Introduction to Machine Learning - Final Project

Carl George-Lembach, Jelena Meyer

December 2024

Carl (50%) and Jelena (50%). We both did all exercises, compared the results and chose the visualisations and implementations that we liked best to present.

1 Business Understanding

In this report, the goal was to use a dataset based on information from the United States Behavioral Risk Factor Surveillance System to address a cardiovascular disease risk prediction problem using machine learning methods.

We began by descriptively investigating the data set to gain an overview of all 19 features and their meanings. Further, we experimented with different versions of the dataset to evaluate the influence of missing values and various preprocessing techniques on the results. We applied supervised machine learning algorithms in multiple configurations to derive comprehensive prediction rules for heart disease. Finally, we employed unsupervised algorithms to form clusters within the data and explained the patterns and their connection with heart disease.

Ultimately, the project demonstrated how machine learning can contribute to a deeper understanding of data, specifically factors that predict heart disease, offering actionable insights for early diagnosis or disease prevention.

2 Data Understanding

Our data was part of a subset derived from the 2021 annual Behavioral Risk Factor Surveillance System dataset, provided by the Center for Disease Control and Prevention. The data was collected via self-reports gathered through telephone interviews.

The original dataset was preprocessed by Lupague et al. (2023), whose preprocessing steps included data cleaning and feature selection.

The final subset we worked with comprised 19 health-related features and 308,854 rows.

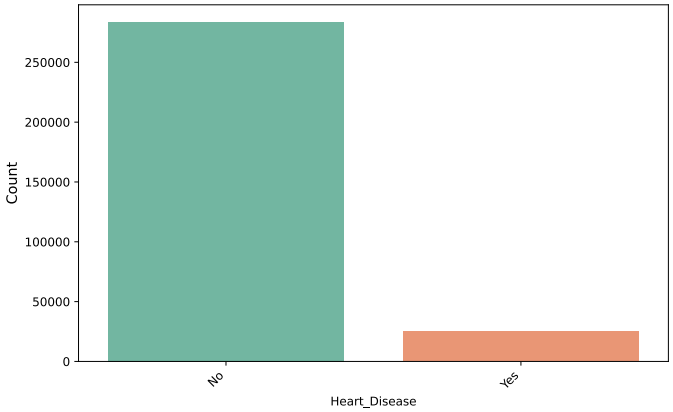
2.1 Summary of the Dataset

Of the 19 features, 12 were categorical, and 7 were numeric. As a first step, we examined each feature along with its manifestations and distribution.

Categorical variables:

- The most important variable in the data set which served as the criterion of our classifications is "Heart_Disease". Only 8% of all cases have heart disease, 92% have none (Figure 1). Although this is very good news in reality, this heavily unequally distribution likely causes considerable problems in the following analyses. In section 3 we are taking measures to deal with this issue.

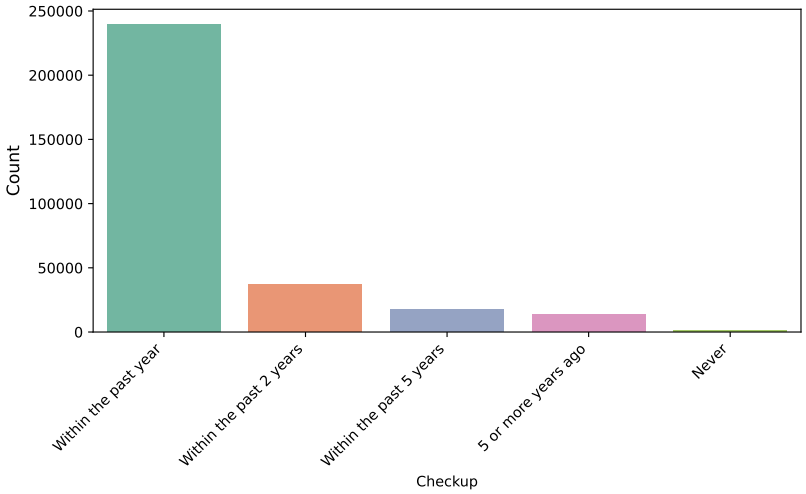
Figure 1: Distribution of the criterion Heart Disease



- "General_Health" was measured on a five point scale from "poor" to "excellent". The distribution is skewed towards better health, with the majority of respondents rating their health as "Very Good" or "Good" (35.7% and 30.9%, respectively), while fewer rated it as "Poor" (3.7%) or "Fair" (11.6%).

- "Checkup" reflects the time since respondents' last routine medical checkup. The distribution shows that the vast majority (77.5%) had a checkup within the past year, with a smaller proportion (12.0%) having one within the past 2 years. Only 5.6% and 4.3% had their last checkup within the past 5 years or 5 or more years ago, respectively, and a very small minority (0.5%) reported never having had a checkup. This highly uneven distribution (see Figure 2) and resulting low variance suggest that the variable may not be a strong predictor, as it provides limited variability to distinguish between individuals.

Figure 2: Distribution of Checkup

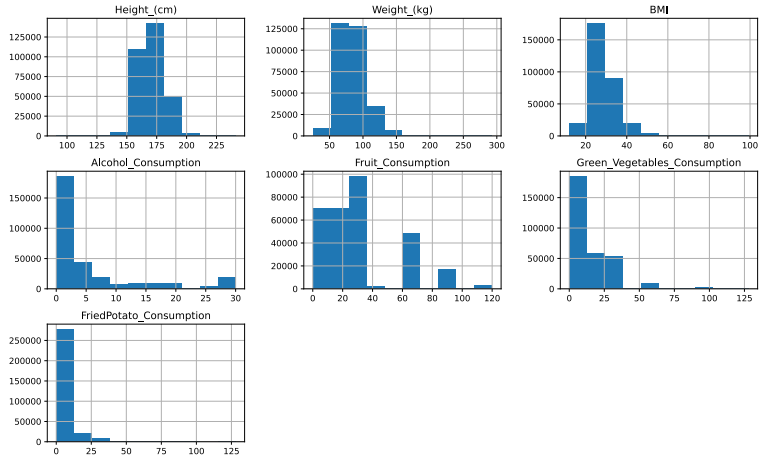


- The variable "Exercise" is a binary variable asking whether participants of the study had partaken in any physical exercise in the last 12 months. It is also rather unequally distributed with 77.5% indicating that they had exercised.
- "Skin_Cancer", also binary, indicates whether respondents have been diagnosed with skin cancer. The distribution shows that the vast majority (90.3%) have not had skin cancer, while a small minority (9.7%) have been diagnosed with it.
- The next binary feature "Other_Cancer" indicates whether respondents have been diagnosed with a type of cancer other than skin cancer. The percentages are almost identical to the "Skin_Cancer" column.
- "Depression" also binary, measures whether respondents have been diagnosed with a depressive disorder. Approximately 20% report having depression.
- "Diabetes" captures respondents' diabetes status. Most respondents (83.9%) report no diabetes, while 13.0% have diabetes. A smaller fraction report pre-diabetes (2.2%) or pregnancy-related diabetes (0.9%) (which we will summarize to "no" and "yes" respectively).
- The binary variable "Arthritis" measures whether respondents have been diagnosed with arthritis. Here the distribution is not as uneven as in the other binary variables, 32.7% report having arthritis.
- "Sex" was also operationalized in a binary way. The sample is balanced, with 51.9% female and 48.1% male.

