02452 Machine Learning

Project 1 – South African Heart Disease Dataset Analysis

Technical University of Denmark

Group 121 – Tuesday 30th September, 2025

Section	Contributors	Contribution (%)
2. Dataset	Vladyslav Horbatenko	30%
	Juan M. Rodriguez	**40%**
	Aryan Mirzazadeh	30%
3. PCA (Principal Component Analysis)	Vladyslav Horbatenko	30%
	Juan M. Rodriguez	30%
	Aryan Mirzazadeh	**40%**
4. Discussion	Vladyslav Horbatenko	**40%**
	Juan M. Rodriguez	30%
	Aryan Mirzazadeh	30%

Name and email		Student number
Vladyslav Horbatenko	s254355@dtu.dk	s254355
Juan Manuel Rodriguez	s253505@dtu.dk	s253505
Aryan Mirzazadeh	s204489@dtu.dk	s204489

Contents

1 Introduction

Cardiovascular diseases remain one of the leading causes of mortality worldwide, and identifying their risk factors is essential for early prevention and treatment. In this project, we analyze the South African Heart Disease dataset, with a particular focus on coronary heart disease (CHD) incidence and systolic blood pressure (SBP). Our objective is to explore how lifestyle factors (such as tobacco and alcohol consumption), psychosocial indicators (Type-A behavior), and biological measures (cholesterol levels, adiposity, obesity, age, family history) contribute to predicting both the occurrence of CHD and variations in blood pressure.

2 Dataset

2.1 Dataset: Feature Overview

The dataset is a retrospective sample of adult males from a heart-disease high-risk region in the Western Cape, South Africa. Each individual was assessed for multiple clinical and lifestyle factors, with CHD status recorded as the primary outcome. There are roughly two controls per case of CHD. Many of the CHD-positive men have undergone blood pressure reduction treatments and other programs to reduce their risk factors after their CHD event; in some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al. (1983) in the *South African Medical Journal* [1]. The dataset was obtained from the *"Elements of Statistical Learning"* repository [2]. Dataset consists of 462 samples and 10 features that are described in Table 1.

Attribute	**Data type**	**Scale**	**Description**
`sbp`	Continuous	Ratio	Systolic blood pressure (mmHg).
`tobacco`	Continuous	Ratio	Lifetime tobacco consumption in kg.
`ldl`	Continuous	Ratio	Low-density lipoprotein cholesterol (mmol/L).
`adiposity`	Continuous	Ratio	Numerical index of body fat content.
`famhist`	Discrete	Nominal	Family history of coronary heart disease.
`typea`	Continuous	Interval	Type-A behavior (psycho-social stress; higher = more Type A). Derived from questionnaire.
`obesity`	Continuous	Ratio	Obesity index.
`alcohol`	Continuous	Ratio	Current alcohol consumption.
`age`	Continuous	Ratio	Age.
`chd`	Discrete	Nominal	Coronary heart disease status.

Table 1: Dataset attributes and descriptions

2.2 Dataset Issues

The dataset has no missing values or inconsistencies. However, the `alcohol` feature is not associated with explicit measurement units or a specified time-frame, which introduces uncertainty. In this analysis, it is hypothesized based on similar datasets and feature range, that it represents grams of ethanol consumed. This limitation must be acknowledged in subsequent steps.

2.3 Previous Work

Several prior studies provide valuable context for our analysis of the South African Heart Disease dataset and the broader challenge of coronary heart disease (CHD) prediction.
The first study *"Coronary Heart disease prediction using machine learning"* [3] proposes a dual-method system that integrates structured patient data with medical imaging. Specifically, a Logistic Regression model was trained on tabular patient information collected via forms, while a Convolutional Neural Network (CNN) was applied to MRI scans of the heart. The reported performance of the Logistic Regression model alone was strong, achieving 83.88% training accuracy and 85.25% test

accuracy. The system is designed to provide a more robust diagnosis by combining form-based and image-based analysis for early and more accurate CHD detection.

A second study *"Exploring Machine Learning Techniques for Coronary Heart Disease Prediction"* [4] explores a comparative evaluation of machine learning techniques applied exclusively to structured clinical data for CHD prediction. The authors implemented a feature engineering pipeline that included statistical feature selection, ANOVA (F-test) ranking, permutation feature importance, and Random Forest feature importance, alongside redundancy checks using Pearson correlation.

