4.77768e-13]

Top 10 by Confidence:

Rule 000014: METS > 6.02 -> WEBER = 1 [Coverage=11.44% (50); Support=11.44% (50); Confidence=100.00%; Leverage=0.08666; Lift=4.12264; p-value=2.81532e-36]

Rule 000019: 3.29 < METS <= 4.28 -> WEBER = 3 [Coverage=11.90% (52); Support=11.90% (52); Confidence=100.00%; Leverage=0.08414; Lift=3.41406; p-value=2.39079e-32]

Rule 000034: STROKE = 0 -> COPD = 0 [Coverage=86.04% (376); Support=83.29% (364); Confidence=96.81%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 00003a: DEATH? = 0 -> PM = 0 [Coverage=49.66% (217); Support=47.14% (206); Confidence=94.93%; Leverage=0.07141; Lift=1.17855; p-value=4.11522e-15]

Rule 000007: 6MWT,DBP1 > 90 -> 6MWT,DBP2 > 80 [Coverage=13.50% (59); Support=12.81% (56); Confidence=94.92%; Leverage=0.0954; Lift=3.91302; p-value=4.3716e-37]

Rule 000035: COPD = 0 -> STROKE = 0 [Coverage=88.33% (386); Support=83.29% (364); Confidence=94.30%; Leverage=0.07295; Lift=1.09599; p-value=2.20288e-29]

Rule 000036: MI = 1 -> STATIN = 1 [Coverage=51.94% (227); Support=46.91% (205); Confidence=90.31%; Leverage=0.07209; Lift=1.18158; p-value=4.77768e-13]

Rule 000000: 6MWT,SBP1 > 151 -> 6MWT,SBP2 > 130 [Coverage=16.71% (73); Support=14.64% (64); Confidence=87.67%; Leverage=0.10708; Lift=3.71964; p-value=2.29825e-39]

Rule 00000d: ANTIPLAT = 1 -> AF = 0 [Coverage=62.70% (274); Support=54.23% (237); Confidence=86.50%; Leverage=0.09325; Lift=1.20763; p-value=7.76026e-19]

Rule 000055: MI = 1 -> DIGOX = 0 [Coverage=51.94% (227); Support=44.39% (194); Confidence=85.46%; Leverage=0.06713; Lift=1.17814; p-value=1.97638e-10]

Rule 000028: ANTIPLAT = 1 -> DIGOX = 0 [Coverage=62.70% (274); Support=53.32% (233); Confidence=85.04%; Leverage=0.07835; Lift=1.17227; p-value=6.13762e-14]

Top 10 by Leverage:

Rule 000000: 6MWT,SBP1 > 151 -> 6MWT,SBP2 > 130 [Coverage=16.71% (73); Support=14.64% (64); Confidence=87.67%; Leverage=0.10708; Lift=3.71964; p-value=2.29825e-39]

Rule 000001: 6MWT,SBP2 > 130 -> 6MWT,SBP1 > 151 [Coverage=23.57% (103); Support=14.64% (64); Confidence=62.14%; Leverage=0.10708; Lift=3.71964; p-value=2.29825e-39]

Rule 000002: 6MWT,SBP1 <= 111 -> 6MWT,SBP2 <= 105 [Coverage=20.59% (90); Support=14.42% (63); Confidence=70.00%; Leverage=0.10505; Lift=3.68554; p-value=1.20415e-36]

Rule 000003: 6MWT,SBP2 <= 105 -> 6MWT,SBP1 <= 111 [Coverage=18.99% (83); Support=14.42% (63); Confidence=75.90%; Leverage=0.10505; Lift=3.68554; p-value=1.20415e-36]

Rules 000004, 000005: 75 < 6MWT,DBP2 <= 80 <-> 75 < 6MWT,DBP1 <= 80 [Coverage=30.89% (135); Support=19.45% (85); Confidence=62.96%; Leverage=0.09907; Lift=2.03813; p-value=1.78362e-21]

Rule 000006: 6MWT,DBP2 > 80 -> 6MWT,DBP1 > 90 [Coverage=24.26% (106); Support=12.81% (56); Confidence=52.83%; Leverage=0.0954; Lift=3.91302; p-value=4.3716e-37]

Rule 000007: 6MWT,DBP1 > 90 -> 6MWT,DBP2 > 80 [Coverage=13.50% (59); Support=12.81% (56); Confidence=94.92%; Leverage=0.0954; Lift=3.91302; p-value=4.3716e-37]