- "Smoking_History" indicates whether respondents have smoked in their lives. With "yes" in 40.6% and "no" in 59.4% of the cases the distribution is rather balanced.
- "Age_Category" represents the age groups of respondents. The dataset is evenly distributed across age categories, with the largest groups being 65-69 (10.8%), 60-64 (10.5%), and 70-74 (10.1%). The youngest group (18-24) comprises 6.0% of the sample.

Continuous variables:

Figure 3: Distributions of all continuous features



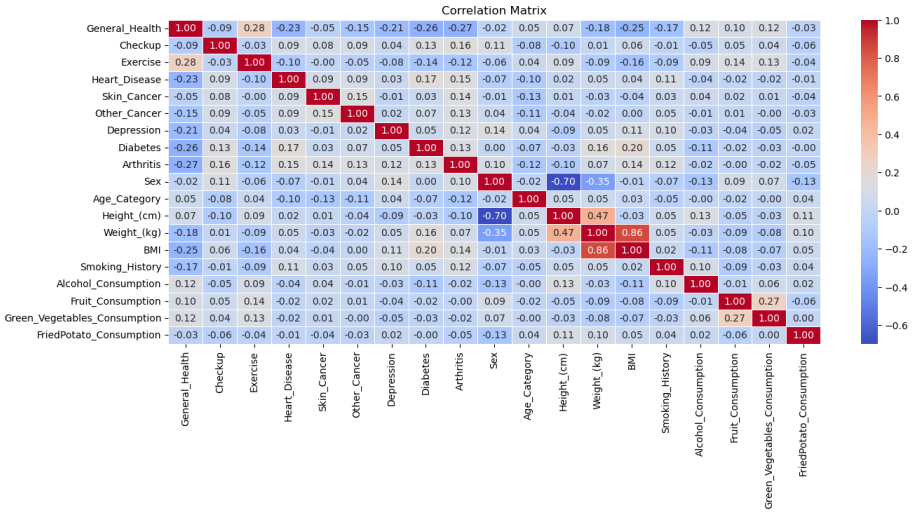
- As is typical for those metrics, "Height_cm", "Weight_cm" and "BMI" are approximately normally distributed.
- Over all cases "Height_cm" has a mean of 171 cm and a standard deviation of 11 cm.
- For "Weight_(kg)" the mean is 84 kg and the standard deviation is 21 kg.
- "BMI" has a mean of 29 and a standard deviation of 7. The official interpretation of 29 as a BMI is "overweight".
- For the last 4 continuous variables we encountered a problem: In the original codebook, several questions are being asked about alcohol consumption, fruit, vegetable and fried potatoes consumption. These questions are continuous and asked for variable time frames (day, week, 30 days) or amounts. However, the study that preprocessed the data (Lupague et al., 2023) must have changed and condensed those variables somehow. They

do not provide a reproducible script or some way to understand their preprocessing steps. Therefore, we cannot interpret the variables reliably. In our analyses we will produce multiple versions of our data set, we will either exclude these variables from the analysis or only interpret its' values in relation to the overall data set (E.g. a person can belong in the first percentile of alcohol consumers).

2.2 Correlations in the data

After having investigated every feature in the data set, we converted all variables to numeric. We did this for two reasons. First, some algorithms (like MLP or K-NN) only work with numeric variables and second, it allowed us to get an overview by checking correlations and regressions between the variables.

Figure 4: Correlations between features of the dataset



As can be seen in Figure 4, the different features do not correlate too highly with each other, with some exceptions:

- Sex and height as well as sex and weight naturally correlate quite highly.
- Weight, height and the BMI of a person have a really high correlation. This is logical, since in the formula that determines the BMI, weight and height are both used ($BMI = kg/m^2$).
- Exercise correlates with general health (probably they even are causes of the other, could work in both directions). Specific health measures like diabetes or arthritis correlate negatively with general health.

- Fruit and vegetable consumption correlate positively (probably both influenced by a third variable "healthy diet").
- Overall, all of the correlations between features are in line with common sense. We will keep the correlations and their possible consequences in mind for our analyses.

From looking at those correlations we can already interpret that general health (-0.23), diabetes (0.17), arthritis (0.15), smoking history (0.11) exercise and age category (both -0.10) have the highest correlations with heart disease. Those features are hence most likely to predict heart disease well.

2.3 Regression and first investigations of feature importance to predict heart disease

To better understand the relationships between the features in our dataset and the criterion variable heart disease, we conducted a logistic regression analysis. In this analysis, heart disease served as the dependent variable, while the remaining 18 features were treated as independent variables.

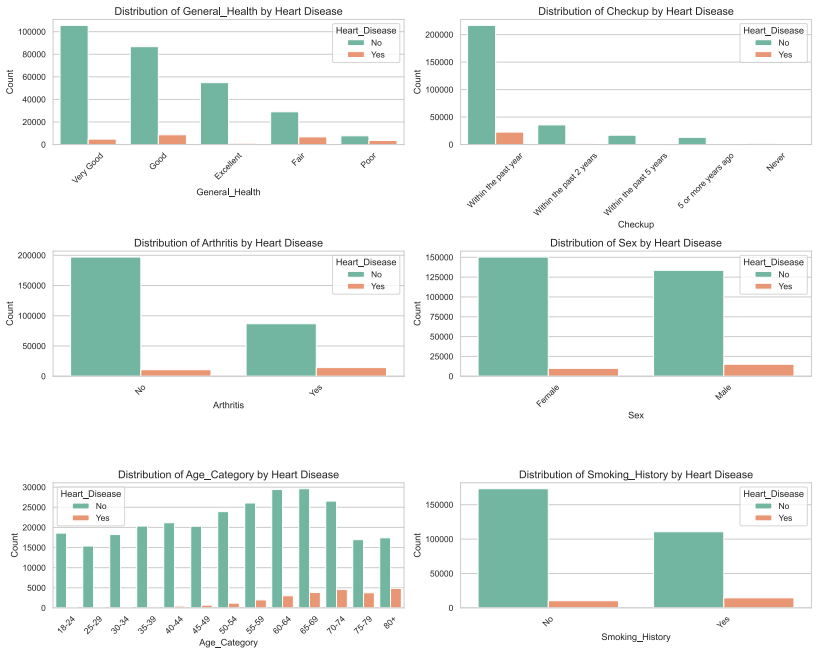
We employed a logistic regression model using the statsmodels library. An intercept term was added to the data to account for the baseline probability of heart disease. The model was fitted to the data, and the regression results, including coefficients, standard errors, and p-values, were examined to assess the significance and strength of the relationships between the features and heart disease.

