UNIVERSITY OF PADUA
UNIVERSITA' DEGLI STUDI DI PADOVA

Prediction of Coronary Artery Disease

Statistical Learning 2, A.Y. 2022/23

Marco Uderzo, 2096998
Francesco Vo, 2079413

Context and Motivation

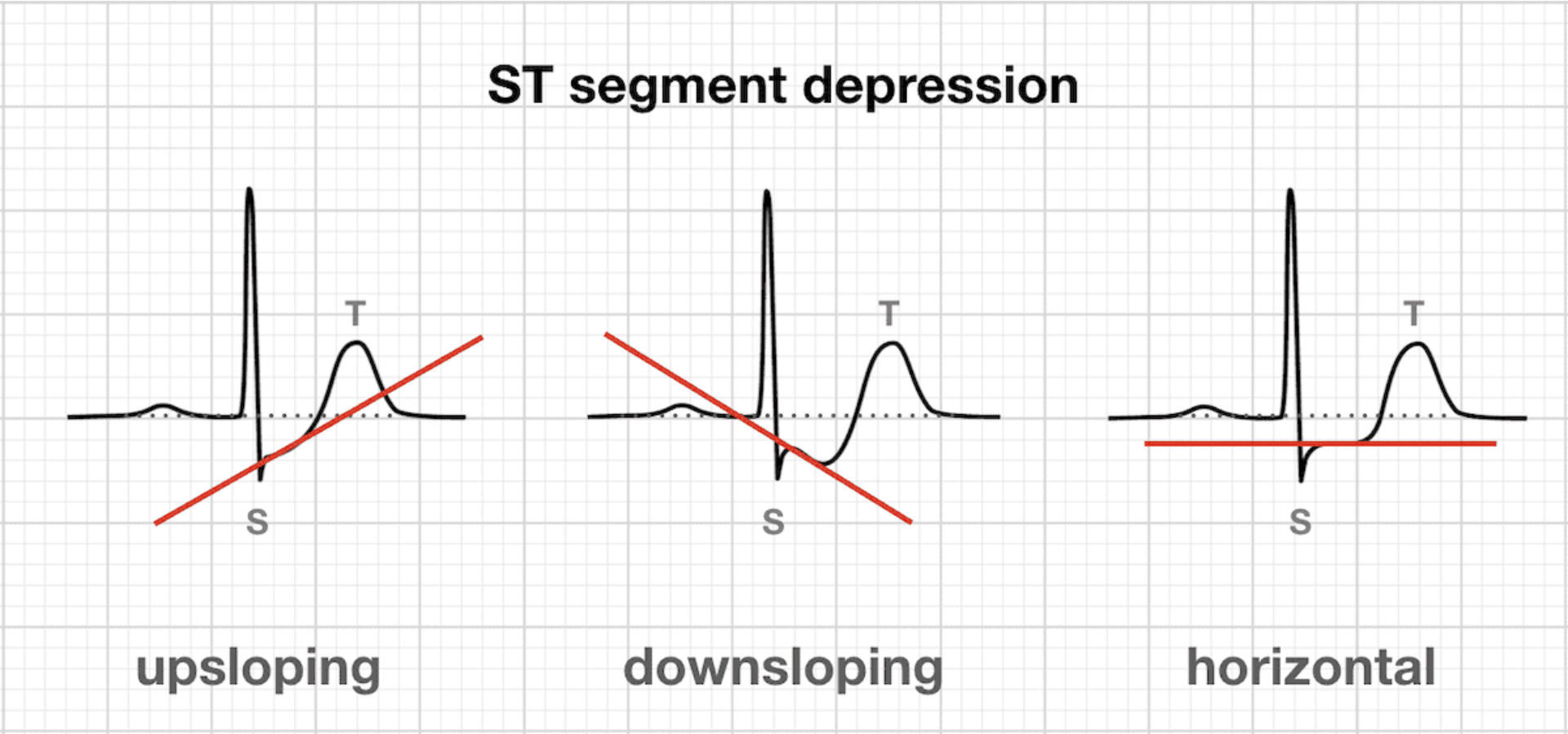
- Cardiovascular diseases (CVDs) are the number one cause of death globally.
- Coronary Artery Disease (CAD) is a common and very deadly occurrence, and this dataset contains 11 features that can be used to predict it.
- People with cardiovascular diseases or who are at high cardiovascular risk need early detection wherein statistical learning models can be of great help.

The Dataset

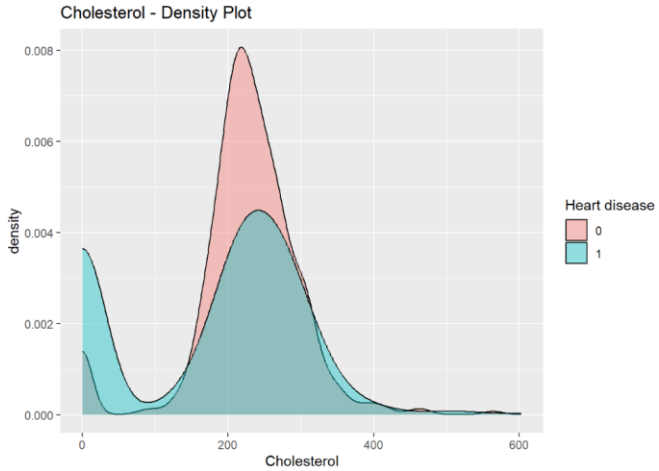
The dataset contains 12 parameters and 918 observations.

