STATISTICAL LEARNING PROJECT

Detecting Heart Disease

Doina Vasilev
Università degli Studi di Milano

TABLE OF CONTENT

01. **THE HEART DATA DATASET**

02. **DATA PRE-PROCESSING**

03. **SUPERVISED LEARNING**
Decision Trees, Bagging & Random Forests

04. **UNSUPERVISED LEARNING**
Hierarchical Clustering

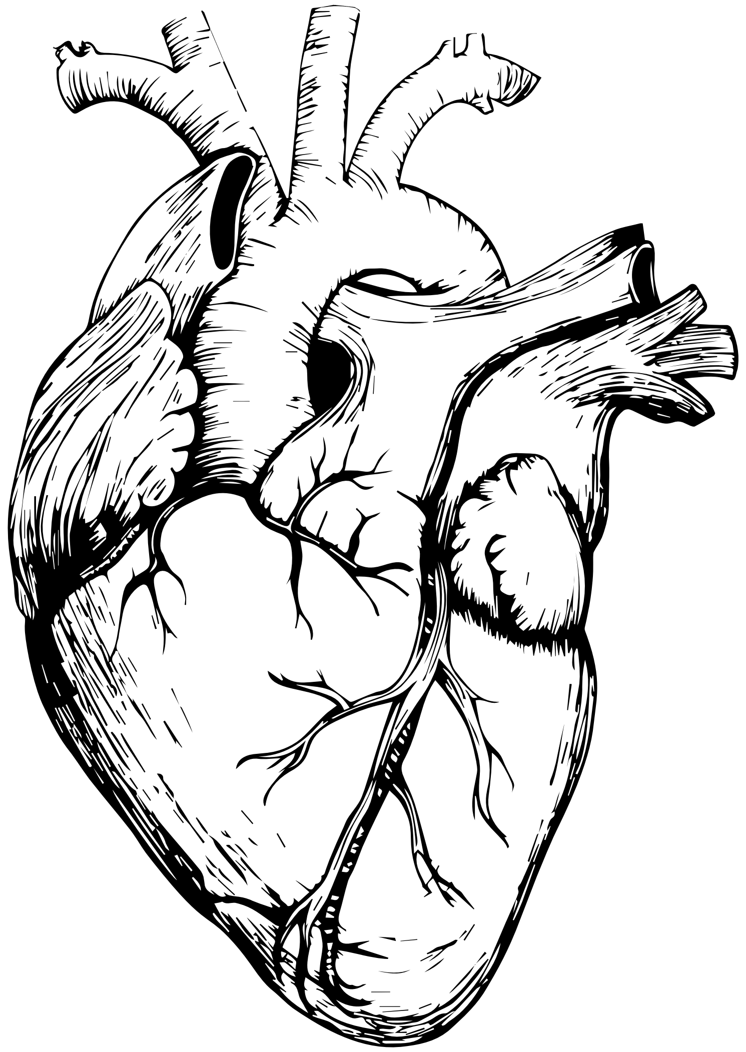
THE DATASET

The dataset is retrieved from the UCI Machine Learning Repository, i.e., a collection of databases, domain theories, and data generators provided by the University of California.

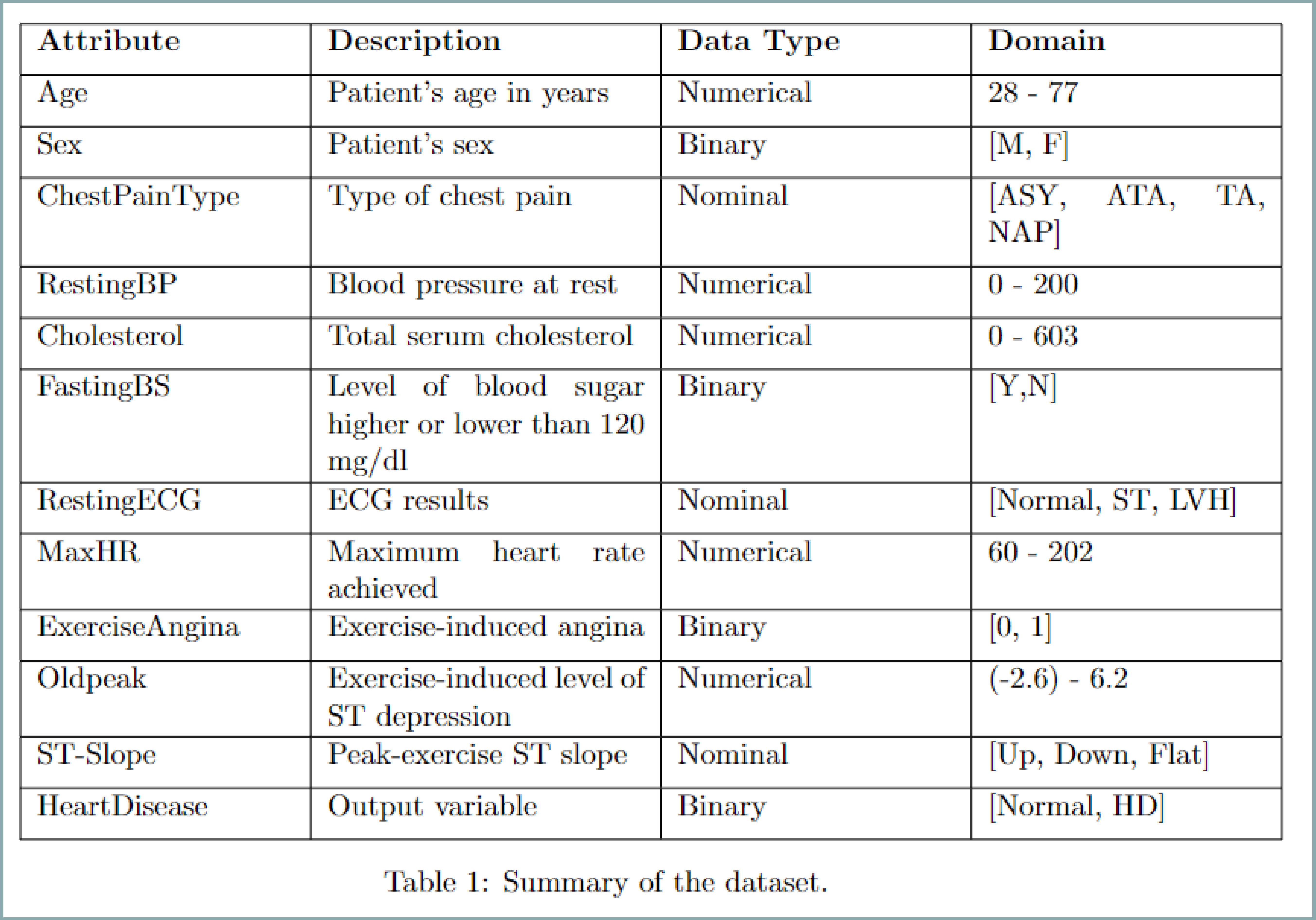
The dataset is made up of 918 observations, 11 predictors and 1 response variable: **HeartDisease**.

Among the predictors, there are **5 numerical** variables and **5 categorical** variables.
The response variable is binary, and the possible outcomes are "**HD**", when the patient is predicted to have an heart disease, and "**Normal**" when it's not.

The aim of my project is to build models that are able to accurately predict the arise of heart disease.