The logistic regression analysis revealed valuable insights about the predictors of heart disease. Significant features include general health, smoking history, and age, which strongly correlate with a higher likelihood of heart disease. Among these, general health (which also correlates the most strongly with heart disease, see Figure 4) was the most critical predictor, with poorer health strongly increasing the odds of heart disease. In contrast, features like height, weight, BMI, fruit consumption and green vegetable consumption were not significant, indicating limited predictive value for heart disease in this dataset. Overall, the models predictive quality is moderate. Even though the accuracy of predictions was 92% the F1-score was only 11% for people with heart disease (precision: 52%, recall only 6%).

In addition, we grouped the data according to the heart disease variable and then compared the distributions of the characteristics for both resulting subgroups.

The features most prominently differing between the subpopulations with and without heart disease are visualized in Figure 5. Individuals with heart disease report lower values for the feature general health. Additionally, all individuals with heart disease had their last check-up within the past year. Among this group, the prevalence of arthritis, smoking history, and male sex exceeds 50%. Furthermore, the age distribution of individuals with heart disease skews older compared to those without the condition.

Figure 5: Feature distributions grouped by heart disease



2.4 Section Summary

From all these first exploratory analyses we conclude that some attributes (general health, sex, age category, smoking history, ...) seem to be important to predict heart disease. They correlate highly with heart disease, play important roles in its regression and have noticeably different distributions for the subpopulations with and without heart disease. We will check the results of the upcoming analyses and see whether they align with these first hypotheses.

As a side note it is noticeable that many features are not normally distributed, instead they are heavily skewed (e.g. Checkup, all cancer types, ...), leaving it questionable if they have a great predictive power.

Most notably, the target feature heart disease is also really skewed, 92% of all cases do not have a heart disease.

3 Data Preparation

We experimented with different data preparation techniques, created multiple different data sets, saved them and conducted the data analyses with each. Due to the large volume, we only report the most important and meaningful results.

- For the performance of supervised algorithms: Balanced data and missing data

- For the performance of the unsupervised algorithms: Reduced data and discretized data

3.1 Normalizing/Standardizing the data

For some algorithms to work properly and not over emphasize variables which have been measured on large scales (having big absolute differences) the data has to be standardized. We standardized the data before we used PCA since it is sensitive to scales and we wanted to make sure that all variances were equal (= 1). We also standardized the data for the Multi-layer perceptron, k-NN, k-Means and DBScan. Since the last three operate on calculating distances between data points, different scales would skew those calculations severely otherwise. The decision trees algorithm defines cut-off points within variables and chooses the ones with the highest entropy reduction. Therefore, it operates independent of scales and not scaling the data allows for an easier interpretation.

We also ran the unsupervised algorithms (k-Means and DBScan) with normalized instead of standardized data to see whether this would reveal interesting and well interpretable results. Interpretations can be found in Section 4 and 5.

3.2 Discretizing the data

Even though many algorithms require numeric inputs, we recognize that for practitioners who apply machine learning techniques and seek actionable interpretations, discretized variables can offer a more intuitive and straightforward understanding of the results.

To this end, we will apply discretization to a version of our dataset and examine the differences in results while searching for easily interpretable rules.

- **Age Category**: Three age categories:
  - "young" = 1 (up to 40 years)
  - "medium" = 2 (41 to 60 years)
  - "old" = 3 (61 years and above)
- **General Health**: Three categories:
  - "High" = 3 ("very good" and "excellent")
  - "Medium" = 2 ("good")
  - "Poor" = 1 ("poor" and "fair")
- **Checkup**: Checkup within the last year:
  - "Yes" = 1
  - "No" = 0
- **Alcohol Consumption**: Discretized into three categories:

  - "None" (33%) = 0
  - "A little" (66% up to 4 drinks) = 1
  - "A lot" (last 33%) = 2

- **BMI**: Categorized according to official standards:
  - "Underweight" = 0 (up to 18.5)
  - "Normal" = 1 (18.5 to 24.9)
  - "Overweight" = 2 (25 to 29.9)
  - "Obese" = 3 (30 and above)

- **Height**: Divided into three categories:
  - "Small" (33%, up to 165 cm) = 0
  - "Medium" (66%, up to 175 cm) = 1
  - "Tall" (last 33%, above 175 cm) = 2

- **Weight**: Categorized as follows:
  - "Light" (33%, up to 73 kg) = 0
  - "Medium" (66%, up to 90 kg) = 1
  - "Heavy" (last 33%, above 90 kg) = 2

- **Fruit Consumption**: Discretized into three levels:
  - "Little" (33%, up to 15 servings) = 0
  - "Medium" (66%, up to 30 servings) = 1
  - "Much" (last 33%, above 30 servings) = 2

- **Green Vegetables Consumption**: Three categories:
  - "Little" (33%, up to 8 servings) = 0
  - "Medium" (66%, up to 16 servings) = 1
  - "Much" (last 33%, above 16 servings) = 2

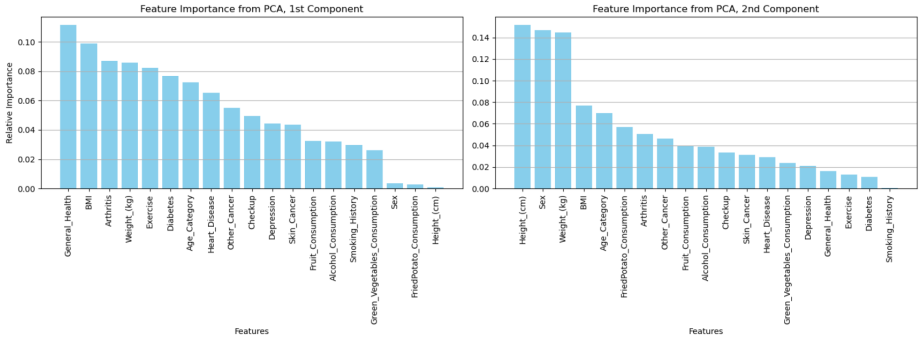
- **Fried Potato Consumption**: Divided into three levels:
  - "Little" (33%, up to 2 servings) = 0
  - "Medium" (66%, up to 6 servings) = 1
  - "Much" (last 33%, above 6 servings) = 2