Rule 000008: 6MWT,HR2 > 80 -> 6MWT,HR1 > 98 [Coverage=24.94% (109); Support=14.42% (63); Confidence=57.80%; Leverage=0.09508; Lift=2.93695; p-value=2.84364e-27]

Rule 000009: 6MWT,HR1 > 98 -> 6MWT,HR2 > 80 [Coverage=19.68% (86); Support=14.42% (63); Confidence=73.26%; Leverage=0.09508; Lift=2.93695; p-value=2.84364e-27]

Rule 00000a: ANTIPLAT = 0 -> AF = 1 [Coverage=37.30% (163); Support=19.91% (87); Confidence=53.37%; Leverage=0.09325; Lift=1.88101; p-value=7.76026e-19]

Rule 00000b: AF = 1 -> ANTIPLAT = 0 [Coverage=28.38% (124); Support=19.91% (87); Confidence=70.16%; Leverage=0.09325; Lift=1.88101; p-value=7.76026e-19]

Top 10 by Lift:

Rule 000056: REST,DBP <= 65 -> 6MWT,DBP2 <= 65 [Coverage=11.90% (52); Support=8.01% (35); Confidence=67.31%; Leverage=0.06702; Lift=6.1278; p-value=3.90018e-28]

Rule 000057: 6MWT,DBP2 <= 65 -> REST,DBP <= 65 [Coverage=10.98% (48); Support=8.01% (35); Confidence=72.92%; Leverage=0.06702; Lift=6.1278; p-value=3.90018e-28]

Rule 000050: METS <= 3.29 -> CPX,TIME <= 6.98 [Coverage=11.67% (51); Support=8.47% (37); Confidence=72.55%; Leverage=0.06731; Lift=4.87753; p-value=1.2324e-24]

Rule 000051: CPX,TIME <= 6.98 -> METS <= 3.29 [Coverage=14.87% (65); Support=8.47% (37); Confidence=56.92%; Leverage=0.06731; Lift=4.87753; p-value=1.2324e-24]

Rule 000014: METS > 6.02 -> WEBER = 1 [Coverage=11.44% (50); Support=11.44% (50); Confidence=100.00%; Leverage=0.08666; Lift=4.12264; p-value=2.81532e-36]

Rule 000015: WEBER = 1 -> METS > 6.02 [Coverage=24.26% (106); Support=11.44% (50); Confidence=47.17%; Leverage=0.08666; Lift=4.12264; p-value=2.81532e-36]

Rule 000042: 6MWT,DBP1 > 90 -> 6MWT,SBP1 > 151 [Coverage=13.50% (59); Support=9.15% (40); Confidence=67.80%; Leverage=0.06898; Lift=4.05851; p-value=2.10481e-22]

Rule 000043: 6MWT,SBP1 > 151 -> 6MWT,DBP1 > 90 [Coverage=16.71% (73); Support=9.15% (40); Confidence=54.80%; Leverage=0.06898; Lift=4.05851; p-value=2.10481e-22]

Rule 000006: 6MWT,DBP2 > 80 -> 6MWT,DBP1 > 90 [Coverage=24.26% (106); Support=12.81% (56); Confidence=52.83%; Leverage=0.0954; Lift=3.91302; p-value=4.3716e-37]

Rule 000007: 6MWT,DBP1 > 90 -> 6MWT,DBP2 > 80 [Coverage=13.50% (59); Support=12.81% (56); Confidence=94.92%; Leverage=0.0954; Lift=3.91302; p-value=4.3716e-37]

Rule 000016: 6MWT,DIST <= 386 -> EXERCISE2 <= 10 [Coverage=19.68% (86); Support=11.44% (50); Confidence=58.14%; Leverage=0.08514; Lift=3.90877; p-value=5.46978e-29]

Top 10 by p-value:

Rule 000062: MI = 0 -> ANTIPLAT = 0 [Coverage=43.94% (192); Support=22.88% (100); Confidence=52.08%; Leverage=0.06495; Lift=1.39634; p-value=1.28443e-08]

Rule 000063: ANTIPLAT = 0 -> MI = 0 [Coverage=37.30% (163); Support=22.88% (100); Confidence=61.35%; Leverage=0.06495; Lift=1.39634; p-value=1.28443e-08]

Rule 00005a: MI = 0 -> DIGOX = 1 [Coverage=43.94% (192); Support=18.53% (81); Confidence=42.19%; Leverage=0.06672; Lift=1.56237; p-value=2.19775e-10]

Rule 00005b: DIGOX = 1 -> MI = 0 [Coverage=27.00% (118); Support=18.53% (81); Confidence=68.64%; Leverage=0.06672; Lift=1.56237; p-value=2.19775e-10]

