Patients Classification to Categories of Heart Failure Risk

Empirical comparison of classification algorithms and best model selection

Alessandro Mirone - 966880

2022-09-01

UNIVERSITÀ DEGLI STUDI DI MILANO

Abstract

The following analysis focus on finding the best statistical model for prediction of risk classes, as well as identifying the ones capable of interpreting the relevance of features in explaining the risk of heart failures. In the process, four methods are compared and their performances over these two tasks are evaluated through analytical considerations and empirical results. In order to find the correct model specification, an initial analysis of the data set's variables is carried out; to provide more information about the data hierarchical clustering is also performed, a method which is capable of giving valuable insights about what combinations of feature are most likely encountered in the population, and what of these combinations are most likely to rise -or lower- the risk of Heart failure. Results show that, once the correct model is specified, Linear Discriminant Analysis and Logistic Regression with quadratic terms tie at 87% accuracy as the best two models given the data and the task at hand.

1- Data description

The data set consists in 302 observations of 14 dimensions, one of which is the binary outcome representing the risk of heart failure. Data were obtained from the University of California, collected in 1988 in Cleveland; it is a popular data set for evaluating machine learning algorithms[1] and it has been used in many publications throughout the years. The following is a brief description of all used variables and their coding, in case of categorical features.

- **Age** : the age of the patient.

[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

- **Sex** : the respondent's gender.
- **Cp** : type of anginal pain recorded; categorical (0 = asymptomatic; 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain).
- **Trstbps** : resting blood pressure (in mm Hg on admission to the hospital).
- **Chol** : serum cholesterol in mg/dl.
- **Fbs** : fasting blood sugar greater than 120 mg/dl; categorical (1 = true; 0 = false). Fasting blood sugar levels between 110 and 125 mg/dl are considered prediabetes.
- **Restecg** : resting electrocardiographic results; categorical (0 = normal; 1 = having ST-T wave abnormality; 2 = hypertrophy).
- **Thalachh** : maximum heart rate achieved.
- **Exng** : whether phisycal exercise induced angina; categorical (1 = yes; 0 = no). By exercise is meant patient's own physical activity, not a stress test.
- **Oldpeak** : ST wave depression depth induced by exercise relative to rest, measured in mm. Greater wave's dives signal heart anomalies.
- **Slp** : slope of ST wave during stress ecg test; categorical (0 = downsloping; 1 = flat; 2 = upsloping). Curves with a positive slope signal greater stress of the heart.
- **Caa** : number of major vessels (0-3) colored by flouroscopy. Obstructed vessels are not colored by the test, thus 3 is the optimal situation, and 0 is the worst case scenario. This variable can be considered either as numerical or as ordinal; this distinction doesn't have an impact on the evaluation of the methods, and is therefore meaningless. One can instead consider this aspect to tailor the preliminary analysis regarding this feature. In this case, it simply means that it will be performed a test of correlation like for other numerical variables, but it won't be presented a boxplot; instead, findings are summarized with the help of a contingency table, like for categorical variables.
- **Thall** : coronary status as revealed by the stress test; categorical (1 = fixed defect; 2 = normal; 3 = reversable defect). "Fixed defect" in the context of this test means that myocardial scarring, caused by previous trauma, is present in the patient. "Reversable defect" is instead referring to coronary stenosis, that is the obstruction of arteries, most likely caused by accumulation of fat tissue in the coronaries.
- **Output** : the predicted attribute; diagnosis of heart disease (angiographic disease status); categorical (0 = less then 50% coronary's diameter narrowing, identify lower risk of heart failures; 1 = greater than 50% coronary's diameter narrowing, signal higher risk of heart failures)

The objective of the analysis is to build a classifier capable of assigning each patient to the correct level of risk, starting from these thirteen predictors; the utility is straightforward: a patient with a higher risk of heart failure should be prioritized with respect to others in a less risky situation. Another aspect of interest is to gain insights about what factors are most relevant in determining such higher risks: this evaluation can be used for prevention, or again as evidence pointing to the need for particular attention in regards of patients that present some of the aspects typically related to high risk class at the moment of the hospitalization, in absence of all the information needed to apply the classifier before specific tests, like the stress test or ecg, are performed. To obtain such insights, along with parametric models capable of estimating the relative importance of each feature, a hierarchical clustering algorithm is carried out and its results outlined in section 1.1.