A range of classifiers was tested, including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) neural networks. Among these, SVM slightly outperformed the other models, with prediction accuracies of 73.8%, compared to 73.4% for MLP, 73.2% for KNN, and 72.7% for Logistic Regression. The relatively modest performance highlights the challenge of working with small datasets (≈400 samples).

To address this, the study applied K-means SMOTE, a synthetic oversampling technique, to expand the dataset to 604 samples. This resulted in an average precision improvement of approximately 11% across all models, underscoring the importance of addressing class imbalance and data scarcity in medical datasets.

2.4 Summary Statistics

In the summary statistics (Table 2), we can observe that the characteristics of obesity and adiposity have very similar scales, mean, and four quantiles. So, in the data visualization section of the report, we examine whether these features are correlated and potentially redundant in the dataset.

	sbp	**tobacco**	**ldl**	**adiposity**	**typea**	**obesity**	**alcohol**	**age**
Mean	138.33	3.6	4.7	25.4	53.1	26.0	17.0	42.8
Variance	420.10	21.1	4.3	60.5	96.4	17.8	599.3	213.4
Std	20.50	4.6	2.1	7.8	9.8	4.2	24.5	14.6
Q25	124	0.1	3.3	19.8	47	23.0	0.5	31
Median	134	2.0	4.3	26.1	53	25.8	7.5	45
Q75	148	5.5	5.8	31.2	60	28.5	23.9	55
Q100	218	31.2	15.3	42.5	78	46.6	147.2	64
Min	101	0.0	0.98	6.74	13	14.7	0.0	15
Max	218	31.2	15.33	42.49	78	46.58	147.19	64

Table 2: Summary statistics of continuous features

2.5 Data Transformation

There are primarily three techniques, which are useful for applying to our dataset.

Standardization: Regression models are sensitive to scale. Since feature ranges vary widely (e.g., `sbp` ranges from 101 to 218, while `ldl` ranges from 0.98 to 15.33), standardization ensures all features are comparable. For each feature k and observation i:

$$\tilde{x}_i^{(k)} = \frac{x_i^{(k)} - \hat{\mu}_k}{\hat{\sigma}_k},$$

leading to mean 0 and variance 1.

Log Transformation: For heavily skewed features (e.g., `tobacco`, `alcohol`), a log transform reduces skewness and makes distributions more symmetric.

Binary Encoding: Binary categorical variables such as `famhist` can be encoded as $\{0, 1\}$ for modeling. The `chd` feature is already binary and requires no encoding.

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	famhist	chd
Standardization	✓	✓	✓	✓	✓	✓	✓	✓		
Log transform		✓					✓			
Binary encoding									✓	