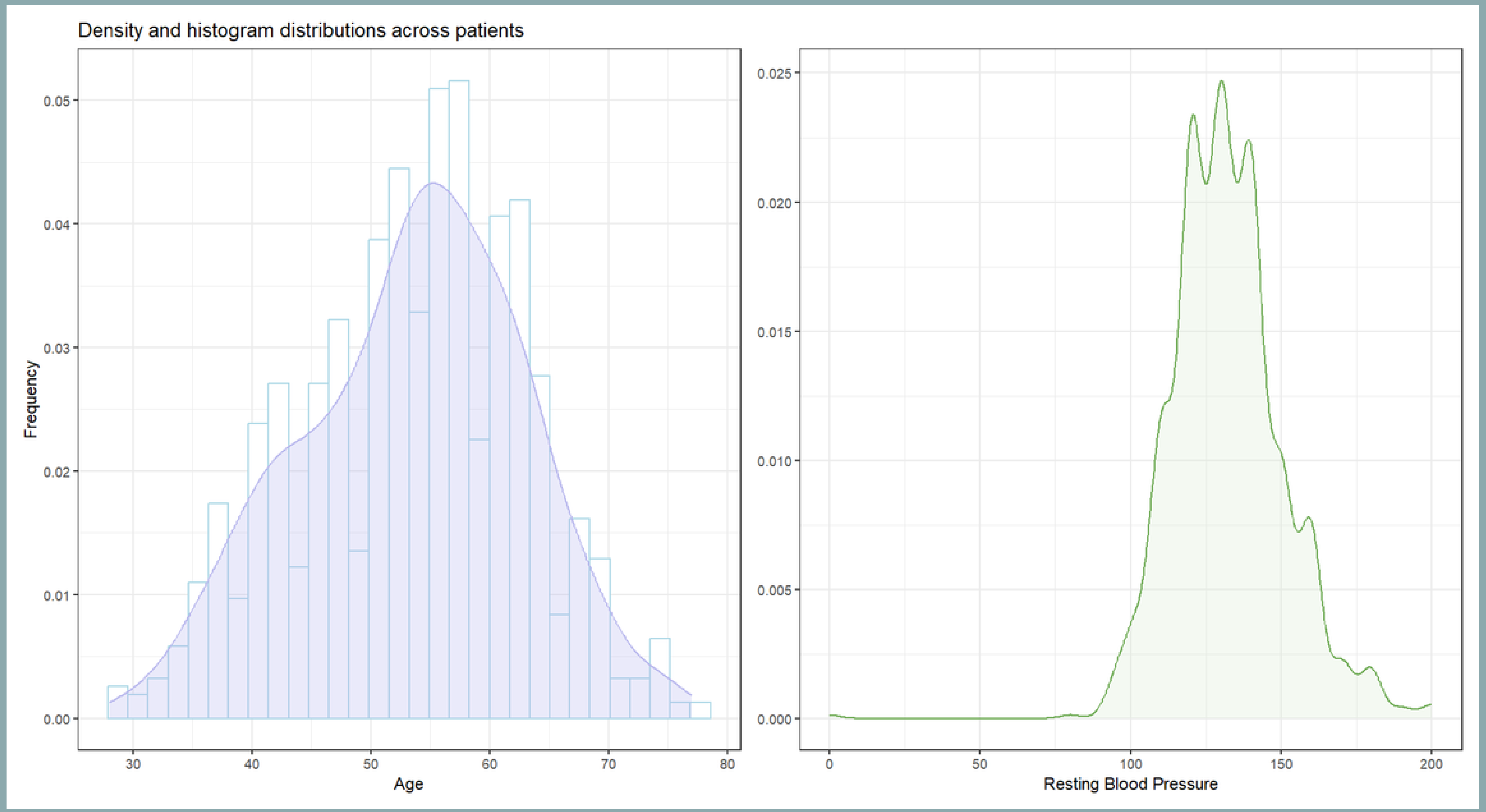


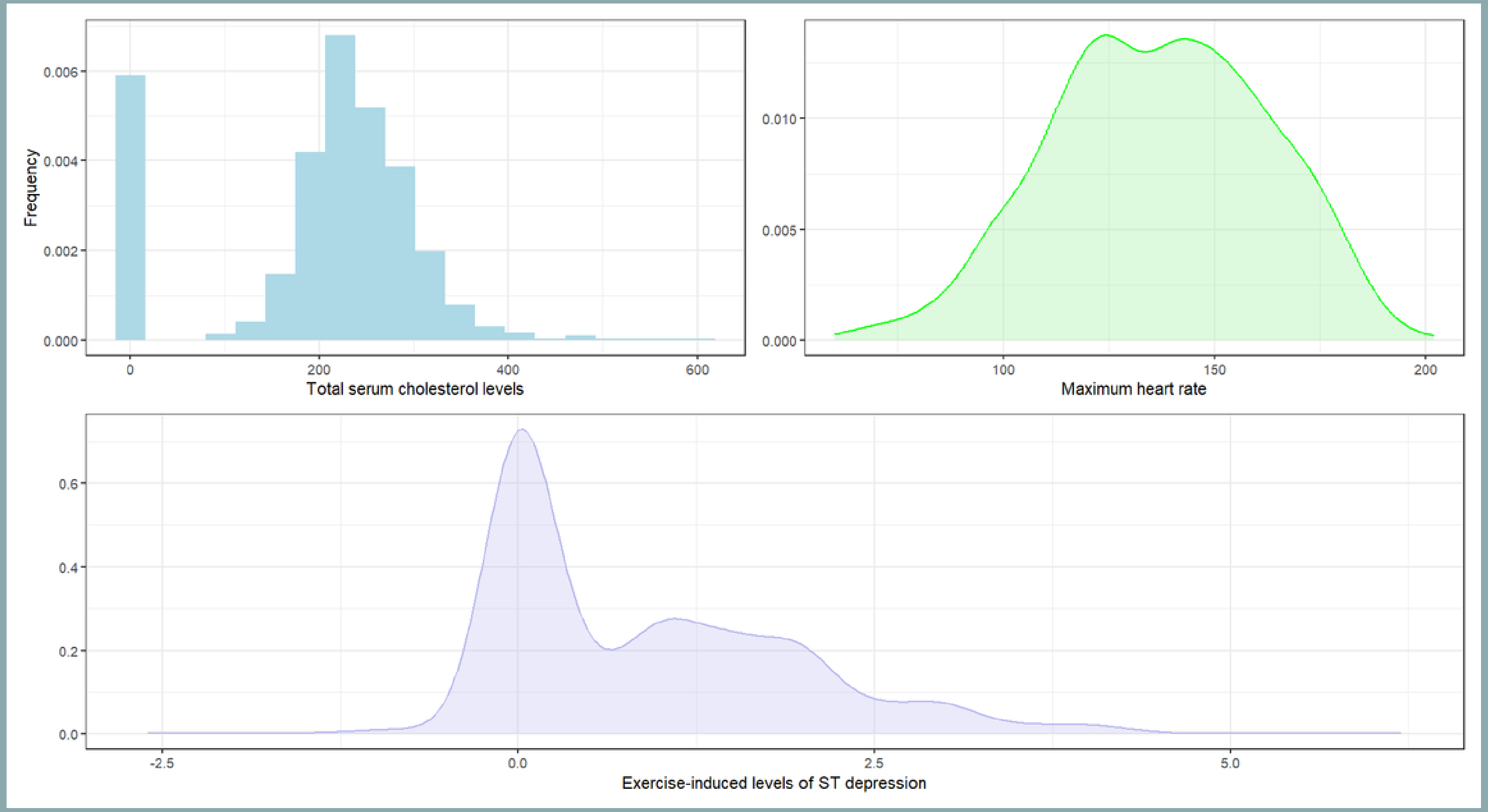
SUMMARY OF THE DATASET



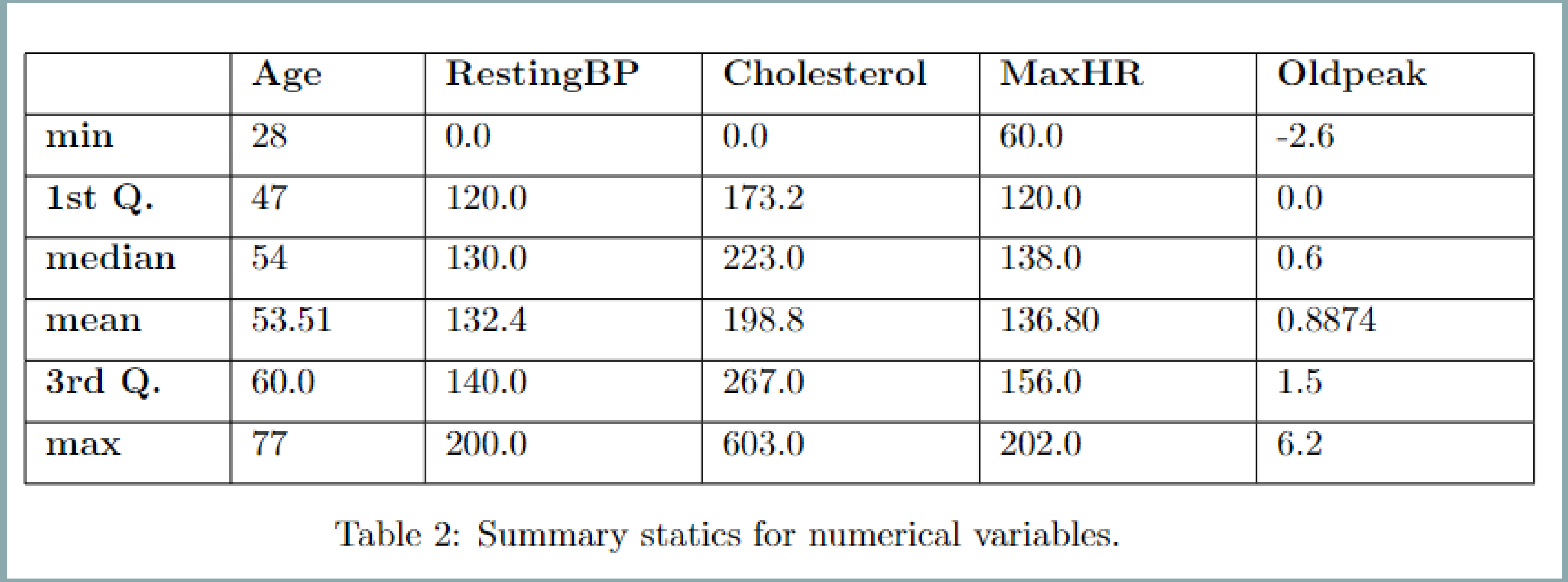
Attribute	Description	Data Type	Domain
Age	Patient's age in years	Numerical	28 - 77
Sex	Patient's sex	Binary	[M, F]
ChestPainType	Type of chest pain	Nominal	[ASY, ATA, TA, NAP]
RestingBP	Blood pressure at rest	Numerical	0 - 200
Cholesterol	Total serum cholesterol	Numerical	0 - 603
FastingBS	Level of blood sugar higher or lower than 120 mg/dl	Binary	[Y,N]
RestingECG	ECG results	Nominal	[Normal, ST, LVH]
MaxHR	Maximum heart rate achieved	Numerical	60 - 202
ExerciseAngina	Exercise-induced angina	Binary	[0, 1]
Oldpeak	Exercise-induced level of ST depression	Numerical	(-2.6) - 6.2
ST-Slope	Peak-exercise ST slope	Nominal	[Up, Down, Flat]
HeartDisease	Output variable	Binary	[Normal, HD]

Table 1: Summary of the dataset.





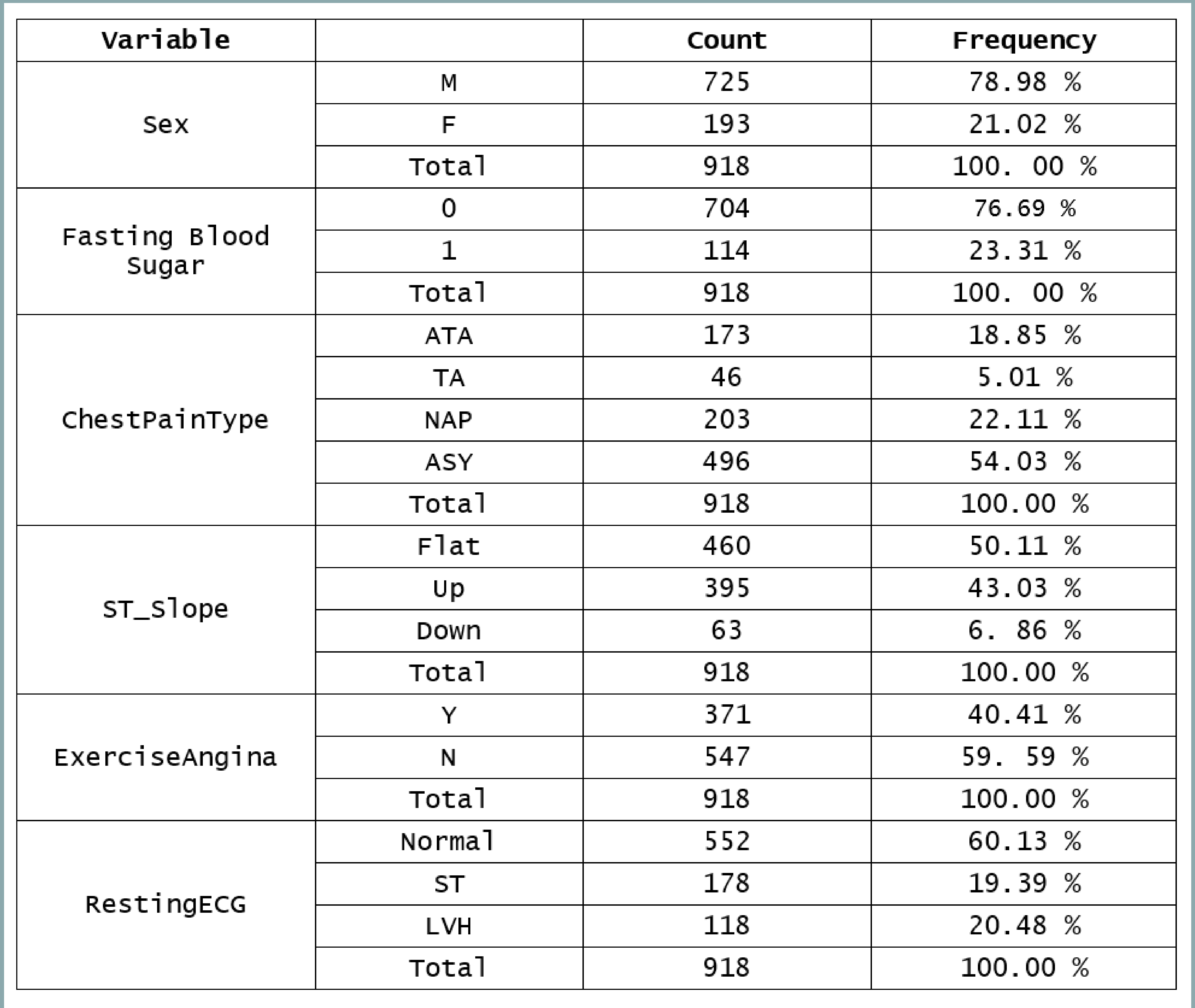
SUMMARY OF THE QUANTITATIVE VARIABLES



	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
min	28	0.0	0.0	60.0	-2.6
1st Q.	47	120.0	173.2	120.0	0.0
median	54	130.0	223.0	138.0	0.6
mean	53.51	132.4	198.8	136.80	0.8874
3rd Q.	60.0	140.0	267.0	156.0	1.5
max	77	200.0	603.0	202.0	6.2

Table 2: Summary statics for numerical variables.