3.3 Reducing the data

3.3.1 Feature extraction: Principal Component Analysis

We employed Principal Component Analysis (PCA) to retrieve information about the variance in the dataset. The first two extracted components explain 13.52% and 12.48% of the overall variance in the data. Those are not particularly high percentages, suggesting that by reducing the data set we would be likely to loose a lot of information. We investigated how much each existing feature loads on the two most important principal components. The results suggest that most of the variance in the dataset is captured by the variables general health, weight, height, BMI, sex, and arthritis. As mentioned in section 2, the variables correlate highly with each other. The multicollinearity is further highlighted by the similar loadings on the principal components. These features seem to contribute most strongly to the overall variance of the data while heart disease only adds a medium contribution to the first factor and a minor contribution to the second.

Figure 6: Results of Principal Components Analysis



3.3.2 Feature selection

We also created a modified dataset by manually reducing the number of features to facilitate more straightforward analyses. Our objective was to strike an optimal balance between dimensionality reduction and retaining sufficient information to support meaningful insights.

The rationale behind the feature selection was as follows:

In the correlation analysis presented in Section 2, we observed that height and weight were highly correlated with each other and with BMI (as well as with sex). This multicollinearity can skew analysis results, so we decided to exclude height and weight from the reduced dataset. By retaining BMI (calculated as $BMI = kg/m^2$), we ensure that the relevant information is still considered, even though the original metrics were removed.

Given that our goal is to identify comprehensive and actionable predictions and clusters, it is essential to maintain features that are interpretable. For the variables "Fruit_Consumption", "Green_Vegetables_Consumption", and "FriedPotato_Consumption", we found that their numerical representations lacked clear, comprehensible meaning. For instance, results showing "FriedPotato_Consumption" values ≥ 10 would be difficult to interpret or utilize effectively. Consequently, we decided to exclude these three features from the analysis. Similarly, the variable "Alcohol_Consumption" was removed due to its ambiguous interpretability.

Additionally, the variable "Checkup" was excluded because nearly all responses indicated that the checkup occurred "Within the Last Year", limiting its usefulness for differentiation. We also removed "Age_Category" due to its strong correlation with arthritis, which could introduce redundancy in our analyses.

Overall the modified reduced data set comprises 11 features.

Additionally, we reduced the sample size of the dataset for some analyses, sampling 2.000 to 10.000 cases randomly for the code to run faster.

3.4 Balancing the data

As visualized in Figure 1, a big problem for the coming analyses is the fact that the probability of not having heart disease in the data set is 92%.

To check if some results (like high accuracies of predicting heart disease) can simply be explained by the skewed distribution, we also built a balanced data set where the probability for heart disease is 50%. The data set includes all 19 features and randomly sampled 2.000 cases (1.000 with heart disease, 1.000 without).

3.5 Adding NAs and handling NAs in the data

As specified in the assignment, we randomly added 10% and 20% of NAs in every predictive feature of the data set (excluding the criterion heart disease to be able to check the true classification).

Since we inserted the missing data compeletely at random (MCAR) we do not have to be very cautious with NA handling techniques. We chose three different techniques and compared their performances:

- Complete Case Analysis (CCA): CCA or listwise deletion is the easiest missing data handling technique. Here, all cases where a single variable from the training data is missing are excluded and, all cases with missing data in the test data are not classified. When doing CCA, there is a steep slope of loss in information, resulting in one having almost no classified data anymore when 20-30 % of every predictor are missing.

- Mean Imputation: In this technique we used the sklearn.impute SimpleImputer to always insert the mean of a feature in every missing cell.

- Multiple Imputation: Multiple Imputation was applied via the sklearn.impute IterativeImputer. Here, for every case with a missing value a regression is calculated with the missing value as dependent and all other values of the case as independent variables, predicting the best fitting value for the missing cell. This technique is widely known as the most computationally demanding but therefore often gives the best results.

We used all different techniques and compared their performances in the classification tasks with the supervised algorithms. As expected, the CCA technique excluded more than 90% of the data, thereby disqualifying its results. Mean and multiple imputation were both successful in NA data handling.

We used the multiple imputation data in the unsupervised algorithms and compared the results with clusters that were built with complete data in section 4 and 5.

4 Modeling

4.1 Supervised Learning

To predict heart disease, the dataset was divided into 70% training data and 30% final test data. The final test data was reserved exclusively for evaluating the models' performance at the end of the process. Within the training data, a further split was performed: 70% of the training data was used to develop the models, while the remaining 30% served as pre-test data. This pre-test data was used to optimize hyperparameters and enhance the performance of each algorithm.

Three different algorithms were trained on the training data, and their performance was evaluated on the test data using accuracy, precision, recall, F1-score, and confusion matrices. For each algorithm, the key hyperparameter influencing its performance was optimized: maximum depth for decision trees, neural network architecture and learning rate for multi-layer perceptrons (MLPs), and the value of k for k-nearest neighbors (k-NN).

When applied to the original dataset, all three algorithms achieved high accuracy (92%) on both the training and test sets. However, this accuracy matched the baseline distribution of the target variable (92% "healthy"), raising concerns about its validity. Further analysis revealed that recall and F1-score were both 0, and the confusion matrices showed that all algorithms classified every instance as "healthy."

Since a simple model that always predicts "healthy" would yield the same result, these algorithms provide no meaningful insights. They fail to identify the 8% of individuals with heart disease, rendering them unhelpful for such a highly imbalanced dataset.

Hyperparameter optimization also yielded no meaningful improvements. Variations in tree depth, MLP architecture, learning rate, and kk had no effect on performance, as all algorithms defaulted to the trivial classification strategy.

To evaluate the algorithms' potential for learning meaningful classification rules, we applied them to the balanced version of the dataset:

4.1.1 Decision Trees

Decision trees are highly susceptible to overfitting, making the optimization of the "max depth" hyperparameter crucial. This parameter, available in the DecisionTreeClassifier function, defines the maximum number of nodes the decision tree can contain.

Overfitting occurs when the algorithm models the training data too closely, capturing noise rather than underlying patterns. As illustrated in Figure 7, increasing the "max depth" parameter improves the accuracy on the training data. However, this comes at the cost of reduced accuracy on the test data, as the model becomes overly specialized. To mitigate this, we optimized "max depth" to achieve the best test performance. Across all trials, the optimal depth ranged between 1 and 10 nodes.