Table 3: Data transormation techniques applied to features

2.6 Data Visualizations

Distribution

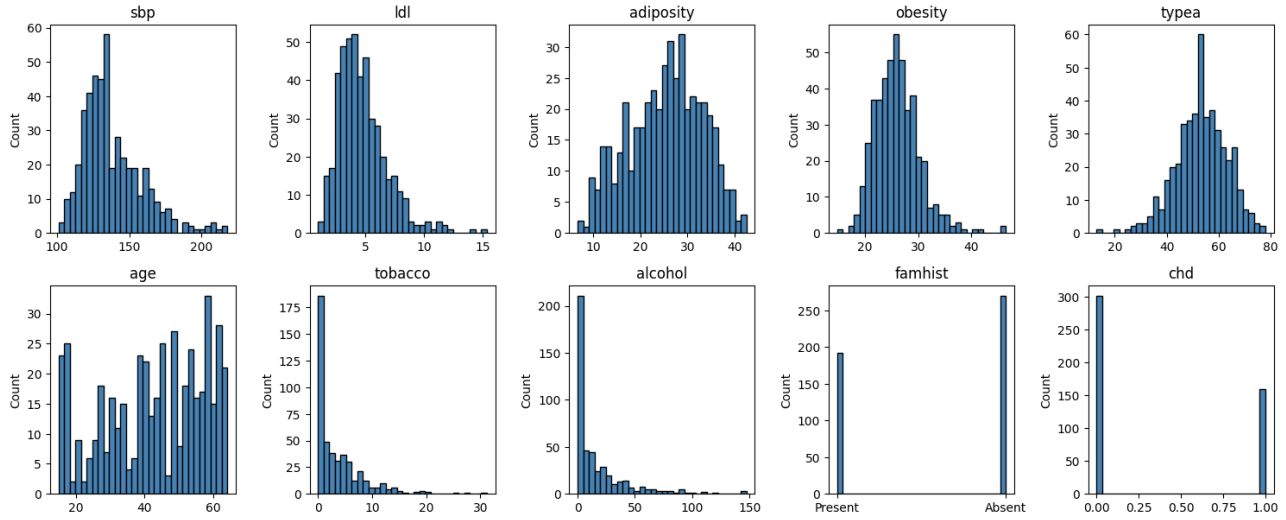

Figure 1: Distributions of Standardized Features

The subplots display the value distributions for each feature. For continuous variables, the histograms show how frequently values fall within certain ranges, while for categorical/binary variables like `famhist` and `chd`, the plots show the frequency of each category.

The features `obesity`, `ldl`, and `sbp` roughly follow a bell-shaped distribution, though they are slightly left-skewed. `Adiposity` is the closest to normal. `Age` appears almost uniformly distributed. `Alcohol` and `tobacco` are heavily skewed with many zeros. Finally, `famhist` and `chd` are binary variables, each showing only two possible outcomes.

Outliers

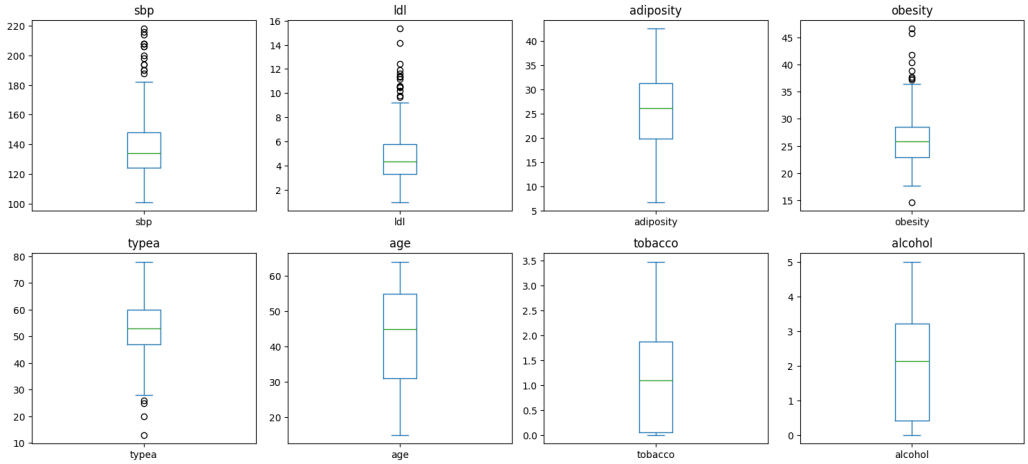

Figure 2: Boxplots of raw features before transformation, highlighting potential outliers.

Outlier analysis was performed on the raw dataset to preserve the original scales and variances of each feature. The results show clear extreme values in `sbp`, `tobacco`, `alcohol`, `ldl`, and `obesity`. By contrast, the binary variables `chd` and `famhist` do not display outliers, as expected, and their boxplots provide little additional insight.

Several patterns are evident. Systolic blood pressure (`sbp`) includes individuals with exceptionally high readings above 180–200 mmHg. Tobacco consumption is mostly concentrated near zero, but a few participants report unusually high lifetime usage (around 25–30 kg). Similarly, `ldl` cholesterol features occasional extreme values above 12–15 mmol/L, while `obesity` shows a small number of cases exceeding 35. Alcohol consumption is heavily right-skewed, with extreme values above 60–140 units. In contrast, `adiposity` is fairly symmetric with only minor high-end outliers around 40, and `typea` has a few low-end outliers below 30 but remains concentrated in the mid-range. Finally, `age` is evenly distributed between 15 and 65 years, with no pronounced extremes.

Correlation

Based on the correlation matrix (Figure 3), three notable feature combinations exhibit moderate to high correlation:

1. *Adiposity and Obesity*: These features show a strong positive correlation (0.72), which is expected given that a higher body fat index (adiposity) is a key factor in defining obesity levels.

2. *Age and Adiposity*: A correlation coefficient of 0.63 suggests that adiposity tends to increase with age.

3. *Age and Tobacco*: With a correlation coefficient of 0.53, this indicates a moderate positive relationship, potentially suggesting that older individuals may have higher levels of tobacco consumption.

These correlations provide initial insights into the relationships between the variables within the dataset.