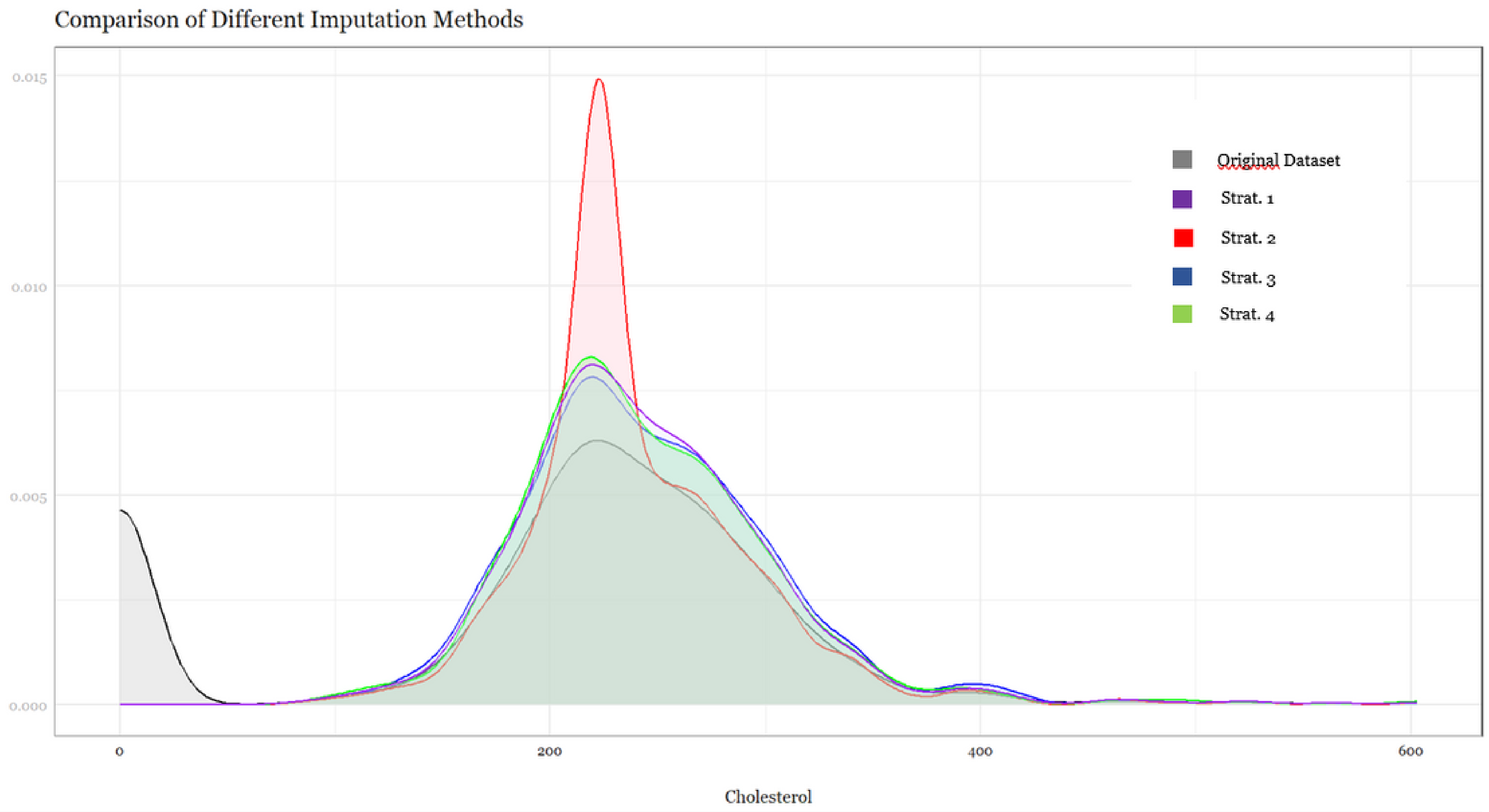
SUMMARY OF THE QUALITATIVE VARIABLES

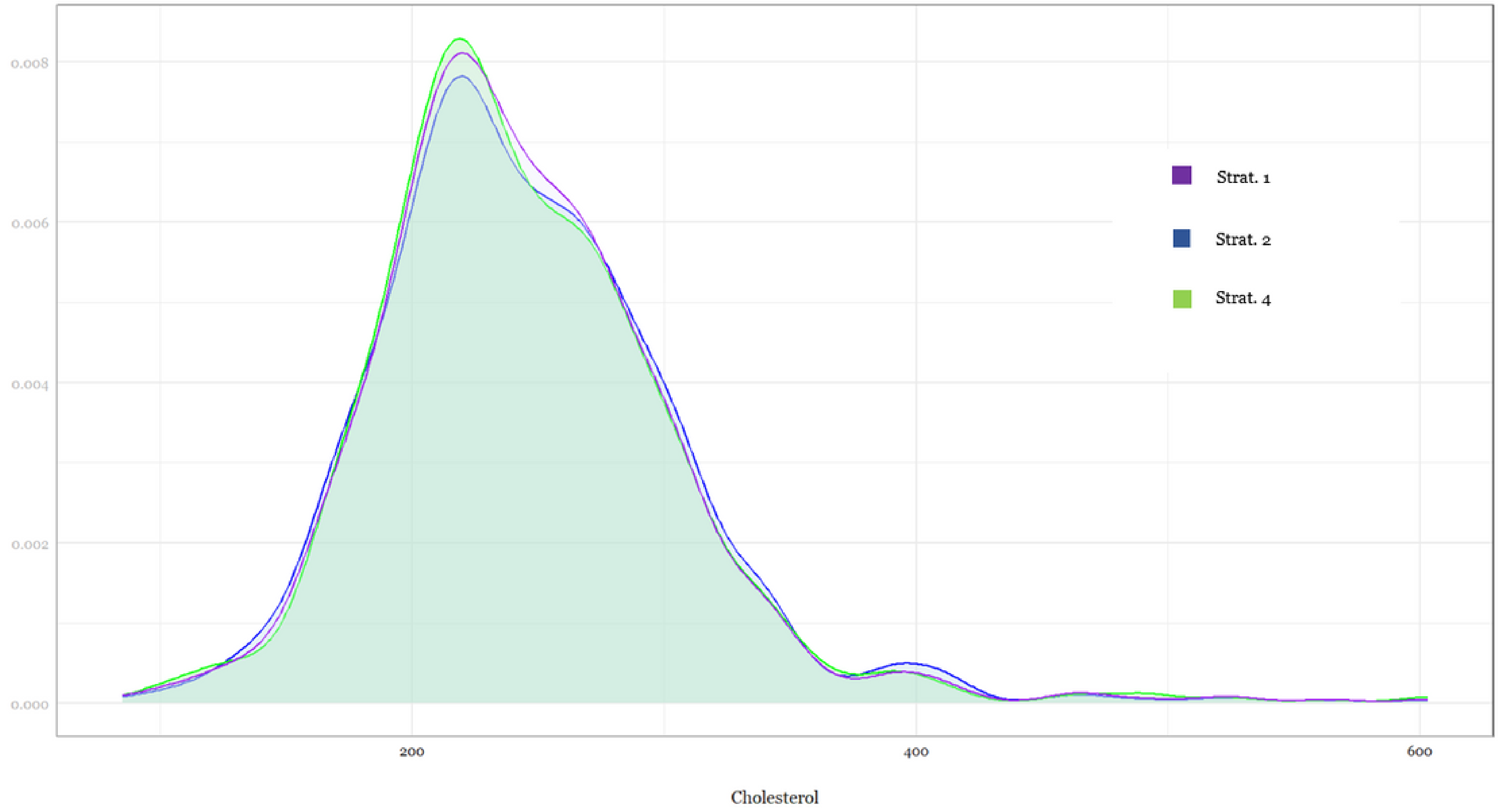


Variable		Count	Frequency
Sex	M	725	78.98 %
	F	193	21.02 %
	Total	918	100. 00 %
Fasting Blood Sugar	0	704	76.69 %
	1	114	23.31 %
	Total	918	100. 00 %
ChestPainType	ATA	173	18.85 %
	TA	46	5.01 %
	NAP	203	22.11 %
	ASY	496	54.03 %
	Total	918	100.00 %
ST_Slope	Flat	460	50.11 %
	Up	395	43.03 %
	Down	63	6. 86 %
	Total	918	100.00 %
ExerciseAngina	Y	371	40.41 %
	N	547	59. 59 %
	Total	918	100.00 %
RestingECG	Normal	552	60.13 %
	ST	178	19.39 %
	LVH	118	20.48 %
	Total	918	100.00 %

DATA PRE-PROCESSING

- **Removed the "dead" person**, i.e., observation with RestingBP equal to 0.
- Applied **random forests** to replace the 0s to the variable Cholesterol.
- Cleaned the dataset from the most extreme values.
- Applied resampling methods to balance **Sex** and **FastingBS**.