Using the optimal depth, we constructed the best-performing decision tree. A key advantage of decision trees is their interpretability, as demonstrated in Figure 8. In all resulting models, the first split was based on either age or general health, as these features provided the greatest reduction in entropy.

Figure 7: Train and test accuracy depending on max depth of the decision tree

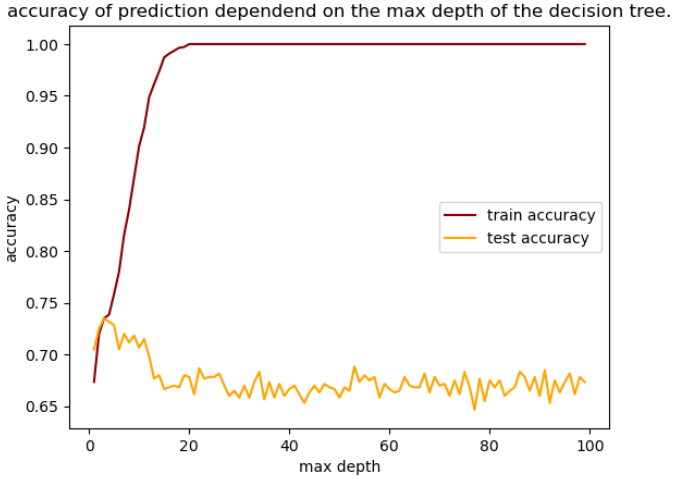
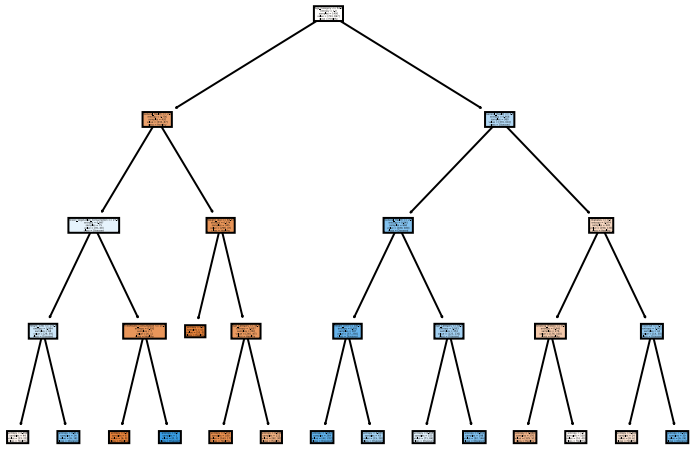


Figure 8: Decision tree with best accuracy for balanced data



4.1.2 Multi-layer Perceptrons

To optimize the performance of our multi-layer perceptron (MLP), we experimented with various learning rates and neural network architectures.

The learning rate controls the magnitude of updates to the model's weights during training in response to errors. A learning rate that is too high risks causing the model to converge prematurely to a suboptimal solution, while one that is too low may lead to stagnation during training. We evaluated the accuracy across a range of learning rates and selected the value that yielded the best performance.

The architecture of a neural network, defined by the number of hidden layers and neurons per layer, significantly affects the model's classification capabilities and predictive quality. We tested configurations with one to three hidden layers, each containing between one and 50 neurons. The architecture that produced the highest accuracy was chosen as the optimal design.

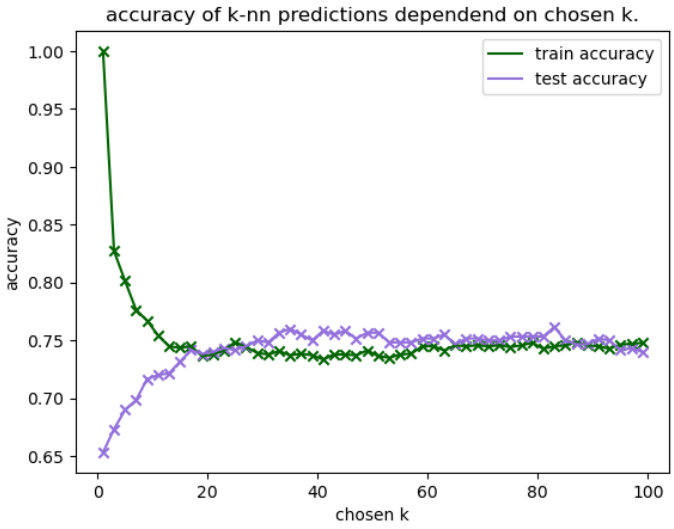
4.1.3 K-Nearest Neighbours

The k-nearest neighbors (k-NN) algorithm is a non-parametric, instance-based learning method that classifies a new data point by comparing it to the k closest points in the training set and assigning it to the majority category among those neighbours.

The performance of k-NN is highly sensitive to the choice of k. A small k can make the model overly sensitive to noise and outliers, leading to high variance and potential over-fitting. Conversely, a large k incorporates too many neighbours, including points from different classes, resulting in high bias. This smoothing effect reduces the model's ability to capture subtle class distinctions and can cause under-fitting.

To address this, we optimized k by selecting the value that maximized the classifier's prediction accuracy, as shown in Figure 9.

Figure 9: Accuracy depending on chosen k



4.1.4 NAs in supervised learning

We experimented with various missing data handling techniques on both the original and balanced datasets, considering scenarios with 10% and 20% missing data.

For the original dataset, the results were consistent with those obtained from the complete data. The algorithms continued to classify all cases as "healthy," resulting in unchanged performance scores. Although the models had less information due to missing data, their accuracy remained unaffected. All other metrics (precision, recall, and F1-score) remained at zero.