If we observe the feature of interest, `CHD`, seems to have the highest correlation with age compared to all other features, even though it is not as highly correlated as the examples above.

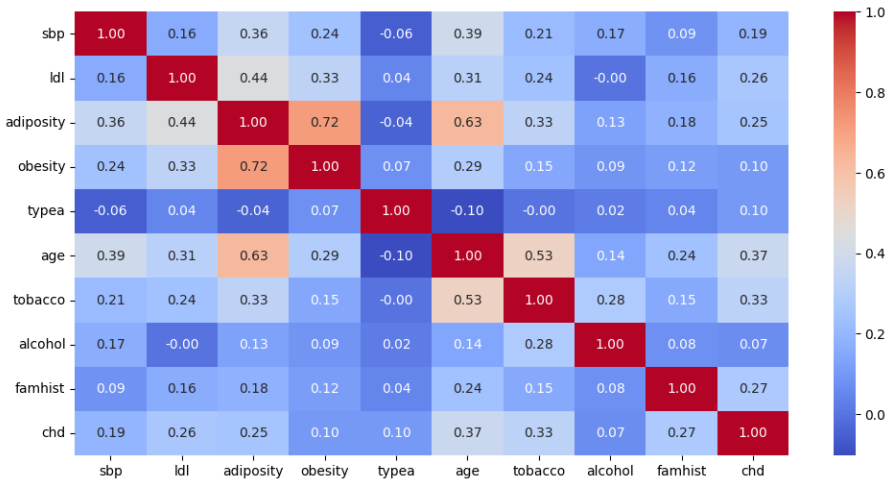

Figure 3: Correlation Matrix

3 PCA (Principal Component Analysis)

We standardized the data prior to applying PCA. The result is the following rearrangement that allows all centered values to fall between -5 and 5.

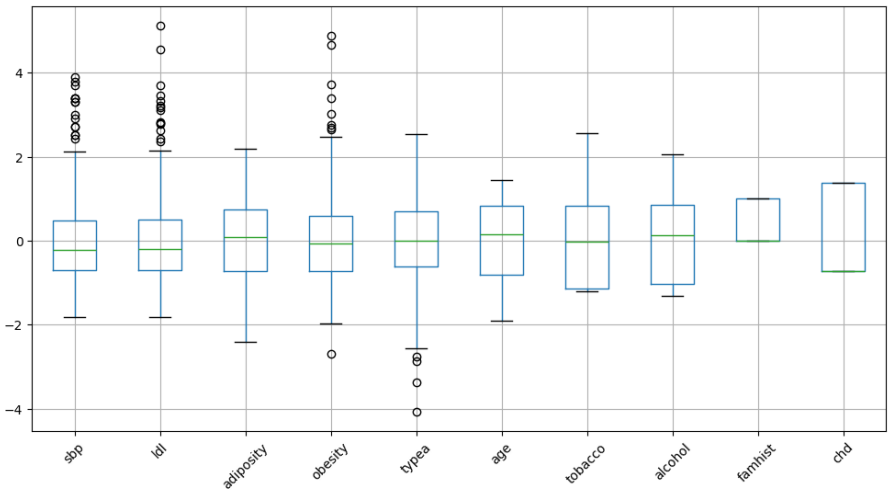

Figure 4: Boxplots of Standardized Features

Since the goal is to classify `chd`, this variable is treated as the target and, therefore, removed from the feature set before applying PCA.

Next, we compute the variance explained by each principal component to evaluate how much information they retain. By examining the cumulative explained variance, we identify the minimum number of components required to surpass the 90% threshold.