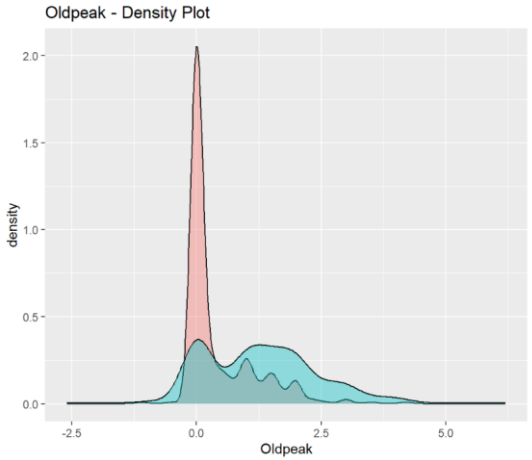
1. **Age**: age of the patient [years]
2. **Sex**: sex of the patient [M: male, F: female]
3. **ChestPainType**: chest pain type [TA: typical angina, ATA: atypical angina, NAP: non-anginal pain, ASY: asymptomatic]. Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary heart disease.
4. **RestingBP**: resting blood pressure [mmHg]
5. **Cholesterol**: serum cholesterol [mm/dl]
6. **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]. This measures the blood sugar level after an overnight fast.
7. **RestingECG**: resting electrocardiogram results [Normal: normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR**: maximum heart rate achieved [numeric value between 60 and 202]
9. **ExerciseAngina**: exercise-induced angina [Y: yes, N: no]
10. **Oldpeak**: oldpeak = ST [numeric value between -2.6 and 6.2]. ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline. Oldpeak measures the depression of the ST slope induced by exercise relative to rest.
11. **ST_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease**: output class [1: heart disease, 0: normal]