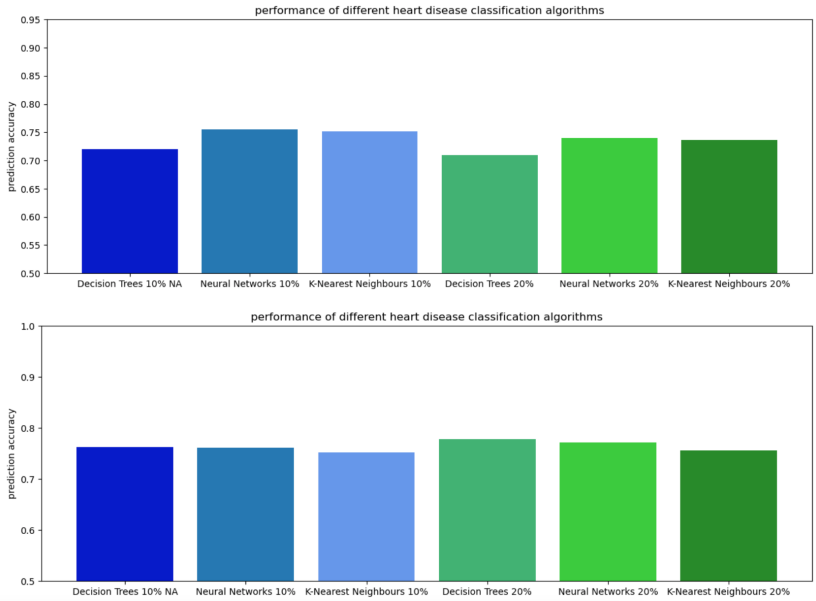
The balanced dataset provided more insightful results:

- Complete Case Analysis (CCA): While performance scores across all algorithms remained relatively high (70%–80% for accuracy, precision, recall, and F1-score), closer analysis revealed that only 2.7% of the original data was retained after dropping rows with missing values. This substantial data loss rendered the results unreliable.

- Mean Imputation: As shown in Figure 10, mean imputation enabled all algorithms to achieve strong performance scores, with precision, recall, and F1-score comparable to accuracy. Among the three algorithms, the decision tree classifier performed slightly worse for both 10% and 20% missing data.

- Multiple Imputation: Interestingly, multiple imputation led to slightly better classifier performance with 20% missing data compared to 10%. In both cases, the classifiers achieved higher accuracy on the test data than when using the original balanced dataset. This improvement may result from multiple imputation replacing missing values with plausible estimates, reducing randomness and bias introduced by missing data. By providing more consistent and complete information, the models could generalize better than with more noisy original data.

Figure 10: Mean (upper) and multiple Imputation and resulting accuracies of different algorithms



4.1.5 Reduced and discretized data in supervised learning

Discretizing or reducing the original data set did not change the supervised learning algorithms performances.

In the balanced version of the reduced data set all algorithms performed a bit worse (accuracies around 71%), probably due to the loss of information.

Discretizing the balanced data set in contrast, did not impede the algorithms performances.

4.2 Unsupervised Learning

4.2.1 K-Means

We employed the elbow method to find the optimal value for k in the implementation of k means. This yielded a value of k=6.

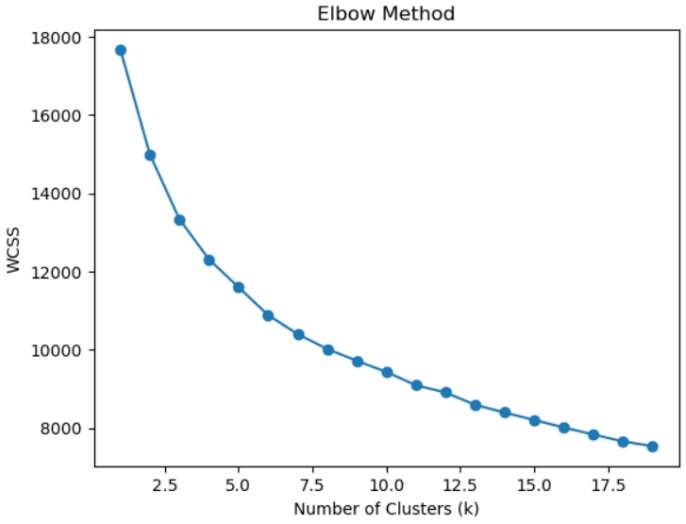


Figure 11: Results of the elbow method

Running k-means on the entire dataset gives us the following six clusters:

- Male respondents with arthritis
- Female respondents with arthritis
- Male, no arthritis, smoking history
- Male, no arthritis, no smoking history
- Female, no arthritis, smoking history
- Female, no arthritis, no smoking history

For a different k chosen, the results are similar. In general, respondents are clustered by the variables Sex, Arthritis, and Smoking_History.

To determine the probability of developing heart disease, sex appears to be the strongest predictor in k-means, as the prevalence of heart disease is by far the highest among the male group, while the female group without arthritis has a significantly weaker chance of having heart disease. These results can be visualized in a 2-dimensional space using the previous results applying PCA to the data (see Figure 12).

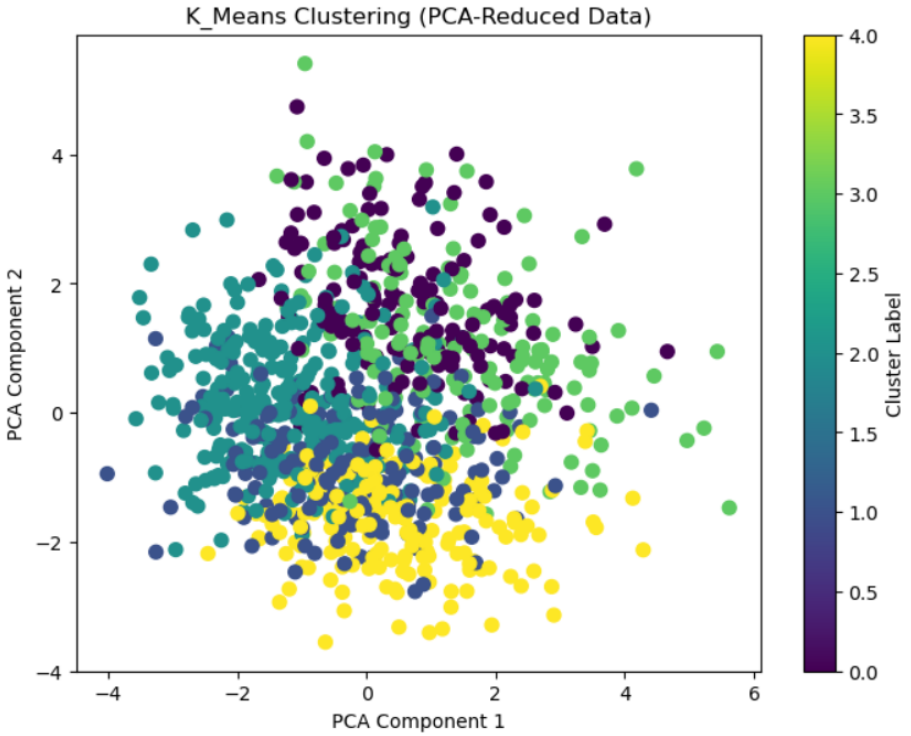


Figure 12: K-Means clusters visualized on PCA reduced data

We employed this algorithm for different versions of the data sets: The complete, unaltered data set; the balanced dataset; the discretized dataset and the dataset with a reduced amount of features. We ran k-means on the datasets with artificially inserted NA values too; however, the results in this case were even splits between respondents with heart disease and those without. This is likely because k-Means is biased towards splitting among binary variables, due to the chosen distance metric. We decided not to further investigate these results.