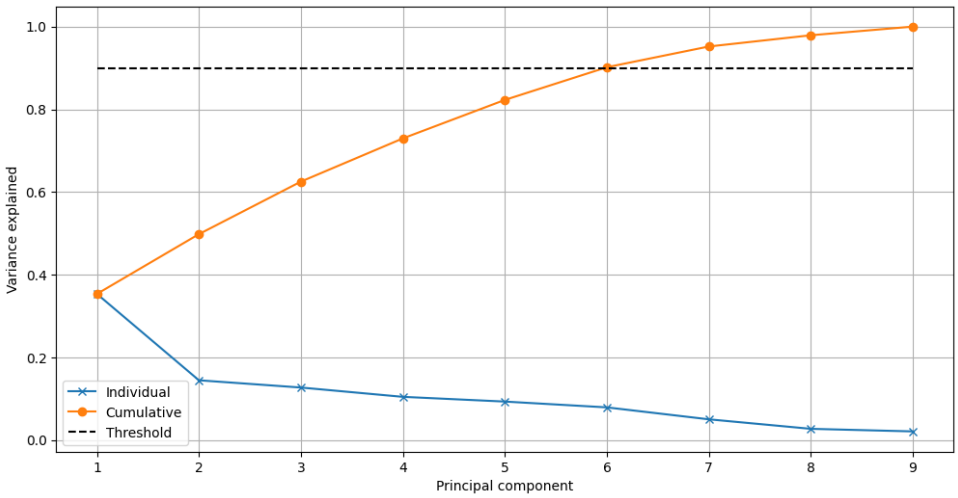

Figure 5: Variance Explained By Principal Components

Variance explained by principal components We observe that after the 7th principal component, the explained variance falls below 5% per component. To hit our 90% threshold we can exactly use the first 6 principal components.

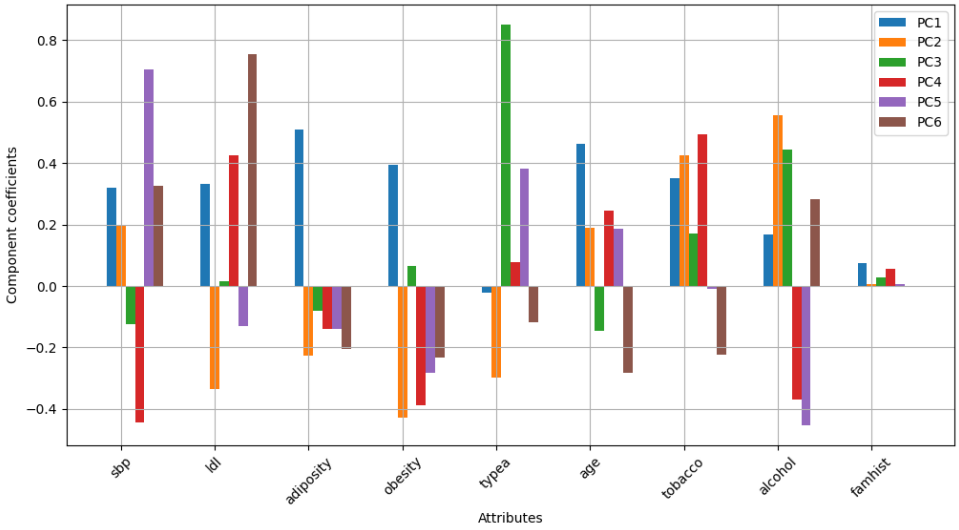

Figure 6: PCA Component Coefficient (K=6) - without CHD feature

We can observe that `adiposity` and `obesity` follow a similar trend in the PCs, aligning with the correlation matrix. The correlation between `age` and `adiposity` is also expressed by their shared direction in PCs 1, 3 and 6.

Another useful plot to observe the direction and the strength with which the features affect each principal component is the following:

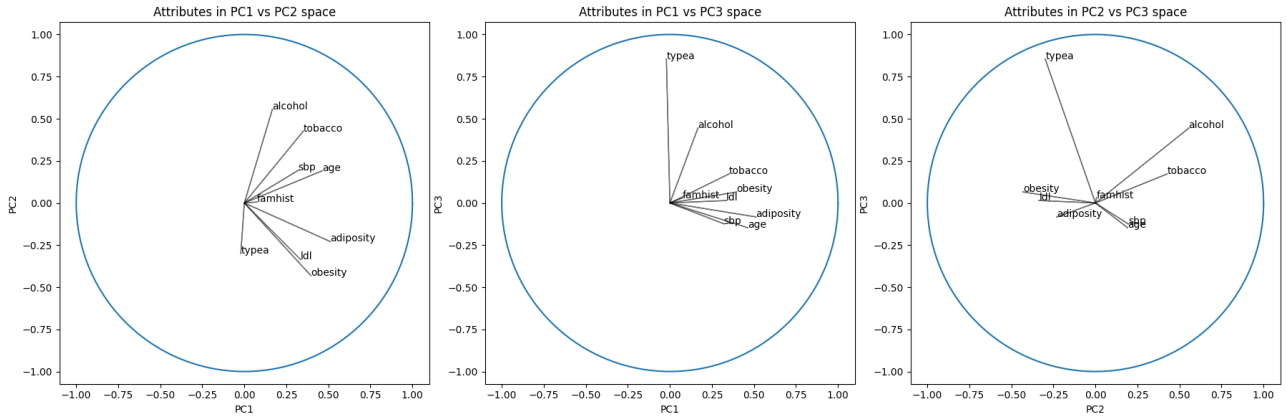

Figure 7: One to one comparison of PC paris

We observe the same trends as in the PC bar plot. In both PC1 and PC2, `obesity` and `adiposity` move in similar directions. `Age` and `adiposity` also follow similar patterns in PC1 and PC2. Meanwhile, PC3 mainly reflects positive contributions from `alcohol` and `typea`.

The 2D scatter plots are presented for the combinations PC1/PC2, PC1/PC3, and PC2/PC3 with points colored by `chd`.

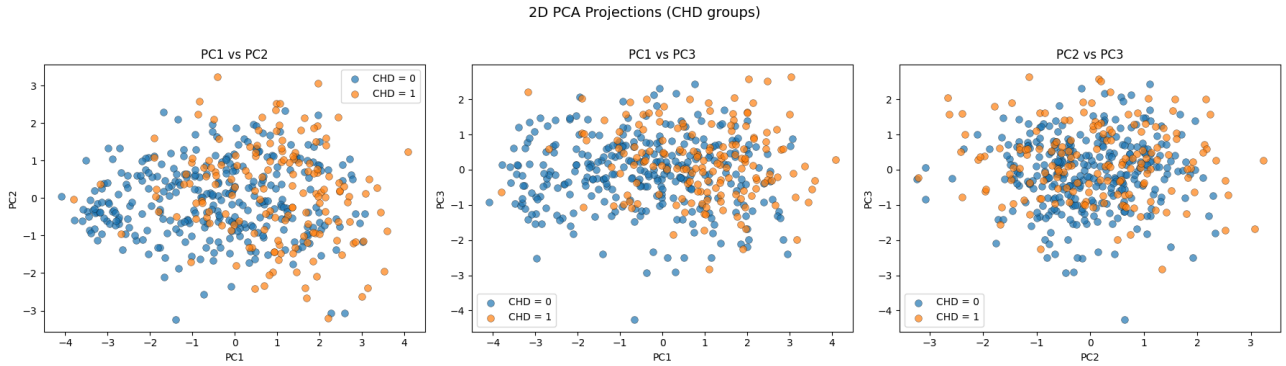

Figure 8: 2D CPA Projections (CHD Groups)

We can see that relatively high values of PC1 are related to larger proportions of positive `chd` (`chd` = 1). In other words, there is dominance of positive `chd` in the positive part of PC1. On the other hand, relatively low values of PC1 are mostly related to negative `chd` (`chd` = 0). However, given that PC1 only explains around 35% of the variance and there is no clear boundary between `chd` values. The plot of the data points with respect to this principal component is not enough to confidently determine the outcome of the feature `chd`.

PCA with blood pressure as target: We applied the same PCA procedure with `sbp` as the target variable (and added `chd` back among the features). Conclusions were very similar, and we needed almost the same number of principal components to reach the 90% explained variance threshold. See the Appendix for the corresponding visualization. For more details, refer to the appendix for the corresponding visualizations - Appendix A.

4 Discussion

The dataset contains several features that are relevant for the classification task: distinguishing individuals who will develop heart disease (`chd` = 1) from those who will not (`chd` = 0).

The correlation matrix provided a first step in identifying variables most closely related to the outcome. For instance, `adiposity` and `obesity` showed a strong correlation, suggesting redundancy; one of these features could potentially be removed in later classification or regression models. `Age` also correlates with `adiposity`, reflecting the tendency for body fat to increase with age as muscle mass declines. A weaker but noticeable correlation exists between `age` and `tobacco`, indicating that older individuals tend to report higher tobacco use. Overall, `age` emerges as the feature most consistently related to both `chd` and `sbp`.

Dimensionality reduction via PCA did not yield a substantial decrease in the number of features; however, we reduced the feature space from nine to six principal components while retaining over 90% of the variance. Relationships seen in the correlation matrix were reflected in the principal components: `adiposity` and `obesity` followed similar trends across several components (PC1, PC2, PC4, PC5, PC6), reinforcing the redundancy previously identified.