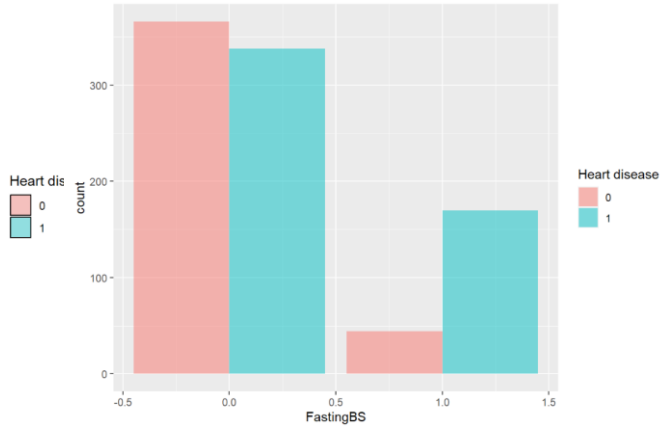
ST Slope

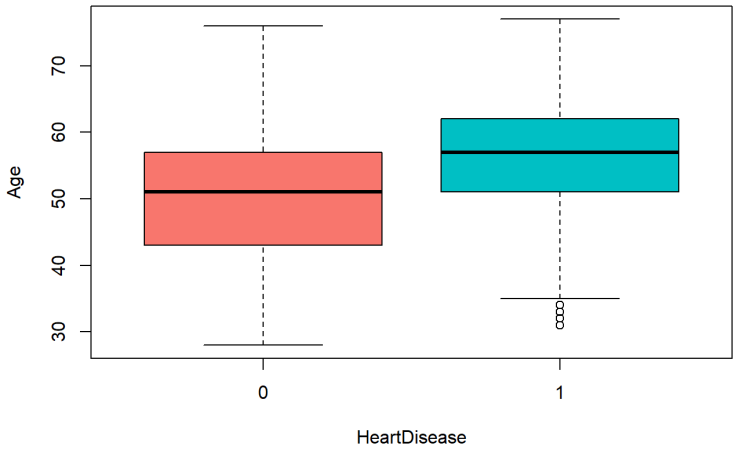

ST segment depression
T
S
T
S
T
S
upsloping
downsloping
horizontal

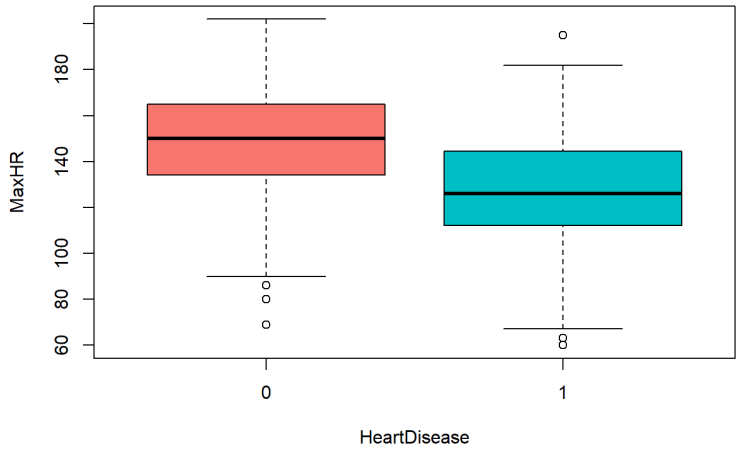
Data Visualization



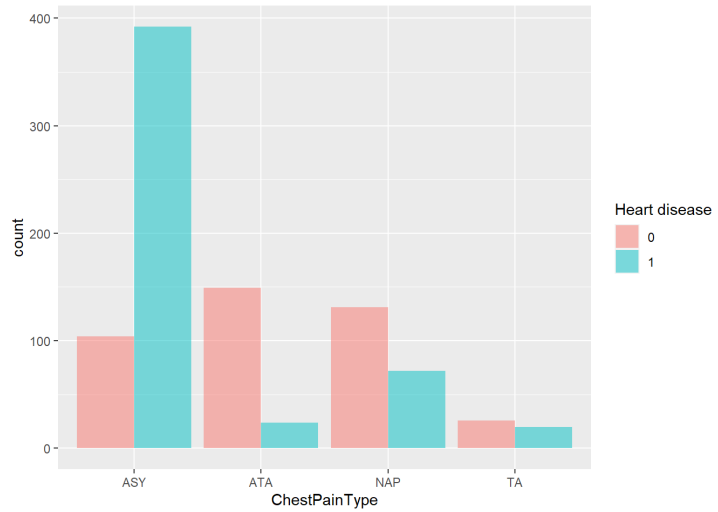


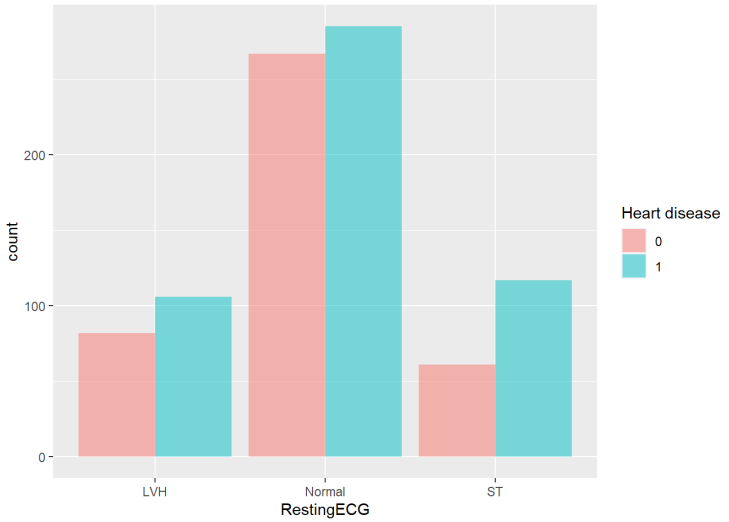


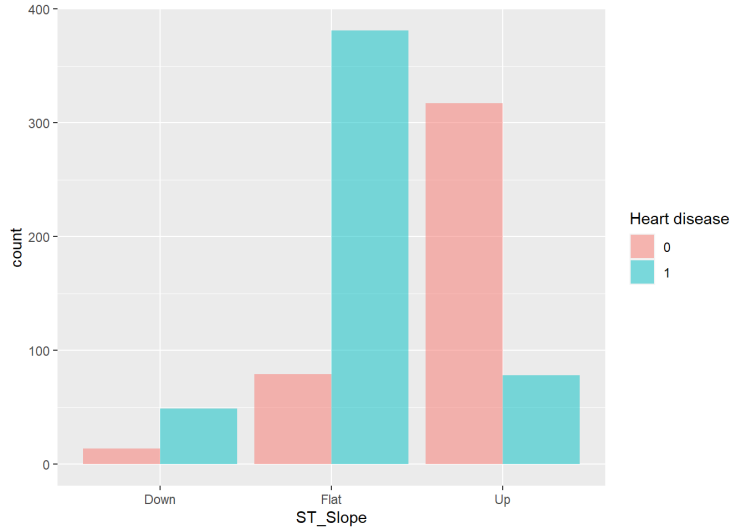


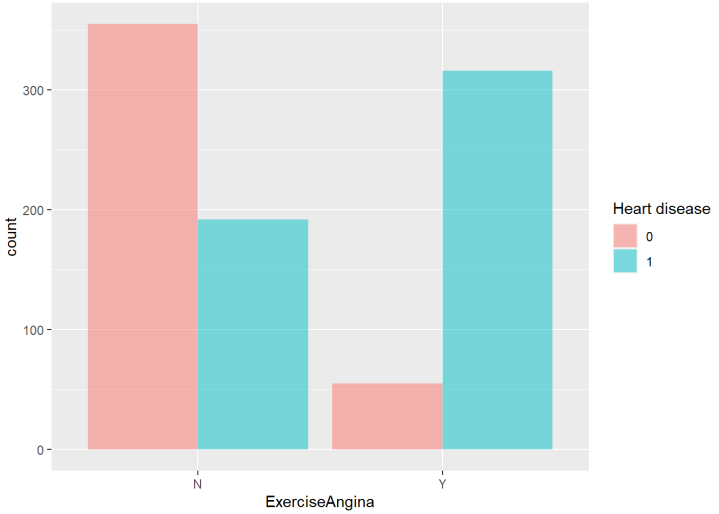


Data Visualization



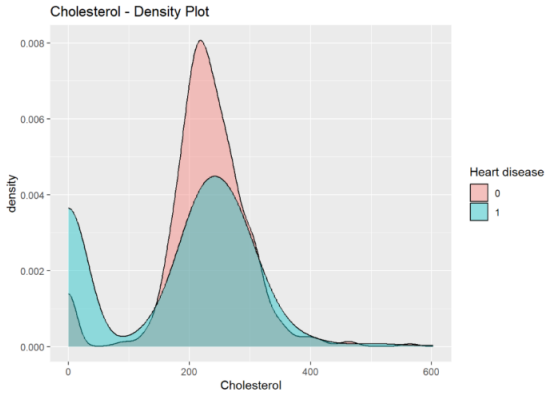






Missing Values & Outliers

- There are 172 variables in Cholesterol that have value equal to 0. Also we have noted that patients that have Cholesterol equal to 0 are very likely to have the heart disease.
    - One possibility is that the measurements were taken after the patient was dead, but if we inspect rows with missing data we can safely assume that these patients are alive and the Cholesterol value was incorrectly recorded.
    - Another guess is that the “serum cholesterol” measured by this variable combines the HDL and LDL values. High HDL and LDL would cancel each other out, making this variable less useful.
- We decided not to use this variable in the models and as we are going to see it doesn’t affect much the predictions.