DIFFERENT IMPUTATION METHODS

I've applied several imputation stategy to replace the zeros in Cholesterol:

1. Removed all the observations where Cholesterol = 0;

2. Replaced the zeros with the *median* value.

3. Used *pmm* algorithm.

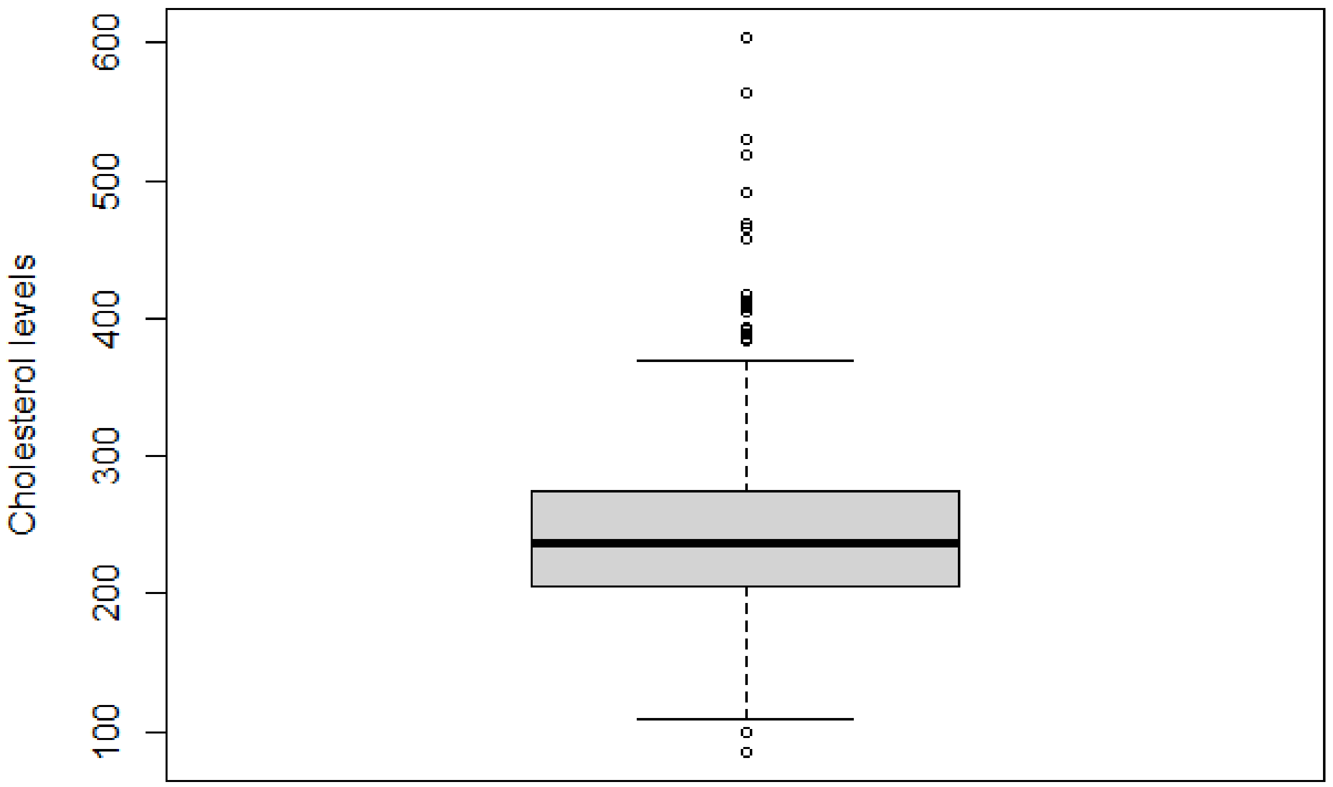
4. Used *random forests* algorithm.

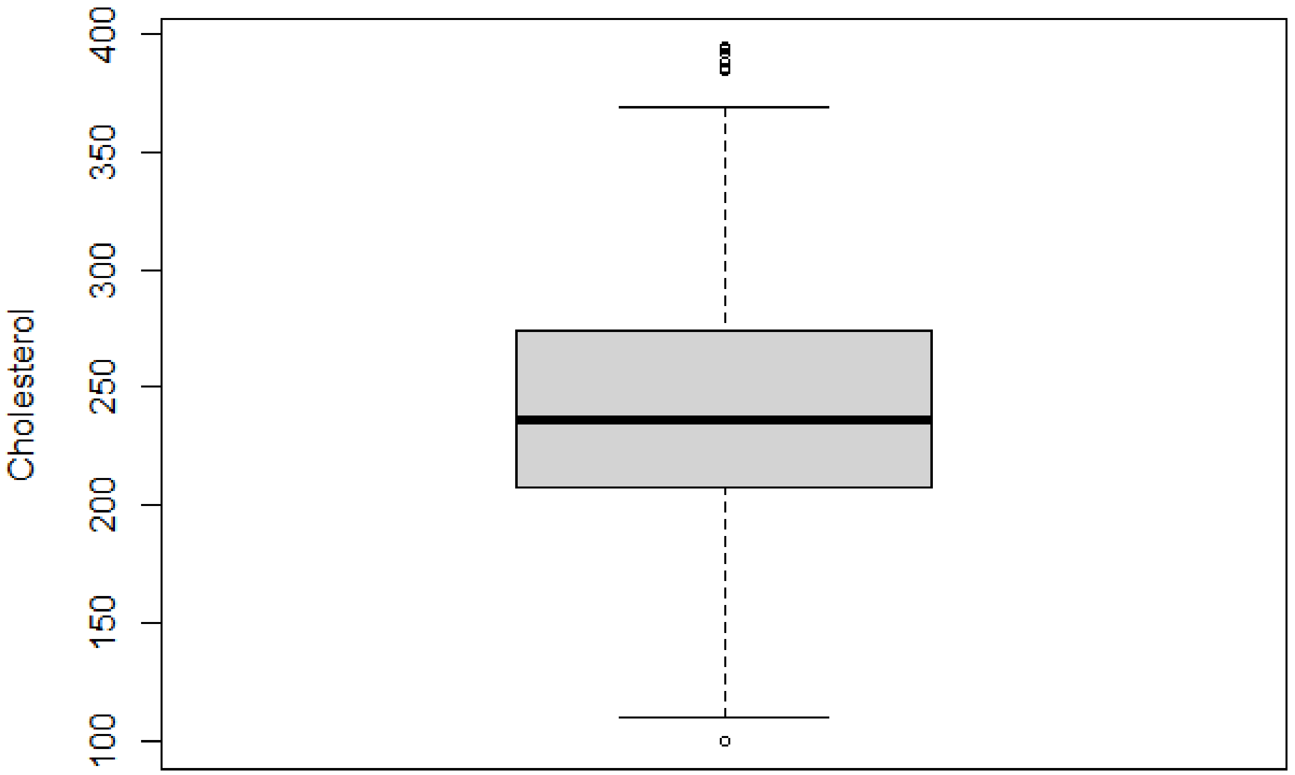
REMOVED ONLY THE MOST EXTREME OUTLIERS

I've only removed the extreme values from Cholesterol, for a few reasons:

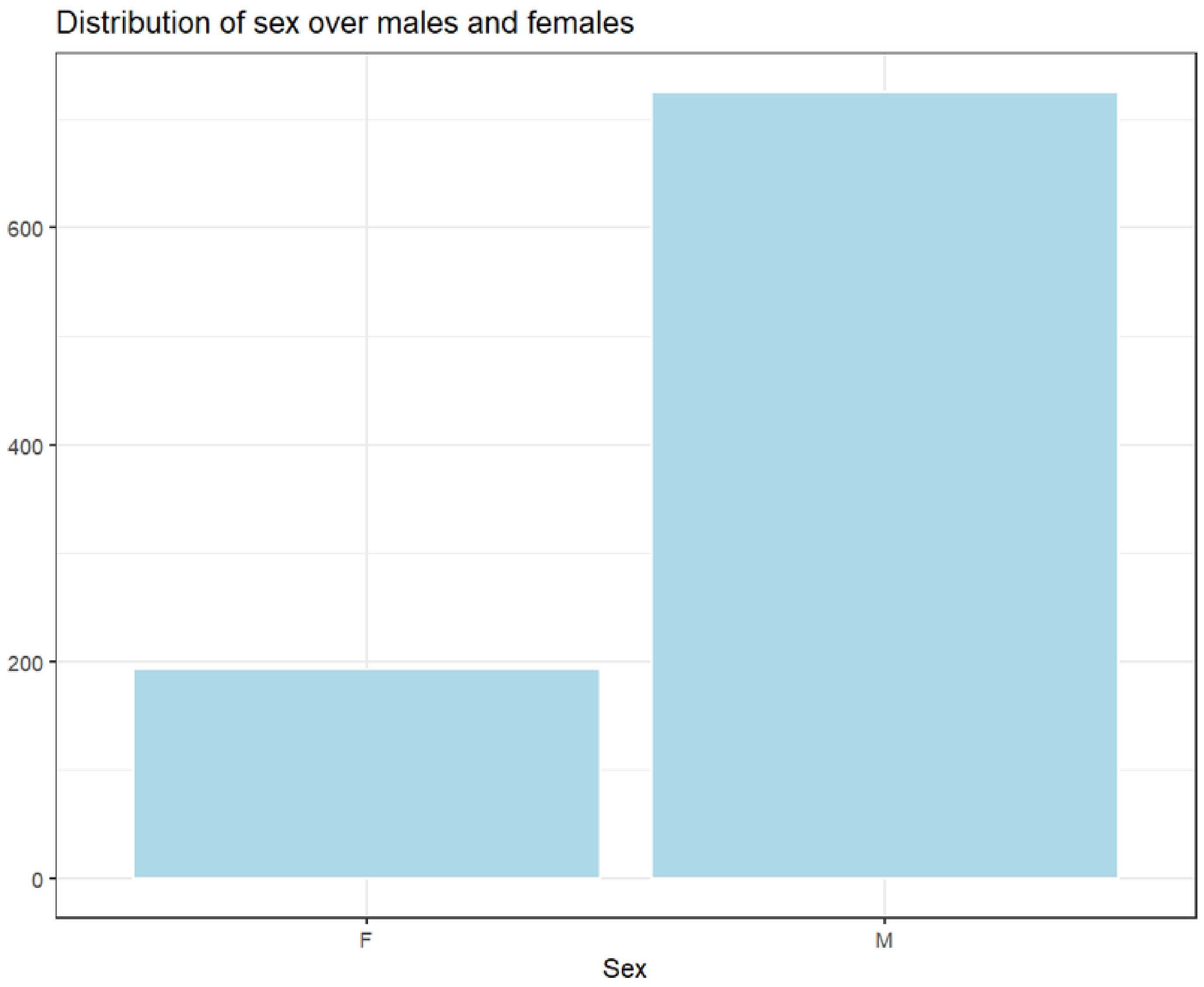
- **I'm going to use decision trees and random forests, which are not sensitive to outliers, thus their performance won't be altered much the presence of some outliers.**