When projecting onto PC1, higher values were associated with a greater proportion of CHD-positive cases. Similarly, in the PC1–PC3 projection, CHD-positive individuals were more dominant when both components had relatively large positive values. This suggests that principal components capture meaningful structure related to CHD occurrence.

References

[1] J. Rousseauw et al. "Coronary risk factor screening in three rural communities. The CORIS baseline study". In: *South African Medical Journal* 64.12 (1983), pp. 430–436. URL: https://pubmed.ncbi.nlm.nih.gov/6623218/.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Data Sets*. https://hastie.su.domains/ElemStatLearn/. Accessed: 2025-09-26.

[3] Veena Potdar, Lavanya Santhosh, and Likith Jadhav. "Coronary Heart disease prediction using machine learning". In: 9 (Dec. 2022), e390–e396.

[4] Hisham Khdair and Naga M Dasari. "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 12.5 (2021), pp. 28–35.

A Appendix

Python notebook for all plots and analysis can also be found by visiting this github repository

PCA

Principal component analysis applied with `sbp` as the target variable.

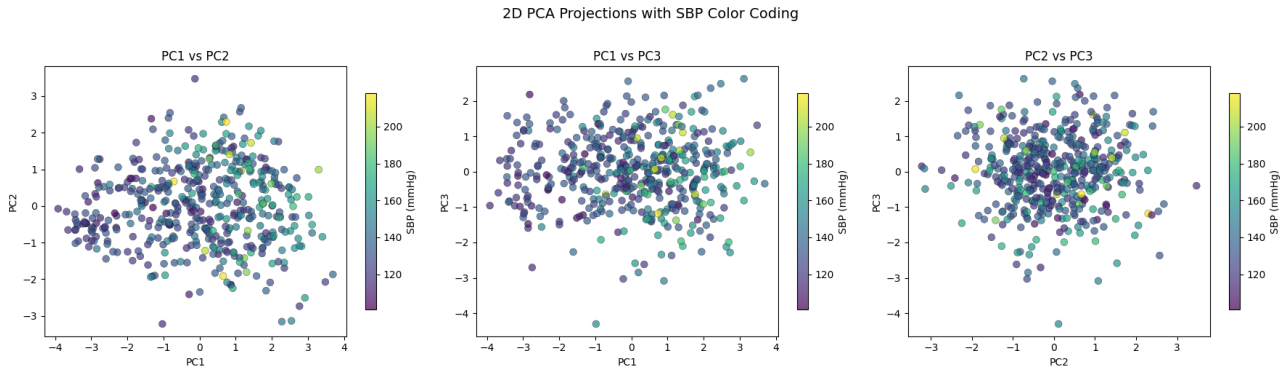

Figure 9: 2D PCA projections of the dataset with `sbp` values used as continuous color coding (PC1 vs PC2, PC1 vs PC3, PC2 vs PC3).

Data classified with following logic, see below:

- Normal: 0–120
- Elevated: 120–130
- High Stage 1: 130–140
- High Stage 2: 140–180
- Crisis: 180–300

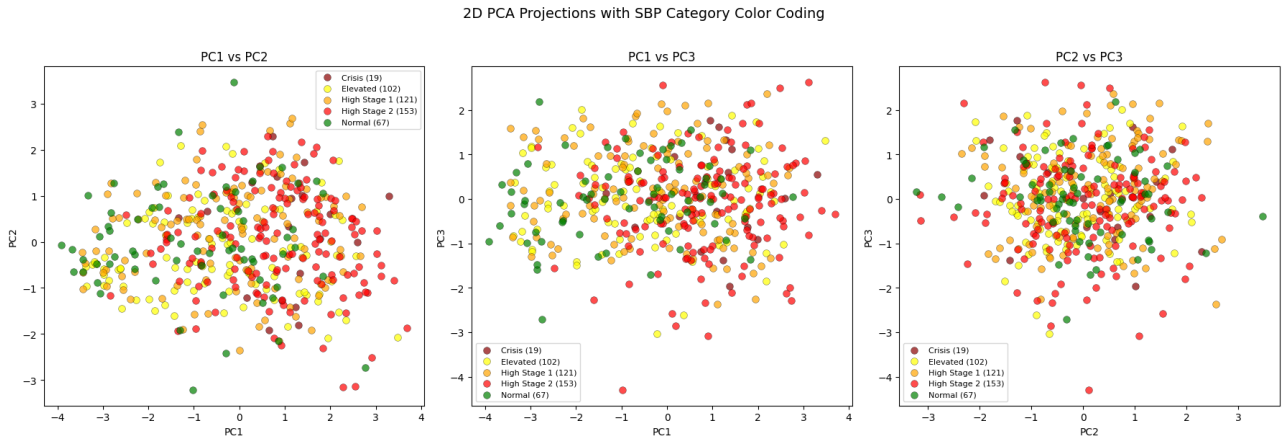

Figure 10: 2D PCA projections of the dataset with categorical `sbp` classes (Normal, Elevated, High Stage 1, High Stage 2, Crisis).