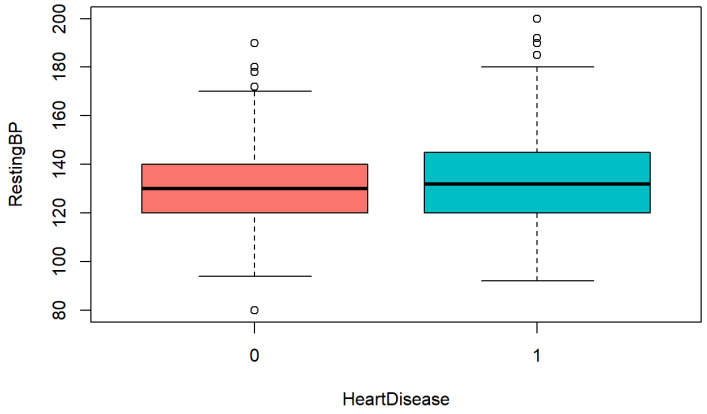


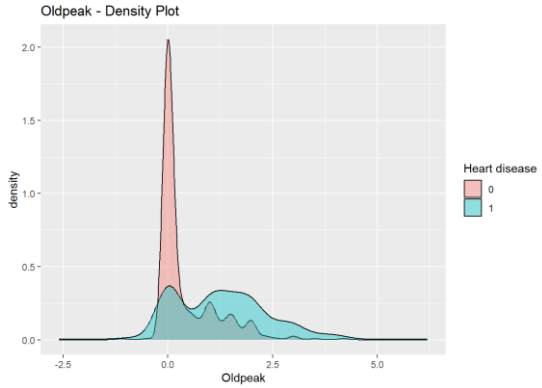
Missing Values & Outliers

Outliers in other continuous variables:

- **`RestingBP`**: outliers were still plausible values
- **`Oldpeak`**: the reason why there are many values as 0 is because Oldpeak refers to the difference between the depression of the ST segment in the ECG that has been measured during rest and then during exercise.

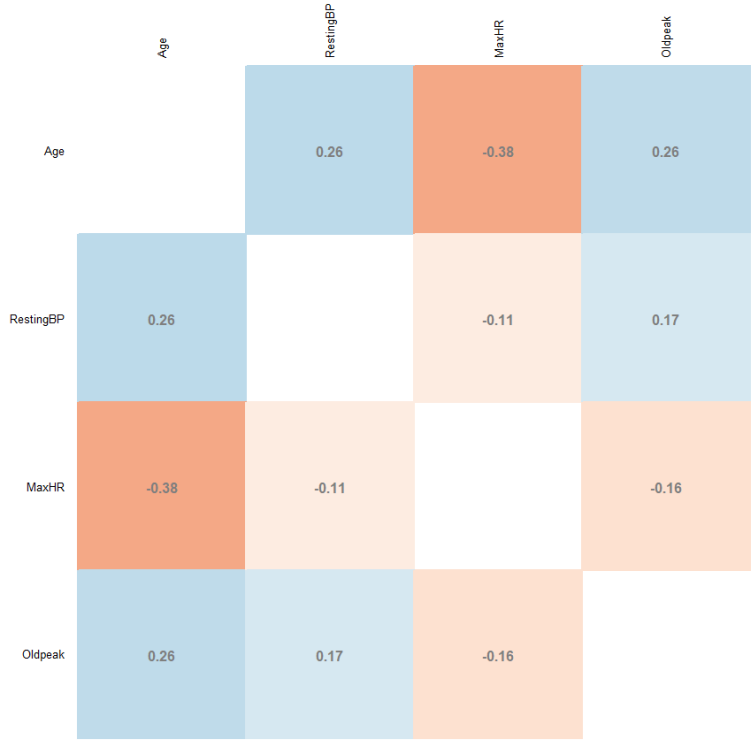
Values around 0 are correlated with no CAD.





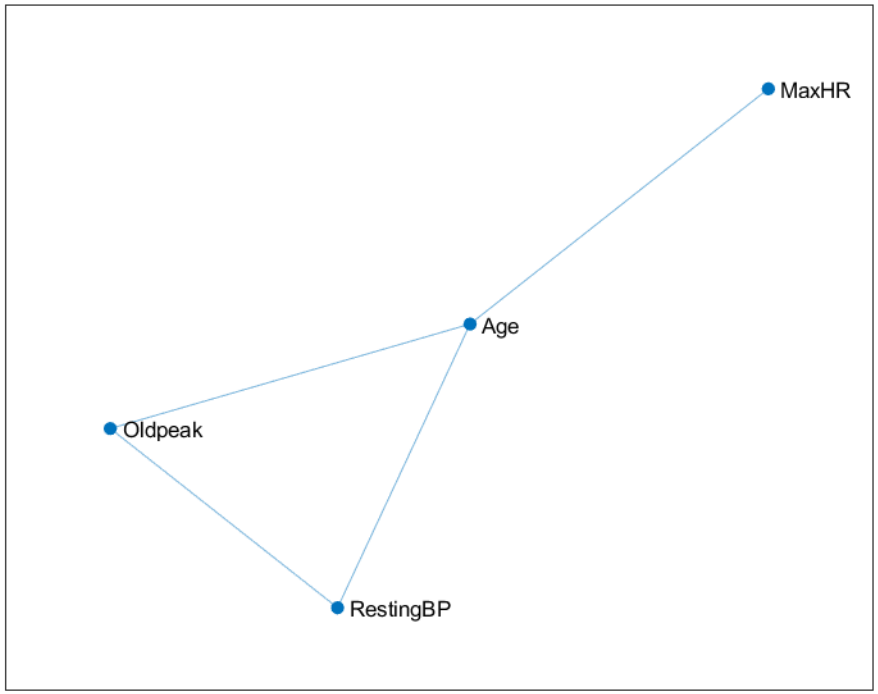
Correlations: matrix

We calculate the correlations for each continuous variable and we plot the correlation matrix.



Correlations: graph

Now we want to visualize the correlations through a graph. We see that Age is somewhat “central” with respect to the other variables.