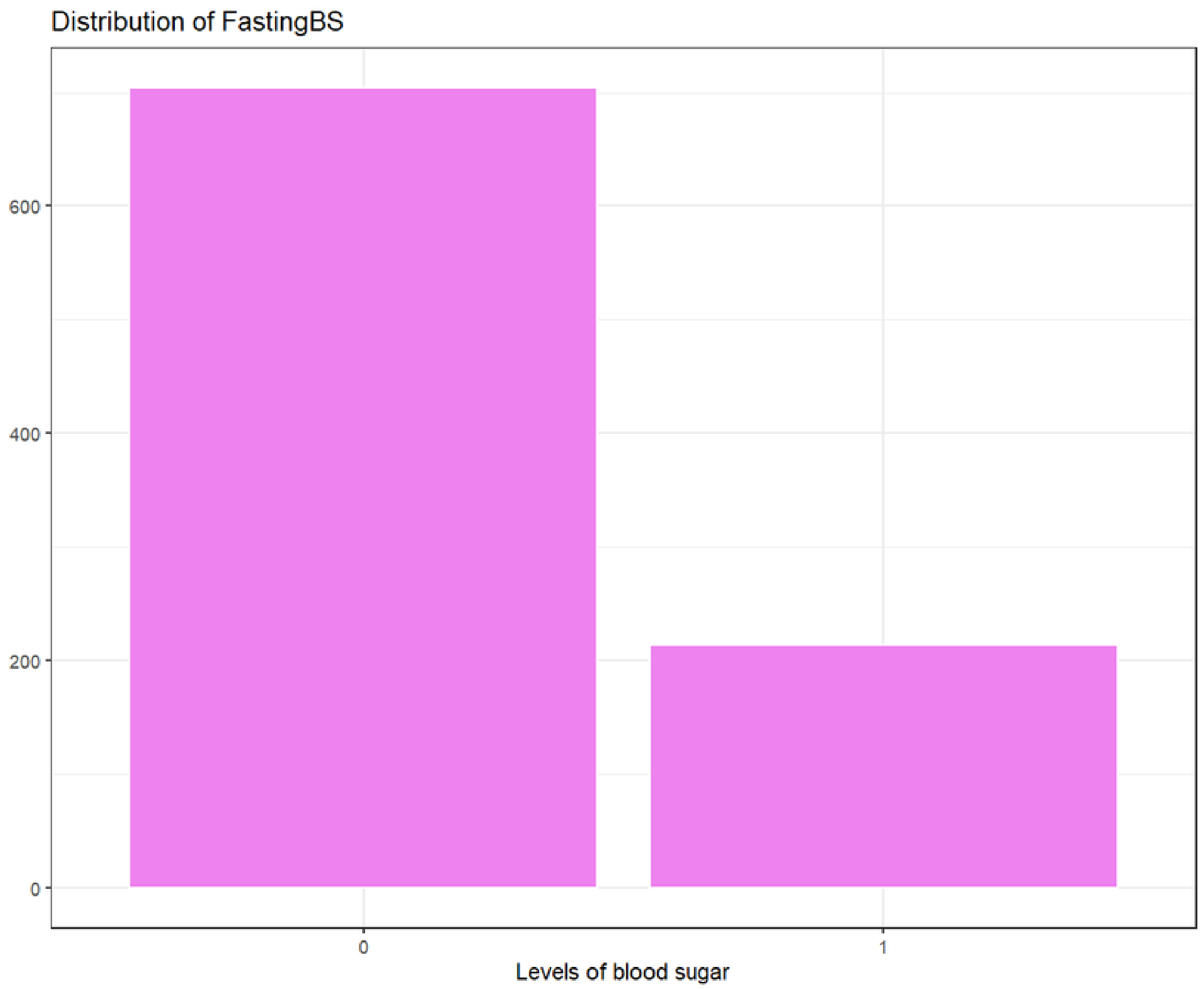
- **In the medical field outliers could represent important information, that should be preserved.**





RE-BALANCED THE CLASSES OF SEX AND FASTINGBS





I've applied random resampling techniques to re-balance the distribution of the observations across the classes of Sex and FastingBS.
I've first applied *random undersampling* to Sex, and then *random oversampling* to FastingBS (*undersampling*, and *oversampling* functions in *caret*).
The final datset contains 610 observations, where the class distributions over Sex and FastingBS are split evenly.

SUPERVISED LEARNING
Decision Trees

I've finally fit a decision tree on each sample I've created:

- *rf* - on which I've only imputed the missing values of Cholesterol.

- *clean* - on which I've removed the extreme outliers.

- *train* - on which I've performed re-sampling over the imbalanced classes.

I've measured the performance of each tree on test samples based on several performance metrics:

- **Accuracy**, as the rate of the correctly predicted classes over the total number of predictions.
- **Sensitivity**, as the rate of correctly predicted positives (HD) over all the predictions of HD.
- **Sensibility**, as as the rate of correctly predicted negatives (Normal) over all the predictions of Normal.
- **Precision**, as the rate of correctly predicted positives (HD), when HD is true.
- **Recall**, as the rate of correctly predicted negatives (Normal), when lack of HD is true.

SUPERVISED LEARNING

Decision Trees

Training Set	Test Set	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
train	rf	0.8451	0.8600	0.8268	0.8789	0.8304	0.8550
clean[train,]	clean[-train,]	0.8333	0.9175	0.7349	0.8018	0.9175	0.856
rf[train,]	rf[-train,]	0.8207	0.7921	0.8554	0.8696	0.7921	0.8296

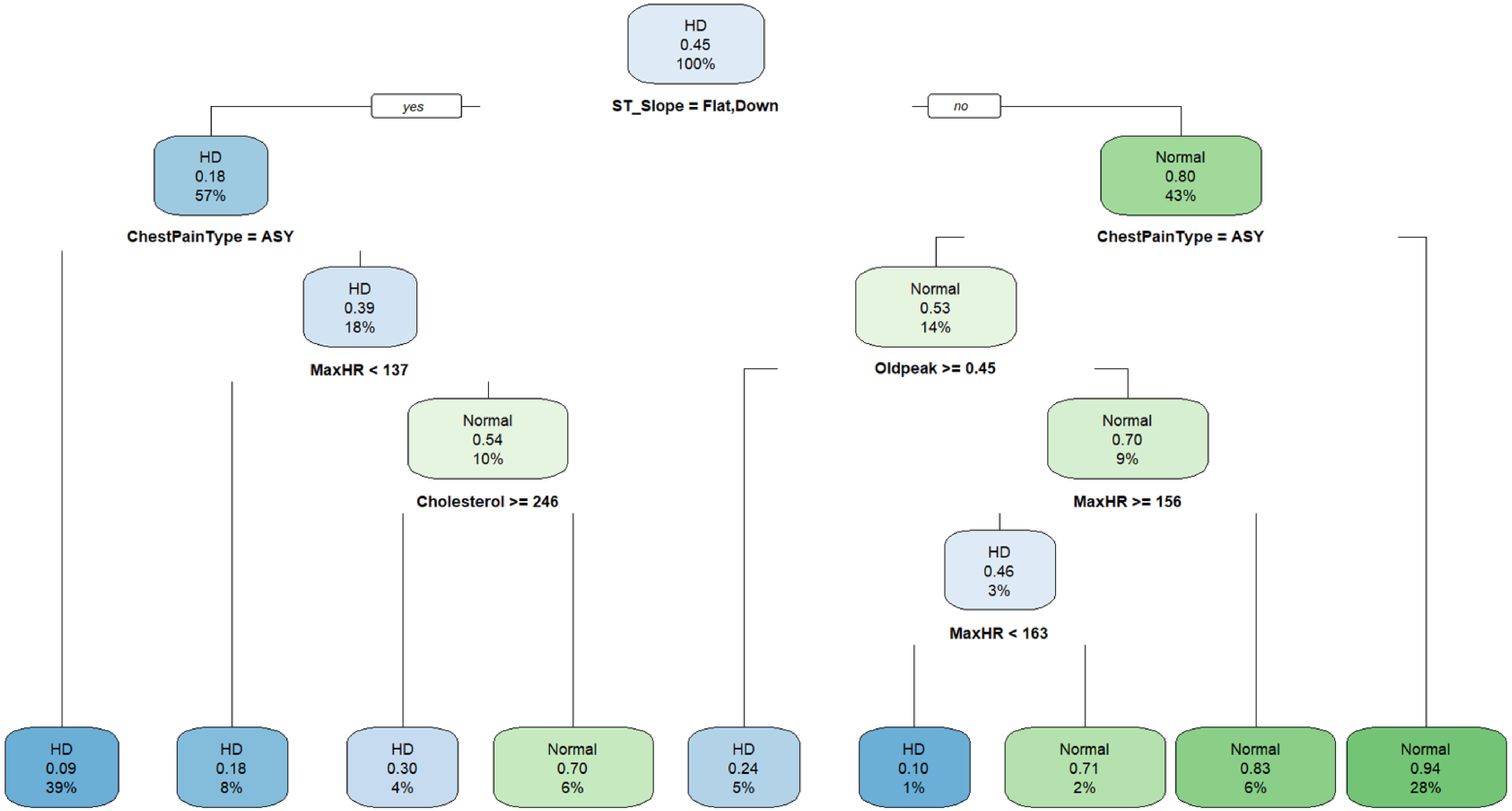
Overall, the best performing tree is the one trained on the *clean* dataset.

SUPERVISED LEARNING

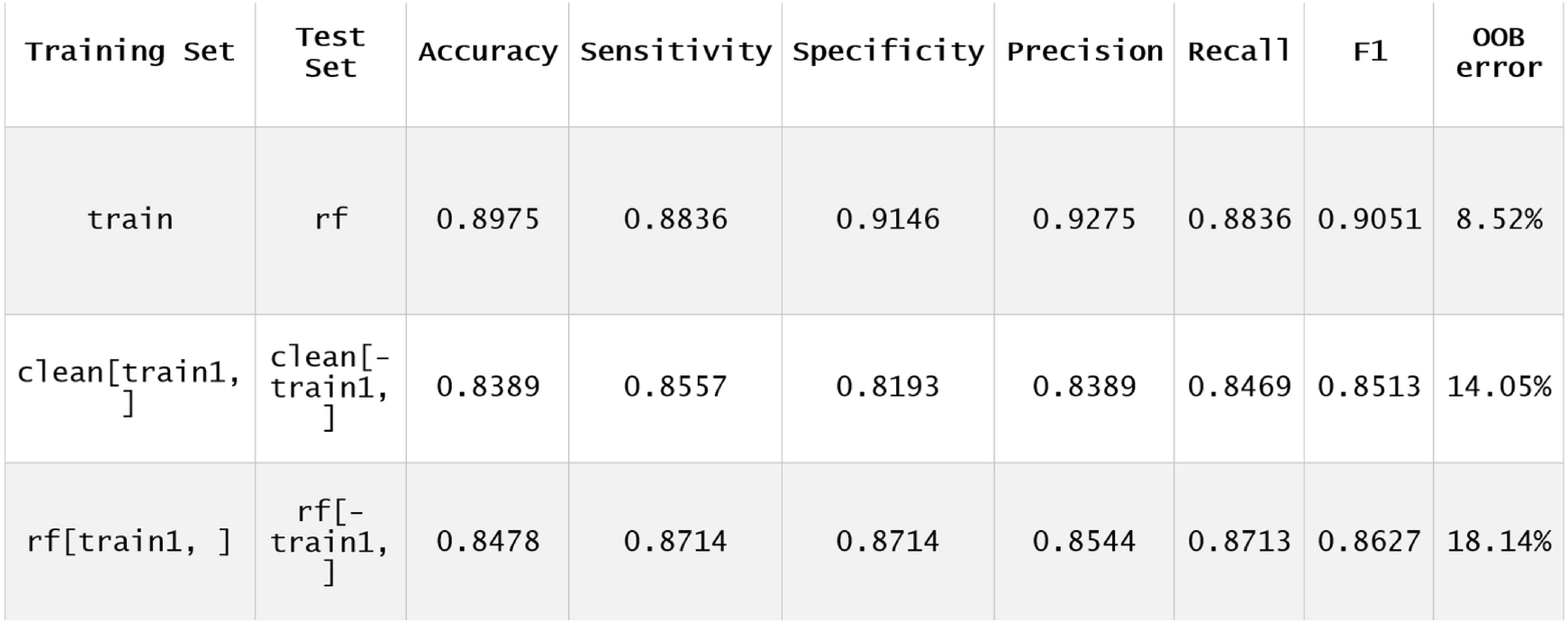
Decision Trees - Best Performing Tree

ST_Slope and **ChestPainType** seem to be quite important variables in the detection of heart disease.