The first five rows of the (unscaled) data set appear as such:

Table 1: initial five examples from the data set

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

The next step is the preliminary evaluation of the data set's variables; for the six numerical variables (including **ca**) Pearson's ρ correlation coefficient is computed.

Table 2: numerical variable correlation with output

	x
age	-0.2307196
trtbps	-0.1424659
chol	-0.1111474
thalachh	0.4210961
oldpeak	-0.4329272
caa	-0.4638856
output	1.0000000

these results can be interpreted as such:

- **Age** is significantly correlated and inversely proportional to output; given that the sample doesn't include a control group (that is, all patients suffer from heart conditions; some have a more severe risk and some do not) it's possible that older patients with more severe symptoms are not present in larger numbers because a death event has occurred
- **Resting blood pressure** has lower correlation in the sample, and a negative dependence with the output. As hearts with a more severe disease can well be compatible with lower blood pressure, as the blood flow of a tired heart is reduced, this result is consistent with the literature.
- **Cholesterol**, as resting blood pressure, has low correlation; it's possible that the dependence is not significant in the sample. It's the first candidate for omission in model specification.
- **Maximum heart rate** is strongly correlated with the output, and positively so, as per typical situations of heart malfunction.
- The severeness of the **depression in the st curve** is strongly and negatively correlated with output; this is quite puzzling, as the real situation is exactly the opposite. We will return on this issue in the conclusions, but the evidence supports the possibility of an error in coding the variable in the data set[2]
- finally, the **number of major vessels**, i.e. how many of the major arteries are not obstructed, is strongly and negatively correlated with the risk of severe heart condition, consistently with medical studies.

The results can be graphically summarized with a boxplot (caa is excluded), presented in *figure 1.*

[2]This dataset is not the original one; it was obtained from the website Kaggle and thus manipulated by other users. A coding error is quite possible.