Table 1 and Table 2 show the cluster identified by k Means as the highest and lowest risk clusters for heart disease, respectively.

Dataset	High-risk group	Increased risk
Complete dataset	Male, Arthritis	113%
Balanced dataset[1]	Arthritis, frequent exercise	100%
Discretized dataset	Smoking history	107%
Reduced dataset	Depression	119%

Table 1: Cluster with the highest risk for Heart Disease, for $k = 6$

Dataset	Low-risk group	Decreased risk
Complete dataset	Female, no arthritis, no smoking history	-60%
Balanced dataset	Male, no arthritis	-100%
Discretized dataset	No checkup in past year	-73%
Reduced dataset	Female, no arthritis, no smoking history	-62%

Table 2: Cluster with the lowest risk for Heart Disease, for $k = 6$

4.2.2 DBSCAN-Clustering

For the parameter min_points in the DBSCAN clustering Algorithm, we use the heuristic that the chosen value should be larger than or equal to the number of features. This gives us the value 12 for the reduced data set and 20 for the complete data set. We went with lower values within this range, as many points were left unassigned.

For the parameter ε in the DBSCAN algorithm we used a k-distance plot. As this leaves some room for experimentation, we went with higher ε values to get multiple clusters of both respondents with and without heart disease, respectively.

When clustering with distance-based algorithms like DBScan and K-Means, binary variables can disproportionately influence the clustering process if all

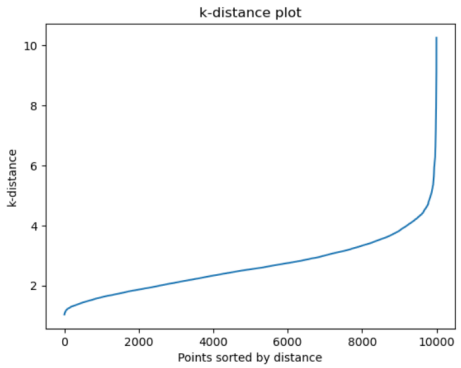


Figure 13: K-distance plot for DBSCAN

features are normalized. This occurs because binary variables (normalized to 0 or 1) have a high maximum pairwise distance, while continuous variables typically exhibit smaller normalized distances. As a result, clusters may be biased toward patterns dominated by binary features, potentially overshadowing the contribution of continuous features.
In addition, we observed that DBscan has a tendency to split the observations among the variables that are the most unevenly distributed. In this data, that variable would be Heart_Disease, as it is the binary variable with the most uneven distribution of positives and negatives among all the features. As a result, as the K-distance plot gives us a range of possible values, we set the ϵ parameter in such a way that we receive multiple clusters of respondents with heart disease, in order to investigate the differences between these clusters.

We found that the two groups centered around respondents with heart disease were separated cleanly along the variable 'Diabetes';the group without diabetes on average reported higher general health, is more likely to be female, and less likely to have arthritis or a history of smoking. Both clusters 5 and 6 consist of vastly more male respondents and respondents with a history of smoking compared to the general population. Figure 14 shows the differences between the mean values of all the clusters identified by DBSCAN as well as the mean value of all unassigned points (labeled '-1') in regards to the features General_Health, Sex, Smoking_History, and Diabetes.

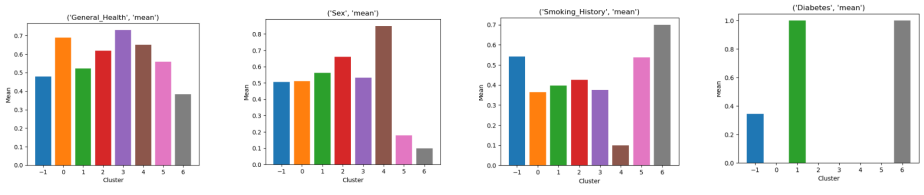


Figure 14: Comparison of DBSCAN Clusters for Various Features; $Heart\_Disease = 1$ is true for clusters 5 and 6

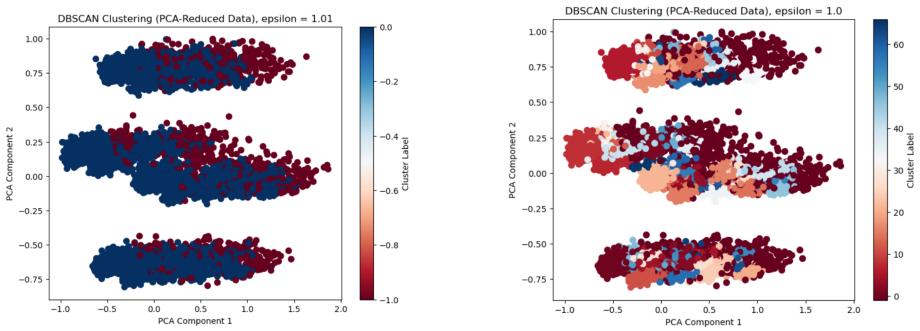
4.2.3 Data selection

We experimented with both normalization and standardization with both unsupervised learning methods. Whereas normalization leads to clearly defined boundaries between clusters, visualizing the standardized data in reduced dimensionality through PCA gives us one large cluster where no clear boundaries are visible. This is due to the binary seperation that is present when we normalize.

For the k-Means algorithm, we chose the normalized data, as this lead to more cleary definable groups being indentified (e.g. male respondent with out arthrits,

female respondents without a smoking history, etc.) In contrast, using standardization for k-Means resulted in clusters that did not give us any meaningful relationship between the clusters and their risk of heart disease.

For the DBScan algorithm, we chose the standardized data, as the normalized data was too volatile and did not give us meaningful results (see Figure 15). This is also due to the behavior of the binary features in our dataset, which standardization better addresses by reducing variance across the board.



(a) DBSCAN for eps=1.01 (b) DBSCAN for eps=1

Figure 15: Results of DBSCAN for different eps values

5 Evaluation

5.1 Supervised Algorithms

In the original dataset, all supervised algorithms achieved a precision of 92%, as they exploited the imbalanced distribution of the target variable, heart disease.