Rule 000054: DIGOX = 0 -> MI = 1 [Coverage=72.54% (317); Support=44.39% (194); Confidence=61.20%; Leverage=0.06713; Lift=1.17814; p-value=1.97638e-10]

Rule 000055: MI = 1 -> DIGOX = 0 [Coverage=51.94% (227); Support=44.39% (194); Confidence=85.46%; Leverage=0.06713; Lift=1.17814; p-value=1.97638e-10]

Rule 000046: AF = 0 -> DIGOX = 0 [Coverage=71.63% (313); Support=58.81% (257); Confidence=82.11%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 000047: DIGOX = 0 -> AF = 0 [Coverage=72.54% (317); Support=58.81% (257); Confidence=81.07%; Leverage=0.06853; Lift=1.13191; p-value=4.8915e-12]

Rule 000048: 75 < 6MWT,DBP2 <= 80 -> 75 < REST,DBP <= 80 [Coverage=30.89% (135); Support=14.42% (63); Confidence=46.67%; Leverage=0.06852; Lift=1.90592; p-value=2.61641e-12]

Rule 000049: 75 < REST,DBP <= 80 -> 75 < 6MWT,DBP2 <= 80 [Coverage=24.49% (107); Support=14.42% (63); Confidence=58.88%; Leverage=0.06852; Lift=1.90592; p-value=2.61641e-12]

Rule 00003e: DIGOX = 1 -> AF = 1 [Coverage=27.00% (118); Support=14.64% (64); Confidence=54.24%; Leverage=0.06983; Lift=1.91143; p-value=1.58497e-12]

Appendix D – Cluster Summary Report

K-means Cluster (k=5) with 5 centroids

Data distribution:

Global: 100% (158 instances)

Cluster 0: 3.80% (6 instances)

Cluster 1: 41.14% (65 instances)

Cluster 2: 29.75% (47 instances)

Cluster 3: 23.42% (37 instances)

Cluster 4: 1.90% (3 instances)

Cluster metrics:

total_ss (Total sum of squares): 5.475960

within_ss (Total within-cluster sum of the sum of squares): 3.937280

between_ss (Between sum of squares): 1.538680

ratio_ss (Ratio of sum of squares): 0.280990

Centroids:

Global: DEATH?: "1", OQLsub1: 33.57595, OQLsub2: 5.46835, AGE: 59.12829, HEIGHT,CM: 173.98101, BMI: 27.49557, LVEF,0: "25", PM: "0", MI: "1", AF: "0", DM: "0", HT: "0", COPD: "0", STROKE: "0", KIDNEY,DIS: "0", ACEI,ARB: "1", BB: "1", MRA: "0", DIUR: "1", ANTIPLAT: "1", STATIN: "1", DIGOX: "0", HB: 14.22342, NA: 141.20253, K: 1.16987, BNP: 3096.26266, CRP: 5.23544, LVEDD: 70.16456, MR: 1.91456, REST,SBP: 118.86076, REST,DBP: 78.13291, REST,HR: 72.53797, EXERCISE1: 5.96899, EXERCISE2: 13.20886, EXERCISE3: 13.90506, 6MWT,DIST: 460.65823, 6MWT,FATIGUE: 6.46203, 6MWT,DYSPN: 3.20886, 6MWT,SBP1: 135.06329, 6MWT,DBP1: 82.68987, 6MWT,HR1: 85.93671, 6MWT,SBP2: 126.07595, 6MWT,DBP2: 81.39241, 6MWT,HR2: 76.89873, EXERCISE4: -1.73418, EXERCISE5: -9.17089, CPX,TIME: 9.65769, RER: 1.12734, SLOPE: 40.09628, METS: 4.41139, WEBER: "3", NYHA_UPDATED: "2,5"