Data Preparation & General Settings

We see that patients without the condition are 410 (45%) and patient with it are 508 (55%).

We see that our dataset is quite balanced so we don't need to apply other methods to rescale it.

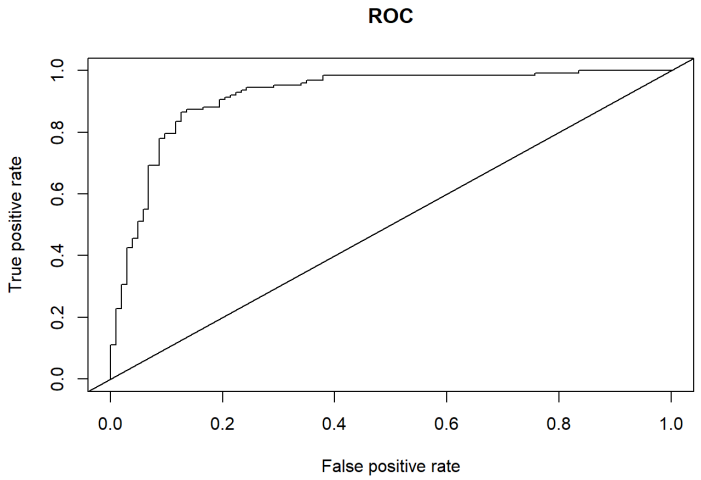
After this we divide the split the data into training set and test set with a ratio of 75 and 25 with respect to our dataset.

We will train our models on the training set and the evaluate them on our test set.

The metrics we are going to use are **Accuracy**, **Precision**, **Recall**, **F1-Score** and **FN Rate**.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classification algorithm, where a discriminant rule tries to divide the data points into K disjoint regions, where K is the number of classes (in this case K = 2).



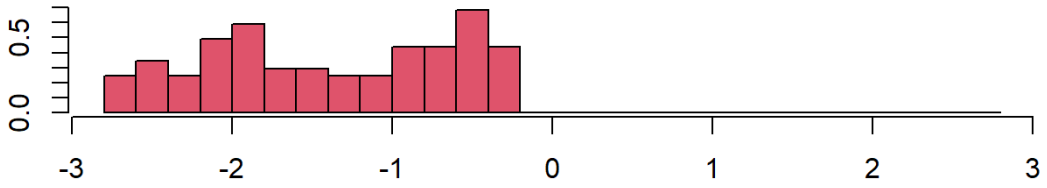
	LD1
Age	0.009930689
SexM	0.744352146
ChestPainTypeATA	-1.182848798
ChestPainTypeNAP	-1.075687933
ChestPainTypeTA	-0.660875986
RestingBP	-0.001194828
FastingBS	0.561717985
RestingECGNormal	0.001854152
RestingECGST	0.036571241
MaxHR	-0.005711702
ExerciseAnginaY	0.319998463
Oldpeak	0.217810362
ST_SlopeFlat	0.719835027
ST_SlopeUp	-0.760627954

Accuracy: 0.86 Recall: 0.87 AUC: 0.92

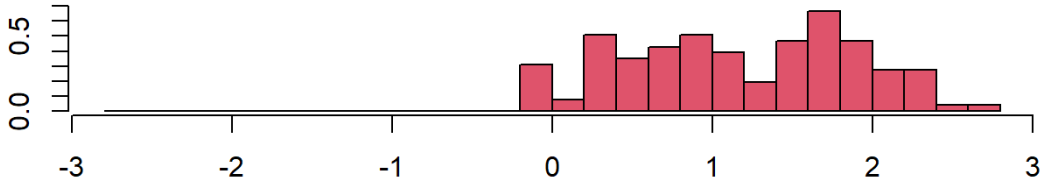
Precision: 0.88 F1-Score: 0.87 FN-Rate: 0.11

Linear Discriminant Analysis

We see that the models divides the data pretty well.



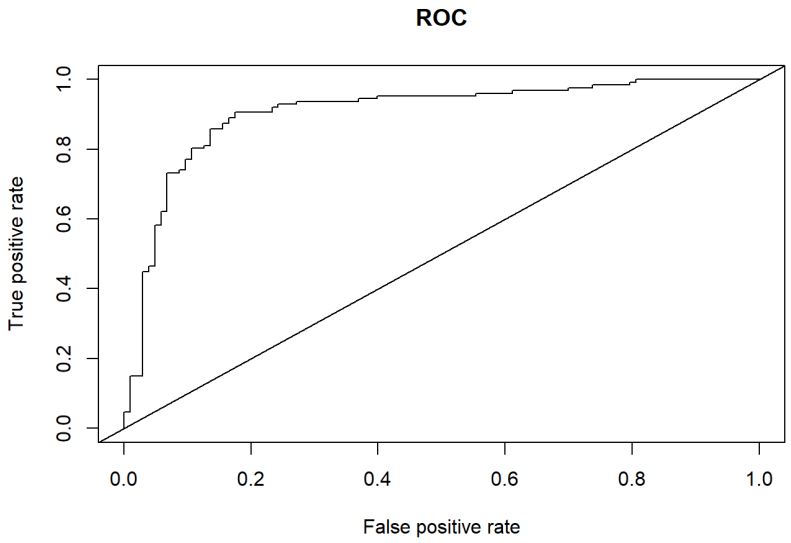
group 0



group 1

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) doesn't assume the equal variance of the classes. For this reason the decision boundary is not linear but quadratic.