However, these results could be due to some further class imbalance on the two variables.



SUPERVISED LEARNING
Bagging

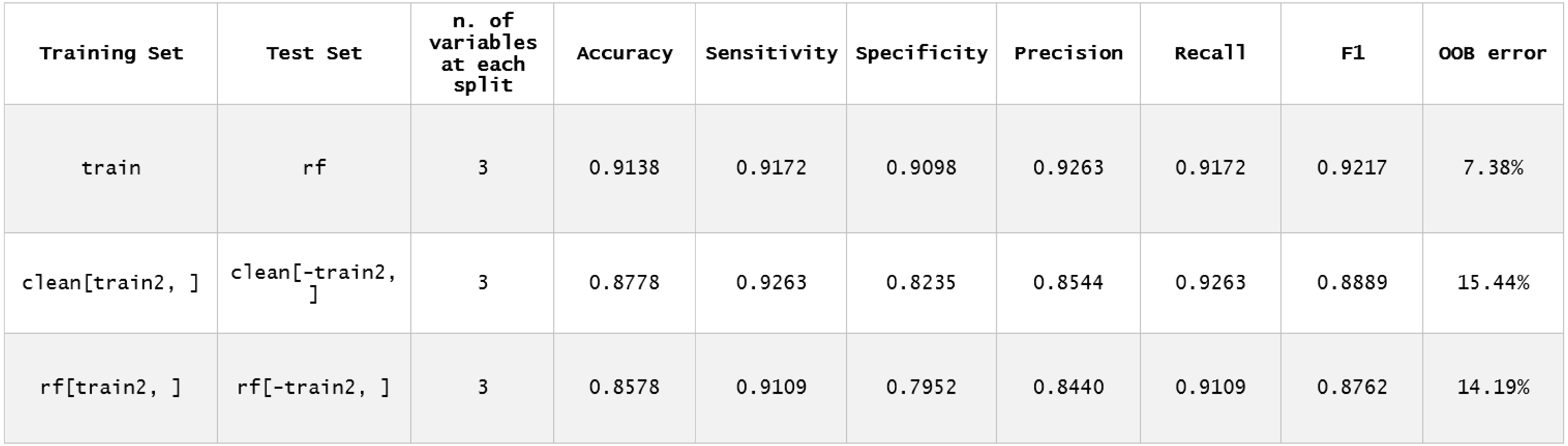


Training Set	Test Set	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	OOB error
train	rf	0.8975	0.8836	0.9146	0.9275	0.8836	0.9051	8.52%
clean[train1,]	clean[-train1,]	0.8389	0.8557	0.8193	0.8389	0.8469	0.8513	14.05%
rf[train1,]	rf[-train1,]	0.8478	0.8714	0.8714	0.8544	0.8713	0.8627	18.14%

Overall, the best performing bagging tree is the one trained on the *train* dataset.
Since it's *bagging*, the number of variables picked for each tree is the total number of variables.

In this case, I've also added the Out-Of-Bag error estimate, which is an estimate of the prediction errors on the *out-of-bag* samples, i.e., the ones on which the trees have not been fit.

SUPERVISED LEARNING
Random Forests



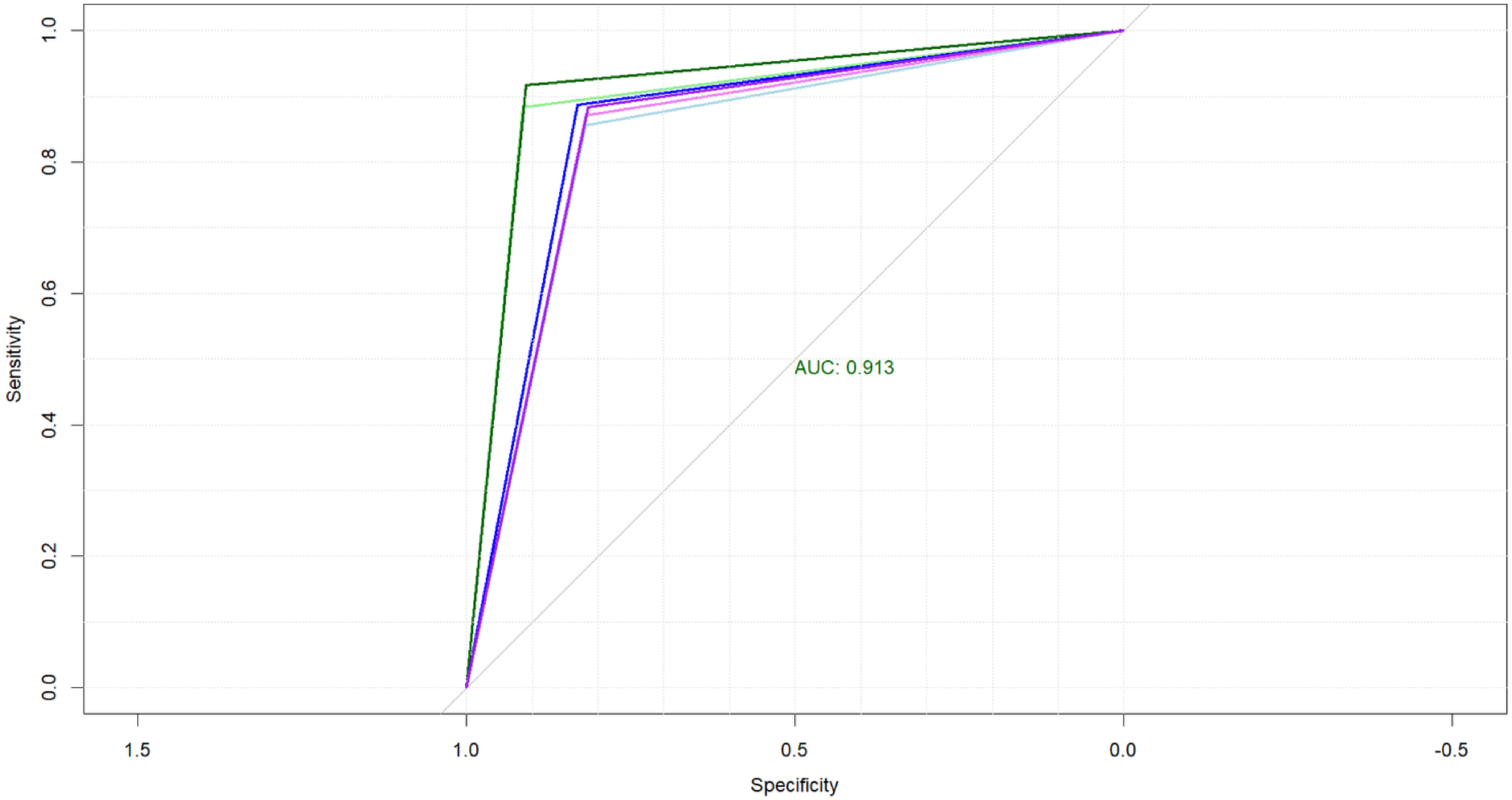
Training Set	Test Set	n. of variables at each split	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	OOB error
train	rf	3	0.9138	0.9172	0.9098	0.9263	0.9172	0.9217	7.38%
clean[train2,]	clean[-train2,]	3	0.8778	0.9263	0.8235	0.8544	0.9263	0.8889	15.44%
rf[train2,]	rf[-train2,]	3	0.8578	0.9109	0.7952	0.8440	0.9109	0.8762	14.19%

Overall, the best performing random forest is the one trained on the *train* dataset.
In this case, the number of variables picked to build each tree on the training set is 3.

SUPERVISED LEARNING

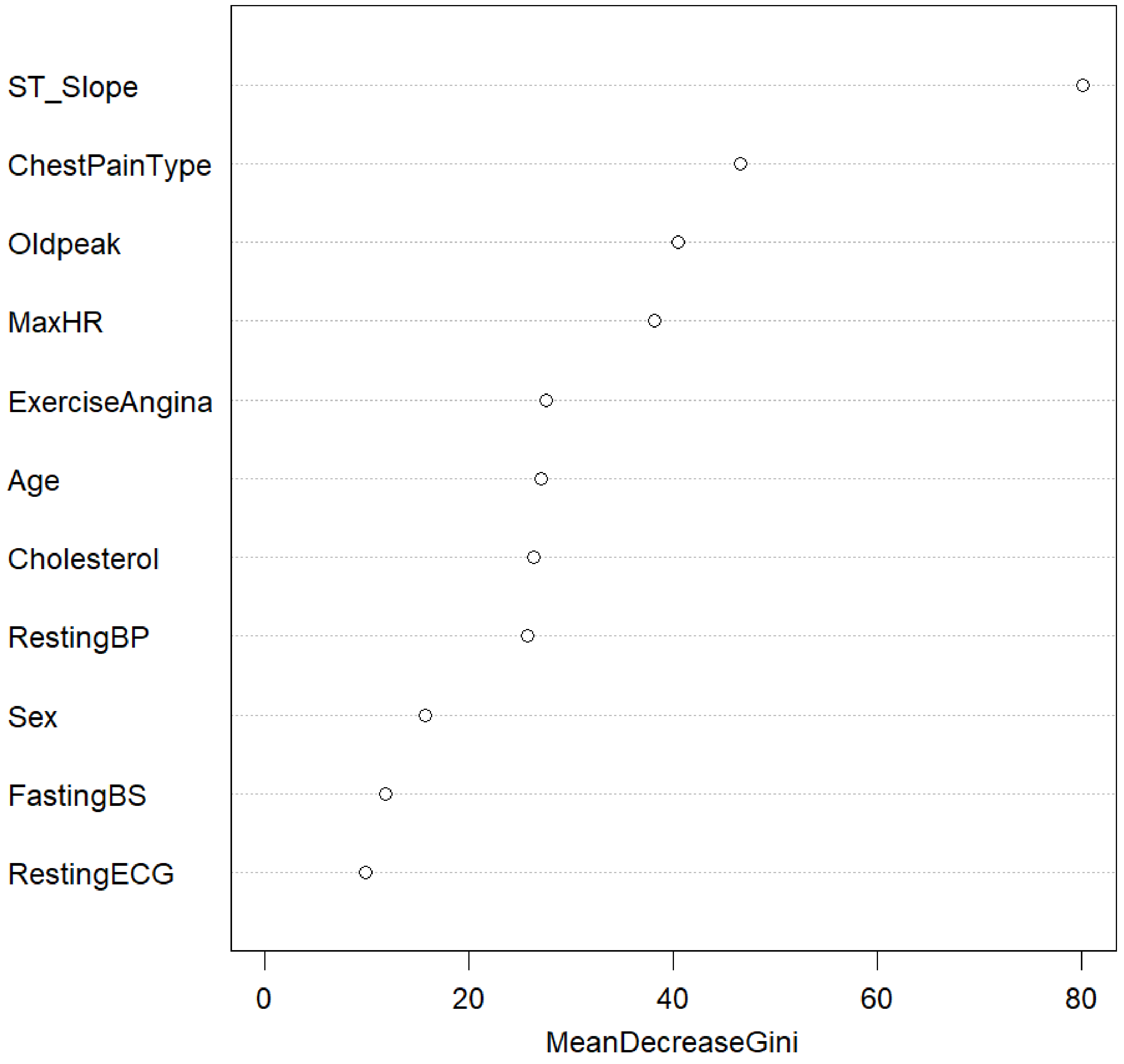
Random Forests & Bagging - Best Performing Models

Among all the possible bagged trees or random forests,
the best performing classifier is the random forest built on the *train* sample.