Cluster 0: DEATH?: "1", OQLsub1: 53.41361, OQLsub2: 15.34031, AGE: 58.8789, HEIGHT,CM: 172.03141, BMI: 25.90995, LVEF,0: "15", PM: "0", MI: "0", AF: "1", DM: "1", HT: "0", COPD: "0", STROKE: "0", KIDNEY,DIS: "1", ACEI,ARB: "1", BB: "1", MRA: "1", DIUR: "1", ANTIPLAT: "0", STATIN: "0", DIGOX: "0", HB: 12.43351, NA: 139.32984, K: 1.43639, BNP: 5137.2733, CRP: 26.45288, LVEDD: 74.65969, MR: 2.50262, REST,SBP: 110.8377, REST,DBP: 67.53927, REST,HR: 72.67016, EXERCISE1: 7.55707, EXERCISE2: 10.1466, EXERCISE3: 11.1466, 6MWT,DIST: 359.42932, 6MWT,FATIGUE: 7.50262, 6MWT,DYSPN: 5.82723, 6MWT,SBP1: 116.75393, 6MWT,DBP1: 70.05236, 6MWT,HR1: 83.97906, 6MWT,SBP2: 117.51309, 6MWT,DBP2: 73.29843, 6MWT,HR2: 77.98953, EXERCISE4: 5.00524, EXERCISE5: -21.27749, CPX,TIME: 7.17207, RER: 1.04969, SLOPE: 51.58264, METS: 3.1527, WEBER: "4", NYHA_UPDATED: "3"

Cluster 1: DEATH?: "1", OQLsub1: 32.30014, OQLsub2: 4.9252, AGE: 57.55059, HEIGHT,CM: 174.25149, BMI: 27.70454, LVEF,0: "30", PM: "0", MI: "1", AF: "0", DM: "0", HT: "1", COPD: "0", STROKE: "0", KIDNEY,DIS: "0", ACEI,ARB: "1", BB: "1", MRA: "0", DIUR: "1", ANTIPLAT: "1", STATIN: "1", DIGOX: "0", HB: 14.42717, NA: 141.59431, K: 1.14857, BNP: 2268.31918, CRP: 4.84821, LVEDD: 70.00229, MR: 1.89146, REST,SBP: 120.53924, REST,DBP: 78.90776, REST,HR: 72.38917, EXERCISE1: 5.41023, EXERCISE2: 14.92749, EXERCISE3: 15.35567, 6MWT,DIST: 497.2313, 6MWT,FATIGUE: 6.38504, 6MWT,DYSPN: 2.94814, 6MWT,SBP1: 138.67829, 6MWT,DBP1: 84.33226, 6MWT,HR1: 87.26388, 6MWT,SBP2: 129.16017, 6MWT,DBP2: 83.21478, 6MWT,HR2: 75.97797, EXERCISE4: 6.32308, EXERCISE5: -4.84626, CPX,TIME: 10.31916, RER: 1.12638, SLOPE: 38.64422, METS: 4.57425, WEBER: "3", NYHA_UPDATED: "2,5"

Cluster 2: DEATH?: "0", OQLsub1: 19.11716, OQLsub2: 2.02481, AGE: 56.68581, HEIGHT,CM: 174.88422, BMI: 27.32364, LVEF,0: "40", PM: "0", MI: "1", AF: "0", DM: "0", HT: "0", COPD: "0", STROKE: "0", KIDNEY,DIS: "0", ACEI,ARB: "1", BB: "1", MRA: "0", DIUR: "1", ANTIPLAT: "1", STATIN: "1", DIGOX: "0", HB: 14.68739, NA: 141.74983, K: 1.07238, BNP: 1696.16533, CRP: 3.01678, LVEDD: 68.82426, MR: 1.66954, REST,SBP: 120.67195, REST,DBP: 80.46175, REST,HR: 73.13301, EXERCISE1: 5.37698, EXERCISE2: 14.25982, EXERCISE3: 15.0765, 6MWT,DIST: 524.46589, 6MWT,FATIGUE: 5.95796, 6MWT,DYSPN: 0.98484, 6MWT,SBP1: 139.82426, 6MWT,DBP1: 85.07236, 6MWT,HR1: 88.35975, 6MWT,SBP2: 128.44935, 6MWT,DBP2: 82.36044, 6MWT,HR2: 79.05169, EXERCISE4: -8.14817, EXERCISE5: -11.58718, CPX,TIME: 11.99434, RER: 1.13657, SLOPE: 33.79929, METS: 5.29946, WEBER: "1", NYHA_UPDATED: "2,5"