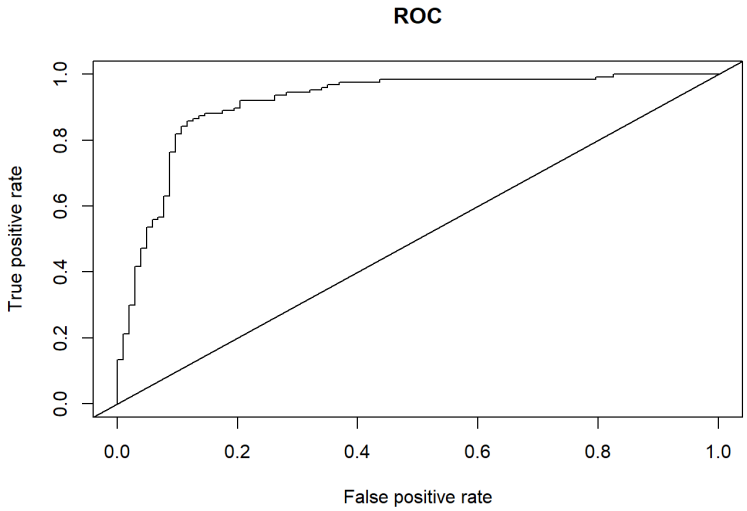


Accuracy: 0.86 Recall: 0.87 AUC: 0.90

Precision: 0.87 F1-Score: 0.87 FN-Rate: 0.13

GLM: Generalized Linear Model

- **Logistic Regression** is the easiest and most common model to perform binary classification. To do so, we use a **Generalized Linear Model**.
- All predictors are initially included
- Decision Threshold initially not optimized
- Variable Selection:
  - **Iterative Backward Selection**, removing weak predictors based on their p-value
- Useful Metrics:
  - **Variance Inflation Factor**
  - **Bayesian Information Criterion**



Accuracy: 0.86 F1-Score: 0.87

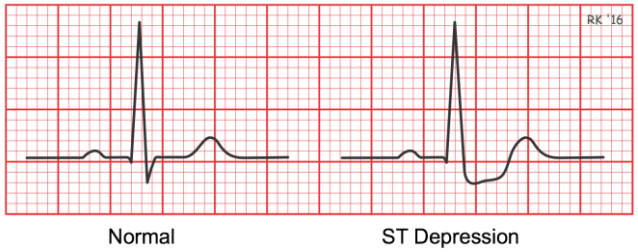
Precision: 0.9 AUC: 0.91

Recall: 0.84 FN-Rate: 0.09

GLM: Backward Selection with p-value

- Start with all predictors in the model and remove one at a time (highest p-value)
- Removed Predictors:
  - `Age, RestingECG, MaxHR, RestingBP`
- Best Predictors (p <= 0.001):
  - `Sex["M"], ChestPainType["ATA", "NAP"], FastingBS, Oldpeak`
- Considerations:
  - `RestingBP` only contains systolic BP, not considering diastolic and pulse pressure, so the variable is not very informative
  - `ST_Slope["Up"]` (upsloping ST segment depression) is a good predictor of CAD, so we kept it despite p-value > 0.05

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.5474     0.5988  -2.584 0.009766 **
SexM              1.6924     0.3074   5.505 3.69e-08 ***
ChestPainTypeATA -2.1879     0.3820  -5.727 1.02e-08 ***
ChestPainTypeNAP -1.8944     0.2929  -6.467 1.00e-10 ***
ChestPainTypeTA  -1.3656     0.4776  -2.859 0.004248 **
FastingBS         1.2972     0.2987   4.342 1.41e-05 ***
ExerciseAnginaY   0.6864     0.2685   2.557 0.010560 *
Oldpeak           0.4639     0.1305   3.556 0.000377 ***
ST_SlopeFlat      1.6074     0.4910   3.274 0.001062 **
ST_SlopeUp       -0.8304     0.5041  -1.647 0.099497 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



GLM: VIF & BIC

- The **Variance Inflation Factor** measures how much the variance of the estimated regression coefficient for a given independent variable is inflated due to multicollinearity.
- The **Bayesian Information Criterion** is a metric that is used to compare the goodness-of-fit of different regression models.

```
## GLM Model  0  -  BIC:  547.7341
## GLM Model  1  -  BIC:  542.4296
## GLM Model  2  -  BIC:  529.3928
## GLM Model  3  -  BIC:  528.7822
## GLM Model  4  -  BIC:  522.2917
```

```
## GLM Model  0  -  VIF:  2.103043
## GLM Model  1  -  VIF:  2.09731
## GLM Model  2  -  VIF:  2.097167
## GLM Model  3  -  VIF:  2.06998
## GLM Model  4  -  VIF:  2.069784
```

- Both VIF and BIC consistently dropped after removal of predictors with high p-value
- VIF indicates some minor correlation between predictors.
- Best GLM Model: `glm.model.4`

GLM: Analysis of Deviance

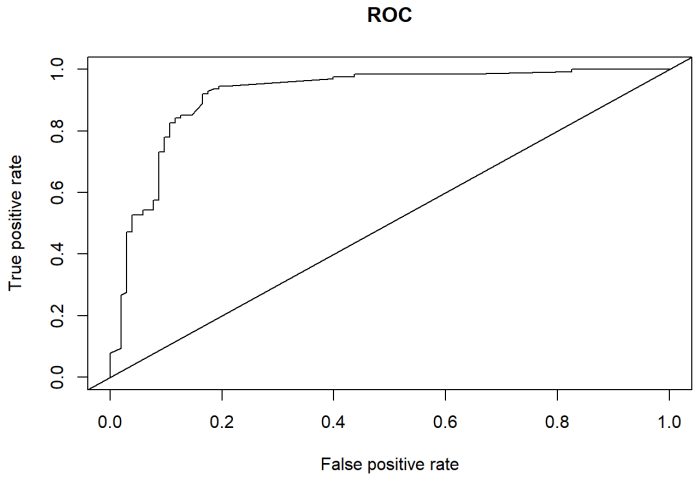
- **Deviance** is a quality-of-fit statistic for a model that is often used for statistical hypothesis testing.
  - Deviance difference test using ANOVA with χ^2 test yields a p-value of 0.2043
    - Reduced model is not significantly different from the full model
    - Removed predictors are indeed not useful

```
Analysis of Deviance Table

Model 1: HeartDisease ~ Sex + ChestPainType + FastingBS + ExerciseAngina +
    Oldpeak + ST_Slope
Model 2: HeartDisease ~ Age + Sex + ChestPainType + RestingBP + FastingBS +
    RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       678     456.95
2       673     449.73  5   7.2266   0.2043
```