SUPERVISED LEARNING

Random Forests - Variable Importance



The **mean decrease in Gini** coefficient is a measure of how each variable contributes to the purity of the nodes and leaves in the resulting random forest.

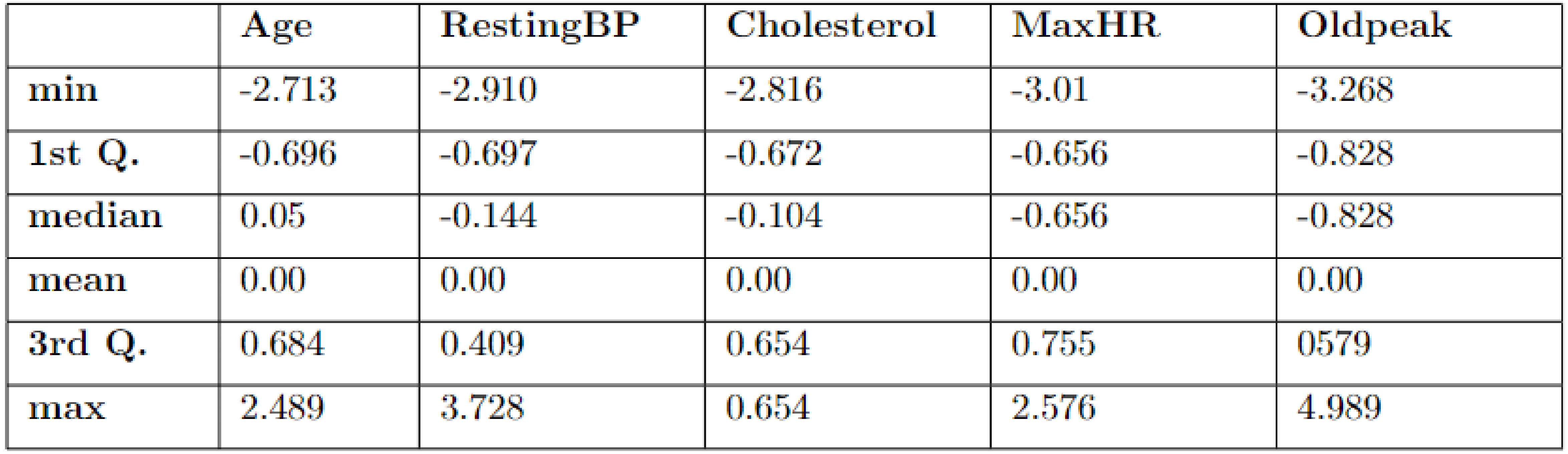
The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model.

In this case, **ST_Slope** and **ChestPainType** are again the 2 most important variables; however, this could be resulting due to the relative class imbalance of the two variables.

UNSUPERVISED LEARNING

Data Pre-Processing

In order to apply Hierarchical Clustering, I first re-scaled all the numerical variables.

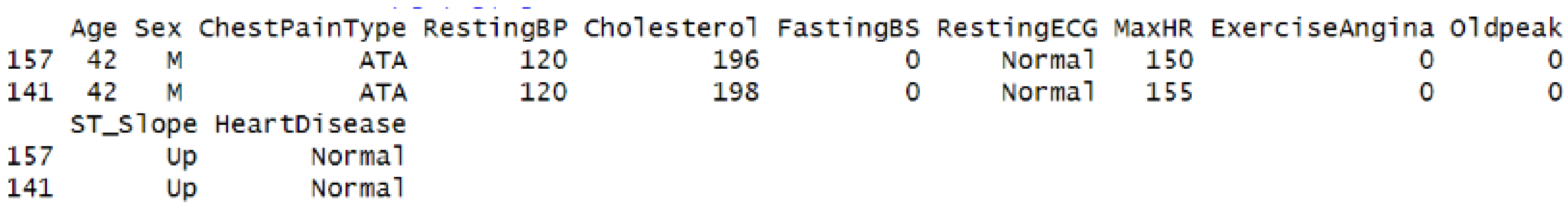


	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
min	-2.713	-2.910	-2.816	-3.01	-3.268
1st Q.	-0.696	-0.697	-0.672	-0.656	-0.828
median	0.05	-0.144	-0.104	-0.656	-0.828
mean	0.00	0.00	0.00	0.00	0.00
3rd Q.	0.684	0.409	0.654	0.755	0579
max	2.489	3.728	0.654	2.576	4.989

Table 7: Example of summary statics for numerical variables after scaling, for *clean.csv* dataset.

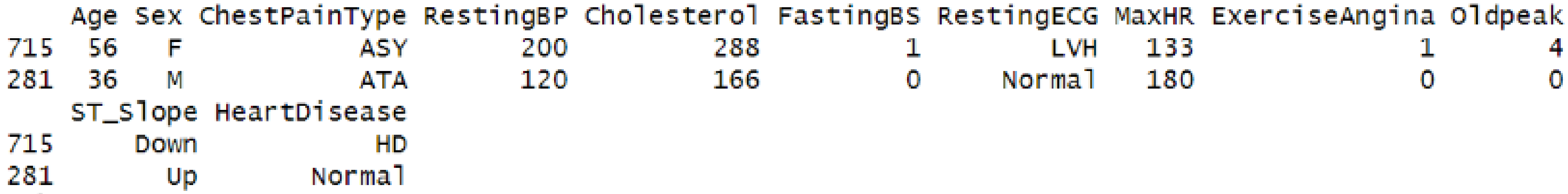
UNSUPERVISED LEARNING
Hierarchical Clustering

Since the dataset is made-up of mixed data types, I've used **Gower distance** to measure the pairwise distances across the distributions.



```
    Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak
157  42   M           ATA       120         196         0     Normal   150              0       0
141  42   M           ATA       120         198         0     Normal   155              0       0
    ST_Slope HeartDisease
157       Up       Normal
141       Up       Normal
```

Figure 18: Least dissimilar pair of observations from *clean.csv*.



```
    Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak
715  56   F           ASY       200         288         1        LVH   133              1       4
281  36   M           ATA       120         166         0     Normal   180              0       0
    ST_Slope HeartDisease
715     Down           HD
281       Up       Normal
```

Figure 19: Most dissimilar pair of observations from *clean.csv*.