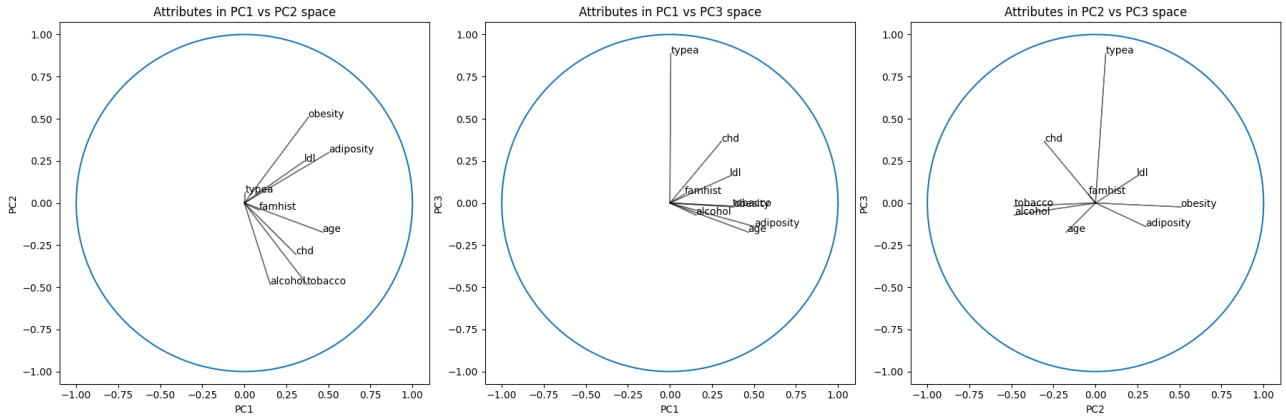

Figure 11: PCA biplots showing the contribution of each attribute in the PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3 spaces.

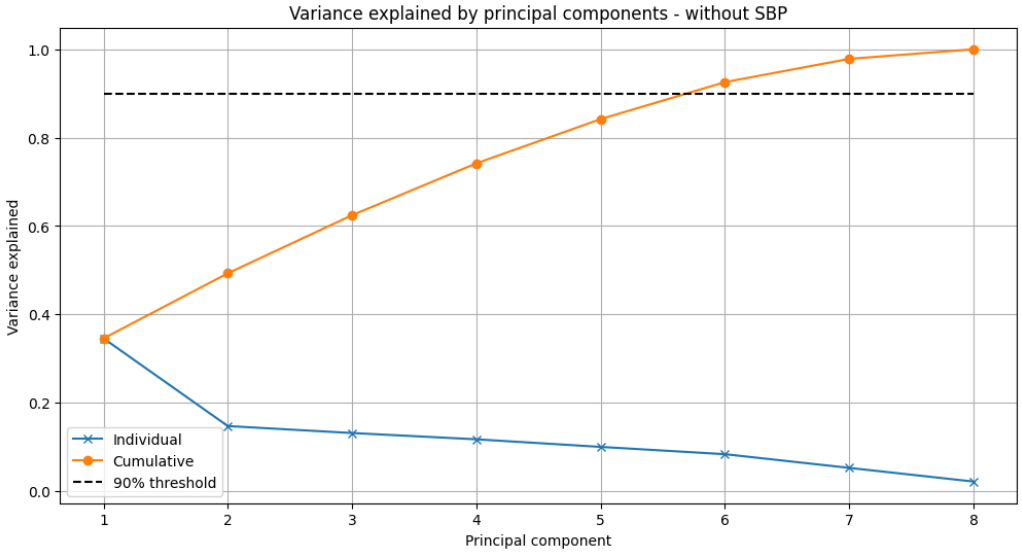

Figure 12: Explained variance ratio by principal components (individual and cumulative) with 90% threshold line, excluding `sbp`.