Cluster 3: DEATH?: "1", OQLsub1: 50.73661, OQLsub2: 8.74451, AGE: 64.69545, HEIGHT,CM: 172.63828, BMI: 27.83881, LVEF,0: "25", PM: "0", MI: "1", AF: "0", DM: "0", HT: "0", COPD: "0", STROKE: "0", KIDNEY,DIS: "0", ACEI,ARB: "1", BB: "1", MRA: "0", DIUR: "1", ANTIPLAT: "0", STATIN: "1", DIGOX: "0", HB: 13.64021, NA: 140.11501, K: 1.21399, BNP: 4151.16541, CRP: 5.58425, LVEDD: 70.82441, MR: 2.13696, REST,SBP: 114.59175, REST,DBP: 75.58385, REST,HR: 72.25637, EXERCISE1: 7.3827, EXERCISE2: 9.50658, EXERCISE3: 10.5417, 6MWT,DIST: 336.0439, 6MWT,FATIGUE: 6.97542, 6MWT,DYSPN: 6.23178, 6MWT,SBP1: 125.71993, 6MWT,DBP1: 78.85865, 6MWT,HR1: 81.1396, 6MWT,SBP2: 118.95961, 6MWT,DBP2: 78.42845, 6MWT,HR2: 76.1475, EXERCISE4: -9.52327, EXERCISE5: -12.29412, CPX,TIME: 6.33245, RER: 1.13163, SLOPE: 46.02294, METS: 3.33305, WEBER: "3", NYHA_UPDATED: "3"

Cluster 4: DEATH?: "1", OQLsub1: 38, OQLsub2: 11.33333, AGE: 66.3, HEIGHT,CM: 174, BMI: 24.43333, LVEF,0: "25", PM: "0", MI: "1", AF: "1", DM: "1", HT: "1", COPD: "0", STROKE: "1", KIDNEY,DIS: "1", ACEI,ARB: "1", BB: "1", MRA: "0", DIUR: "1", ANTIPLAT: "0", STATIN: "1", DIGOX: "0", HB: 13.06667, NA: 140.66667, K: 2.07333, BNP: 26474, CRP: 1.20667, LVEDD: 77.33333, MR: 2.33333, REST,SBP: 120, REST,DBP: 76.66667, REST,HR: 70, EXERCISE1: 7.66667, EXERCISE2: 8.33333, EXERCISE3: 8.66667, 6MWT,DIST: 346, 6MWT,FATIGUE: 7.66667, 6MWT,DYSPN: 1.66667, 6MWT,SBP1: 128.33333, 6MWT,DBP1: 80, 6MWT,HR1: 80, 6MWT,SBP2: 121.66667, 6MWT,DBP2: 76.66667, 6MWT,HR2: 72, EXERCISE4: -8.66667, EXERCISE5: -9.66667, CPX,TIME: 3.724, RER: 1.11333, SLOPE: 75.061, METS: 2.59048, WEBER: "4", NYHA_UPDATED: "3"

Distance distribution:

Global:

Minimum: 0.1036

Mean: 0.18013

Median: 0.17098

Maximum: 0.40583

Standard deviation: 0.04717

Sum: 28.4605

Sum squares: 5.47596

Variance: 0.00223

Cluster 0:

Minimum: 0.13857

Mean: 0.17226

Median: 0.17414

Maximum: 0.20512

```
    Standard deviation: 0.02183
    Sum: 1.03354
    Sum squares: 0.18042
    Variance: 0.00048
  Cluster 1:
    Minimum: 0.11257
    Mean: 0.15619
    Median: 0.15576
    Maximum: 0.24261
    Standard deviation: 0.02592
    Sum: 10.15215
    Sum squares: 1.62862
    Variance: 0.00067
  Cluster 2:
    Minimum: 0.10166
    Mean: 0.14455
    Median: 0.13918
    Maximum: 0.23115
    Standard deviation: 0.02605
    Sum: 6.79369
    Sum squares: 1.01322
    Variance: 0.00068
  Cluster 3:
    Minimum: 0.12708
    Mean: 0.16483
    Median: 0.16573
    Maximum: 0.20844
    Standard deviation: 0.02438
    Sum: 6.09857
    Sum squares: 1.02659
    Variance: 0.00059
  Cluster 4:
    Minimum: 0.13858
    Mean: 0.17022
    Median: 0.185
    Maximum: 0.18709
    Standard deviation: 0.02742
    Sum: 0.51067
    Sum squares: 0.08843
    Variance: 0.00075

Intercentroid distance:
  To centroid Cluster 0
    Minimum: 0.23948805147060057
    Mean: 0.3006127677709317
    Maximum: 0.37719437789573956
  To centroid Cluster 1
    Minimum: 0.12995632418903302
    Mean: 0.22903367074905218
    Maximum: 0.3533663130862467
  To centroid Cluster 2
    Minimum: 0.12995632418903302
    Mean: 0.24367726510990903
    Maximum: 0.3625485185555951
  To centroid Cluster 3
    Minimum: 0.1599635522600565
    Mean: 0.2191153877374382
    Maximum: 0.307725877780602
  To centroid Cluster 4
    Minimum: 0.307725877780602
    Mean: 0.35020877182954585
    Maximum: 0.37719437789573956
```