GLM: Threshold Selection

- Objective: maximize Recall, in order to minimize wrongly discharged patients with CAD. At the same time, all other metrics must be acceptable.
- Thresholds: 0.3, 0.4, 0.5, 0.6
  - 0.3 Maximized Recall but Accuracy and Precision would suffer
  - Best Overall Threshold: 0.4



Accuracy: 0.88 | Recall: 0.92 | AUC: 0.92

Precision: 0.87 | F1-Score: 0.89 | FN-Rate: 0.13

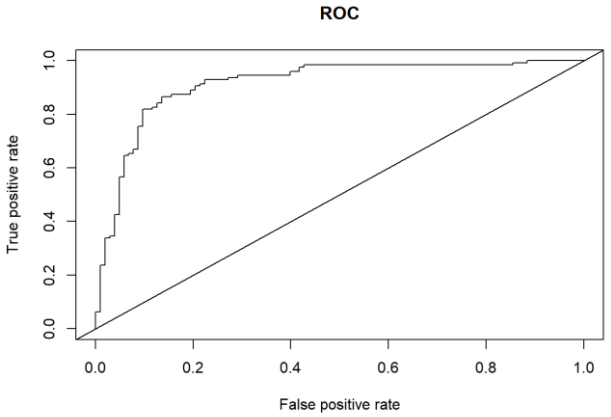
```
Threshold:  0.3
Accuracy, Precision, Rec, F1-Score 0.808695652173913, 0.754601226993865, 0.968503937007874, 0.848275862068965
Threshold:  0.4
Accuracy, Precision, Rec, F1-Score 0.878260869565217, 0.866666666666667, 0.921259842519685, 0.893129770992366
Threshold:  0.5
Accuracy, Precision, Rec, F1-Score 0.865217391304348, 0.869230769230769, 0.889763779527559, 0.879377431906615
Threshold:  0.6
Accuracy, Precision, Rec, F1-Score 0.847826086956522, 0.903508771929825, 0.811023622047244, 0.854771784232365
```

Regularized Regression

- Generalized models with some penalizations according to **λ** value
- Estimators with very large variants and small bias can lead to multicollinearity and produce poor estimates (overfitting)
- Regularized Regression methods use automatic selection of variables through shrinkage on the coefficients of the predictors:
  - **Ridge Regression**: quadratic shrinking, coefficients assume values very close to zero
  - **Lasso Regression**: absolute-value shrinking, coefficients assume values that can even be zero

Lasso Regression

- Cross-validation performed with `glmnet`
  - Best λ: ~ 0.036
- Shrunk Predictors:
  - `RestingECG["Normal", "ST"], RestingBP, ChestPainType["TA"]`
- Best Predictors (largest absolute magnitude):
  - `Sex["M"], ChestPainType["ATA, "NAP"], ST_Slope["Up", "Flat"]`



Accuracy: 0.85, Recall: 0.89 AUC: 0.91

Precision: 0.83 F1-Score: 0.86 FN-Rate: 0.17

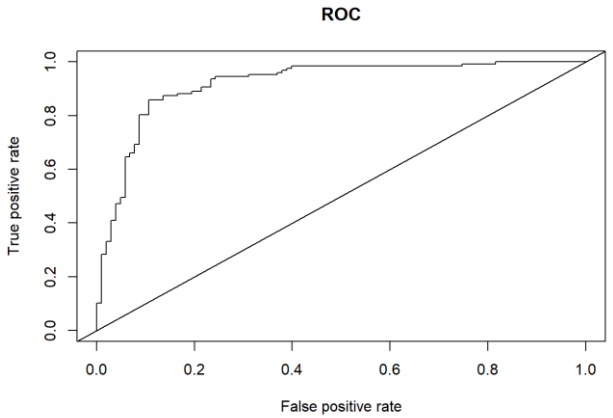
```
Selected variables

 Age
SexM
ChestPainTypeATA
ChestPainTypeNAP
FastingBS
MaxHR
ExerciseAnginaY
Oldpeak
ST_SlopeFlat
ST_SlopeUp
```

```
Age                0.003422359
SexM               0.906708935
ChestPainTypeATA  -1.230503452
ChestPainTypeNAP  -0.932959884
ChestPainTypeTA    .
RestingBP          .
FastingBS          0.572709220
RestingECGNormal   .
RestingECGST       .
MaxHR             -0.008990477
ExerciseAnginaY    0.537545127
Oldpeak            0.260167015
ST_SlopeFlat       0.755638448
ST_SlopeUp        -1.069356700
```

Ridge Regression

- Cross-validation performed with `glmnet`
  - Best λ: ~ 0.031
- Shrunk Predictors:
  - `RestingBP, RestingECG["Normal", "ST"], MaxHR, Oldpeak`
- Best Predictors (largest absolute magnitude):
  - `Sex["M"], ChestPainType["ATA, "NAP"], ST_Slope["Up", "Flat"]`