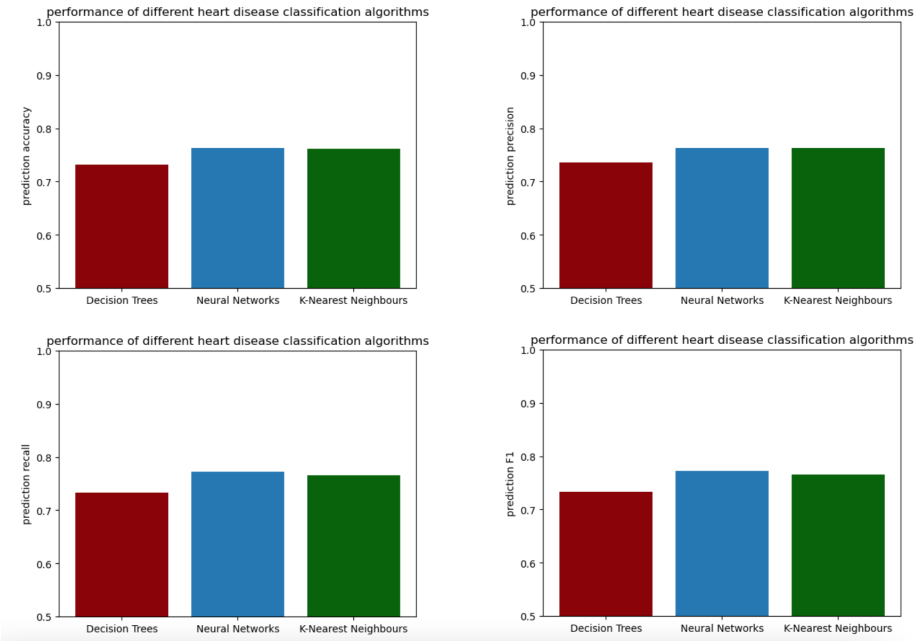
Our necessary application of the algorithms to a balanced data set resulted in the performances shown in Figure 16.

Among the three algorithms, the decision tree exhibited the lowest performance (all metrics $\approx$ 73%). However, if interpretability and simplicity of decision rules are prioritized, the slight performance loss of approximately 2% may be an acceptable trade-off for practical usability.

Both k-NN and MLP performed similarly, with all metrics ranging between 76% and 78%. The choice between these algorithms depends on the user's priorities: k-NN demonstrated better precision, while MLP marginally outperformed in all other metrics.

We also evaluated the algorithms trained on the balanced dataset using the original (unbalanced) test data. Accuracy, recall, and F1-score were consistent with the results shown in Figure 16. However, precision dropped significantly to 25% due to the small number of true positives in the unbalanced dataset.

Figure 16: Different performance metrics of supervised classifications

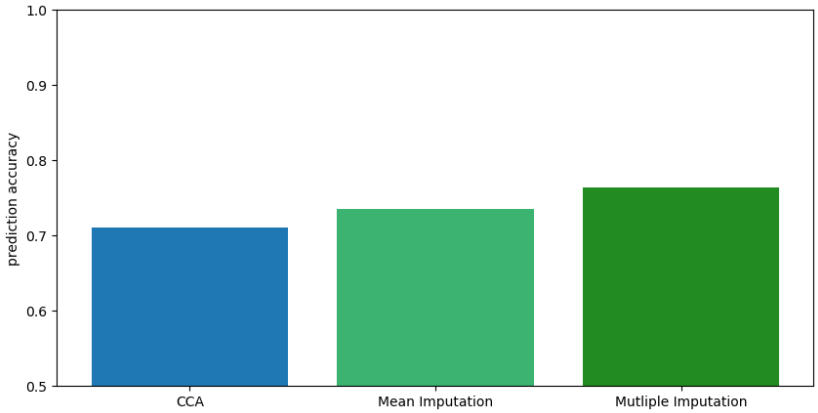


In summary, if accuracy alone were the goal, the original classifiers — predicting "healthy" for all cases — would suffice. However, these models provide no informational value. In contrast, the algorithms trained on the balanced dataset offer meaningful insights: they identify individuals with heart disease in nearly 80% of cases (recall). While the probability of a false positive remains high (75%), these cases would require further medical testing. We believe that these algorithms, trained on balanced data, are the only ones capable of providing relevant and actionable predictions for heart disease.

Our results also highlight the importance of handling missing data appropriately. For datasets with missing values that are missing completely at random (MCAR), multiple imputation emerged as the most effective method. When 20% of data is missing, complete case analysis (CCA) eliminates such a large proportion of cases (retaining only 3%) that its results become unreliable and heavily influenced by the remaining random subset. Consequently, the resulting models perform no better than chance.

Both mean imputation and multiple imputation worked well across all supervised learning algorithms tested. Decision trees experienced a slightly greater performance decline compared to MLPs and k-NNs. Interestingly, multiple imputation often improved overall algorithm performance, likely due to its ability to smooth noisy data and provide more consistent inputs for training.

Figure 17: Averaged accuracies of algorithms trained on data where NAs were handled with one of three different techniques



5.2 Unsupervised Algorithms

For the k-Means algorithm, the choice of dataset had a significant impact on the features identified as influencing the likelihood of having heart disease. Depending on the dataset, the high-risk groups included male respondents with arthritis, respondents with arthritis who exercise frequently, those with a history of smoking, and respondents with depression. The increased risk of heart disease for these groups ranged from 100% to 119%. These findings align closely with the results from the supervised learning section, with the exception of the balanced dataset. However, this discrepancy can likely be attributed to the smaller sample size of the balanced dataset.

Low-risk groups identified by k-means included female respondents without arthritis or a smoking history, male respondents with arthritis, and those who had not had a medical checkup in the past year. The decrease in heart disease risk for these groups ranged from 60% to 100%, with the latter results also coming from the balanced dataset.

For the DBSCAN algorithm, experimentation with different datasets did not produce variations in the clusters formed, so these results were not further pursued. DBSCAN identified two distinct clusters of respondents with heart disease, separated based on diabetes status. Respondents with diabetes and those without were found to exhibit markedly different characteristics.

Future analyses using these methods could explore the effects of varying parameters to better understand their influence on clustering results.

6 Deployment

All scripts which we used are part of the zip folder of this report. They are fully reproducible so that all our algorithms can now be deployed in order to cluster and predict new data cases.

The results of this project provide actionable insights for identifying individuals at risk of heart disease. The predictive models developed could be deployed in healthcare systems to assist physicians in early diagnosis or to guide interventions.

In addition, the clustering analysis could inform public health campaigns by identifying groups with shared risk factors, such as men who are smoking or have arthritis, allowing for targeted interventions.

Practical deployment would require ensuring that the models are updated regularly with new data to maintain accuracy and addressing ethical concerns, particularly with respect to privacy and the handling of sensitive health information.