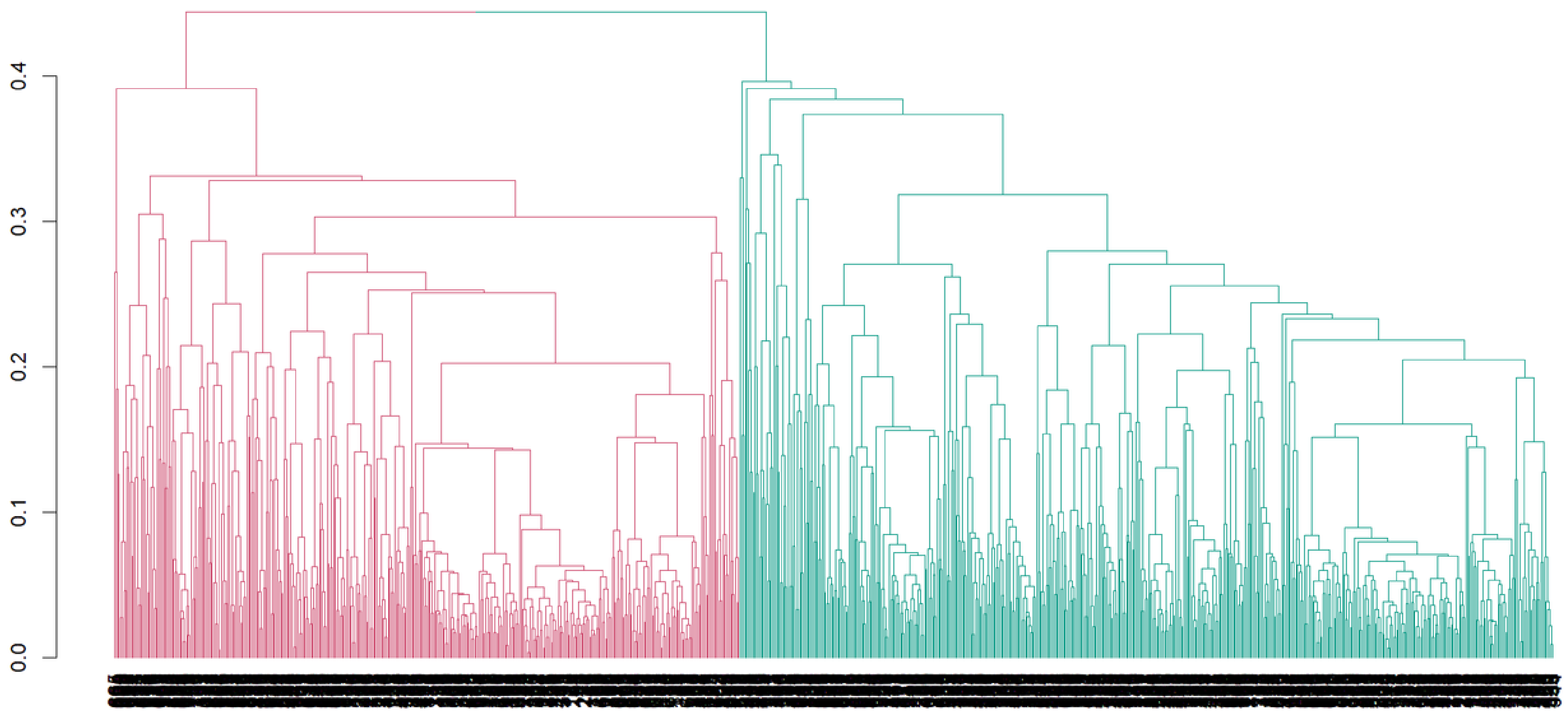
UNSUPERVISED LEARNING

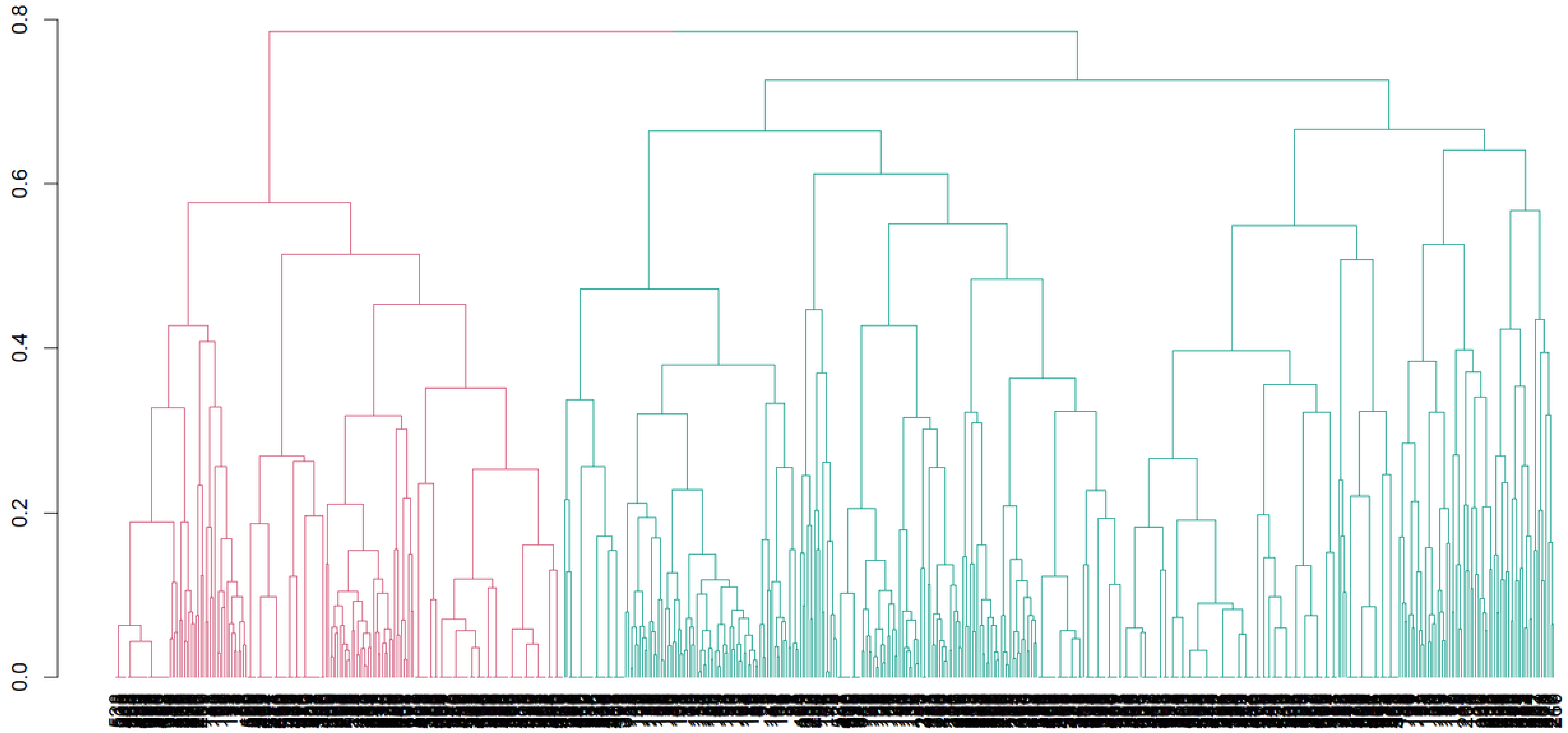
Hierarchical Clustering

I've applied hierarchical clustering on *clean* and *train* datasets.
The optimal number of clusters in both is **two**.

These are two examples of the clusters; on both the splits across the clusters have been determined by *average* linkage.

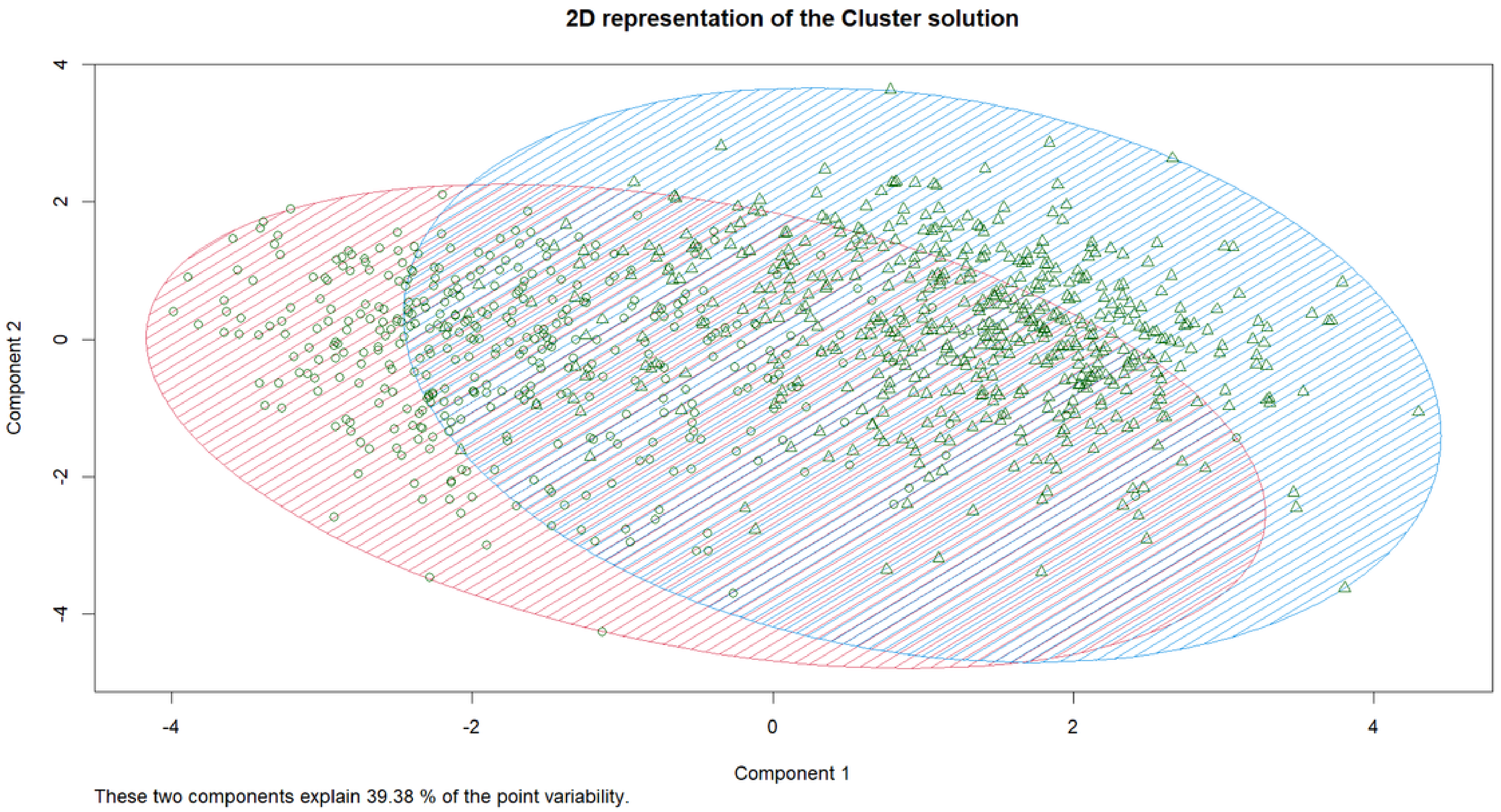
(1) is the dendogram built on the *clean* sample;
(2) is the dendogram built on the *train* sample.

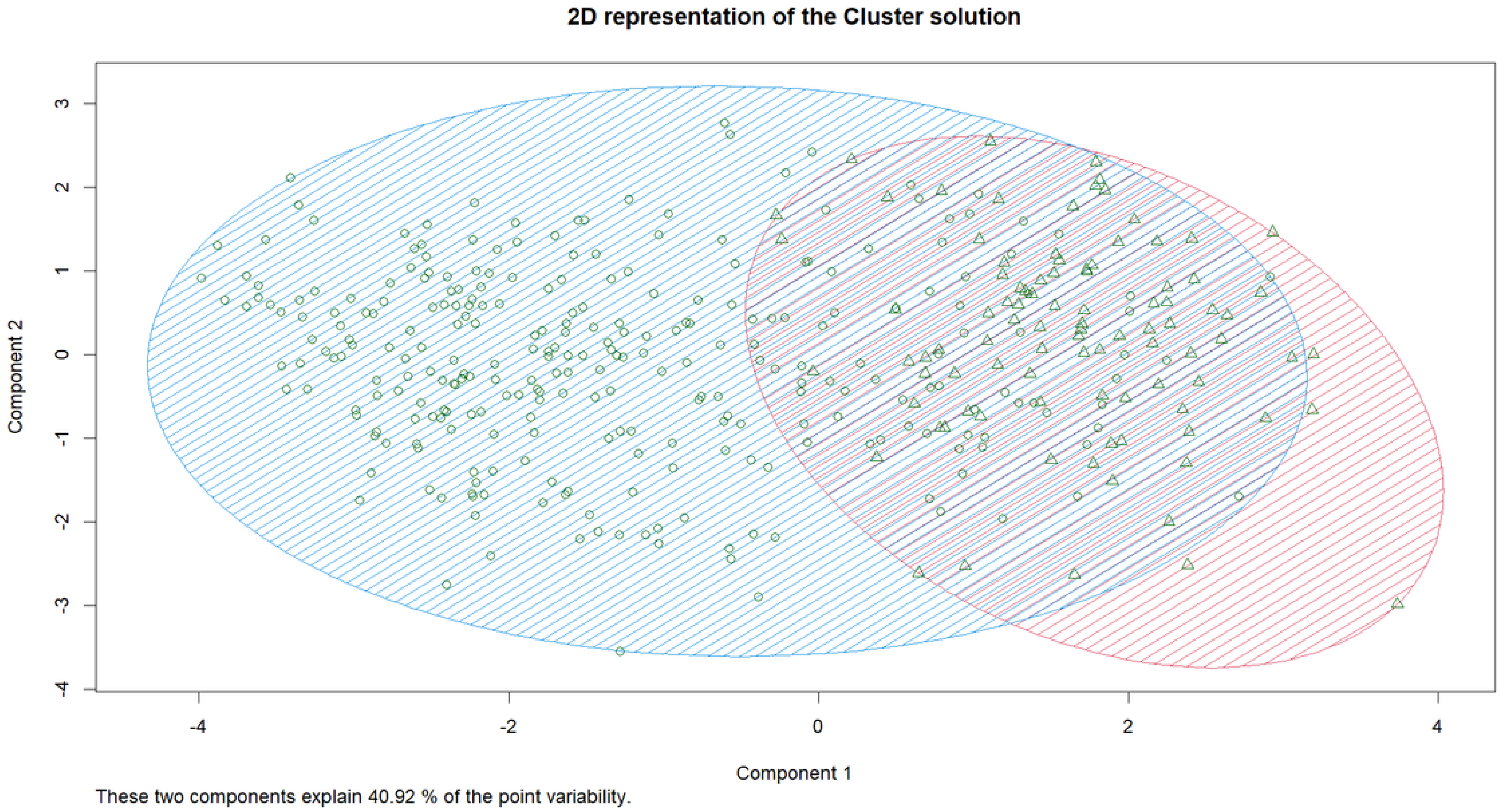




UNSUPERVISED LEARNING

Hierarchical Clustering - 2D Representation of the Clusters





(1) The clusters have been built on the *clean* dataset.

(2) The clusters have been built on the *train* dataset.

Thank you
for your attention!

Doina Vasilev
Università degli Studi di Milano