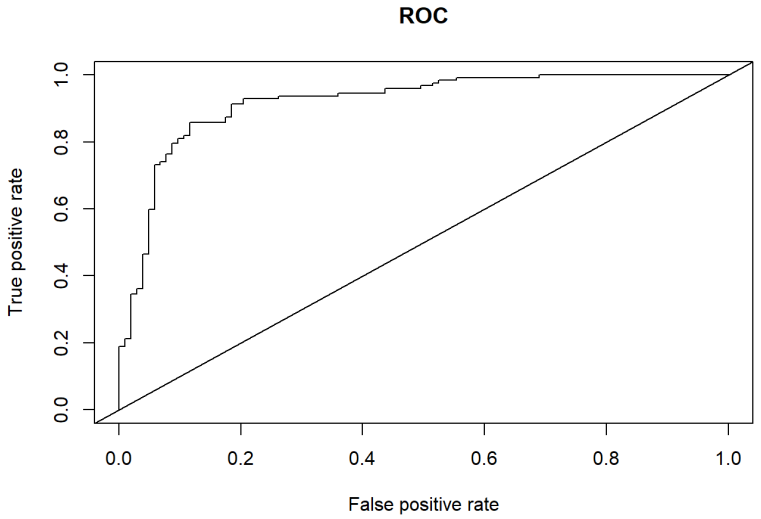
Accuracy: 0.87 Recall: 0.91 AUC: 0.92

Precision: 0.85 F1-Score: 0.88 FN-Rate: 0.15

```
Age               0.0160806001
SexM              1.1635775828
ChestPainTypeATA -1.4366194007
ChestPainTypeNAP -1.2222969859
ChestPainTypeTA  -0.7173331612
RestingBP        -0.0004252259
FastingBS         0.8555164979
RestingECGNormal -0.0134815370
RestingECGST      0.0403147074
MaxHR            -0.0107161566
ExerciseAnginaY   0.5963713906
Oldpeak           0.3521643720
ST_SlopeFlat      1.0208749799
ST_SlopeUp       -0.9436011441
```

Naive Bayes Classifier

- The **Naive Bayes Classifier** is a classification technique based on Bayes' Theorem with an independence assumption among predictors.
- Does not perform explicit feature selection
- But assigns higher importance to features that have a stronger influence on class probabilities
- Best threshold: 0.6

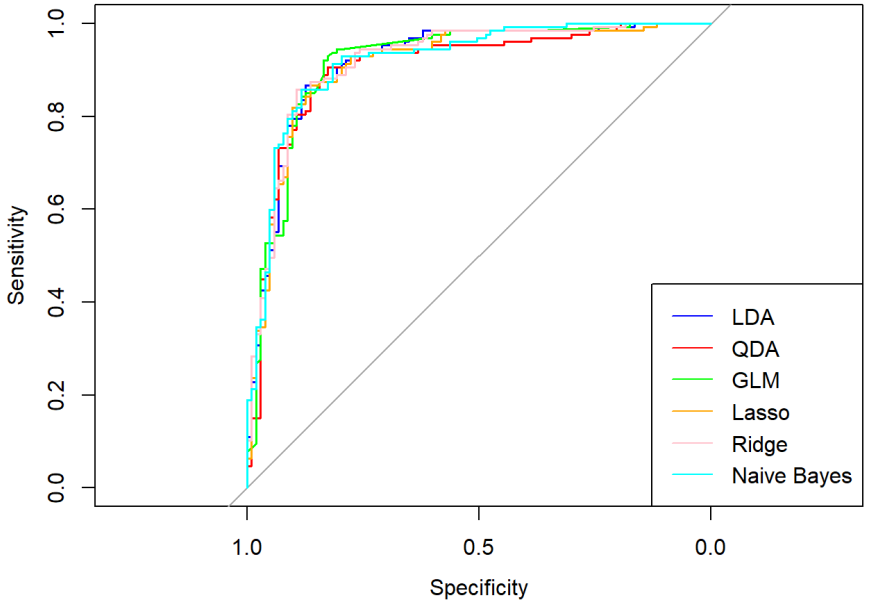


Accuracy: 0.87 Recall: 0.91 AUC: 0.92

Precision: 0.85 F1-Score: 0.88 FN-Rate: 0.15

Conclusions: Best Models

- All models perform relatively good, with similar ROCs and each with their own strengths and weaknesses in their metrics.
- Considering Recall as most important, the best models are:
  - GLM: Logistic Regression (threshold: 0.4)
  - LDA
  - GLM with Ridge



Model <chr>	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1 Score <dbl>	FN Rate <dbl>	AUC <dbl>
LDA	0.8652174	0.8870968	0.8661417	0.8764940	0.1129032	0.9184313
QDA	0.8565217	0.8740157	0.8671875	0.8705882	0.1259843	0.9049767
Logistic Regression	0.8782609	0.8666667	0.9212598	0.8931298	0.1333333	0.9175140
Lasso Regression	0.8478261	0.8267717	0.8898305	0.8571429	0.1732283	0.9130036
Ridge Regression	0.8695652	0.8503937	0.9075630	0.8780488	0.1496063	0.9192722
Naive Bayes	0.8652174	0.8503937	0.9000000	0.8744939	0.1496063	0.9195016

➞ **GLM with Logistic Regression** can predict CAD with 88% accuracy and 92% recall.

➞ **LDA** minimizes False Negatives

Conclusions: Best Predictors

- Medical literature was also briefly checked to confirm our findings
- Considering all models, the very best predictors are:
  - **`Sex["M"]`**: males are more likely to suffer from CAD.
  - **`ChestPainType["ATA", "NAP"]`**: Atypical Angina is a known symptom preceding Myocardial Infarction. Non-Anginal Pain is often non-ischemic related.
  - **`ST_Slope["Up", "Flat"]`**: ST Segment Depression is a known anomaly of the electrophysiology of heart due to coronary blockage and therefore myocardial ischemia.
  - **`Oldpeak`**: variations in ST Segment Depression induced by exercise relative to rest also indicate myocardial ischemia.
  - **Age**: central predictor in the correlation graph, although no model kept it.