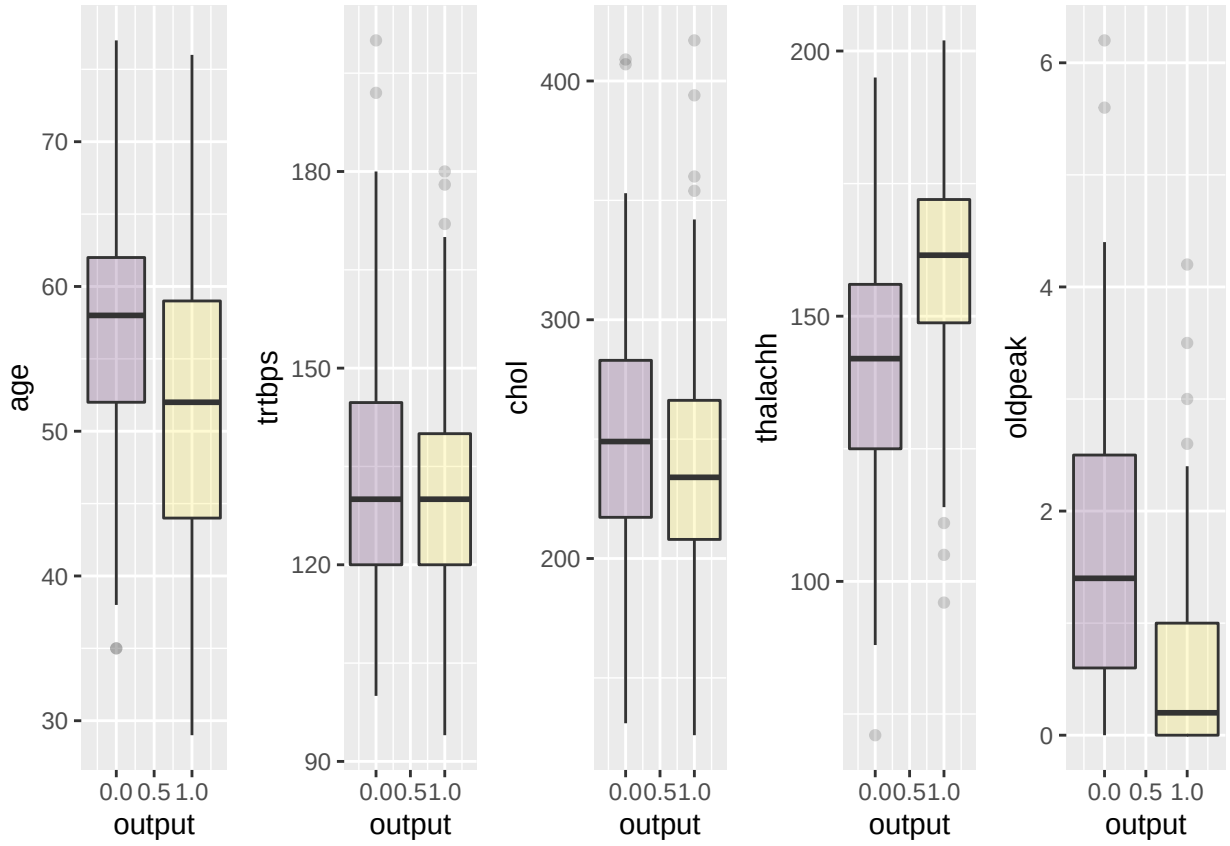


Figure 1: Figure 1

For the remaining seven categorical variables in the dataset (plus caa), it's possible to look at contingency tables and χ^2 - values to retrieve information about the strength and nature of their dependence with output[3]:

Table 3: fasting blood sugar

	0	1
0	116	22
1	141	23

Table 4: exercise induced angina

	0	1
0	62	76
1	141	23

Table 5: resting ecg results

	0	1	2
0	79	56	3
1	67	96	1

Table 6: chi squared test for fbs, exng, ecg

Chi-squared	p-value
0.0924084	0.7611375
55.4562030	0.0000000
10.3509272	0.0056536

- **Fasting blood sugar** appears independent with respect to output. Indeed, the χ^2 - value is low and the relative p-value is not significant.
- **Exercise-induced angina** shows instead strong dependence with the output. Regarding this variable, it should be noted that "no chest pain induced by exercise" doesn't mean that the patient did not experience chest pain; rather that angina wasn't caused by exercise. Thus the inverse relation with output that can be seen in the contingency table can be interpreted accordingly.
- Worse **Resting ecg results** are unsurprisingly positively related to output; the p-value relative to the small χ^2 of the class "normal" is not significant, while when it comes to the class 1 (ST-T wave abnormality) the results show a significant dependence with output.

Table 7: gender

	0	1
0	24	114
1	71	93

Table 8: chest pain type

	0	1	2	3
0	104	9	18	7
1	39	41	68	16

Table 9: slope of the ST-T curve

	0	1	2
0	12	91	35
1	9	48	107

Table 10: chi squared test for sex, cp, slp

Chi-squared	p-value
22.13168	2.5e-06
80.97876	0.0e+00
48.35779	0.0e+00

- From the contingency table, incidence of high risk in women seems statistically significant. This is indeed confirmed by the independence testing : the output dependence with respect to **sex** is statistically significant, but only because the incidence of higher risk in women is much higher than in

[3]to lighten the presentation of the results, not all statistics are presented in the paper. The code for reproducing those results which are not in the χ^2 tables is attached in the appendix.

men. Holding risk constant, the proportion of women experiencing high risk of heart failures is much higher than the one of men (75% of women vs 44.9% of men). This is coherent with the literature; as the sample mean for age is 54, most of the women in the sample do not produce estrogen hormones anymore, which are a natural defense against heart conditions.

- **Chest pain** appears strongly (positively) correlated with higher risk. Note how anginal and asymptomatic are more strongly correlated to the response with respect to non anginal pain. In particular atypical anginal seems a little more indicative, while non anginal χ^2 - value is not significant at the 5% significance level
- The **slope of the ST-T curve** is another predictor strongly related to the outcome. From the contingency table it is evident how upsloping curves, signifying cardiac stress, are three times more common in the high risk class then in the low risk one.

Table 11: number of major vessels

	0	1	2	3
0	46	44	31	17
1	133	21	7	3

Table 12: coronary status

	1	2	3
0	12	36	90
1	6	131	27

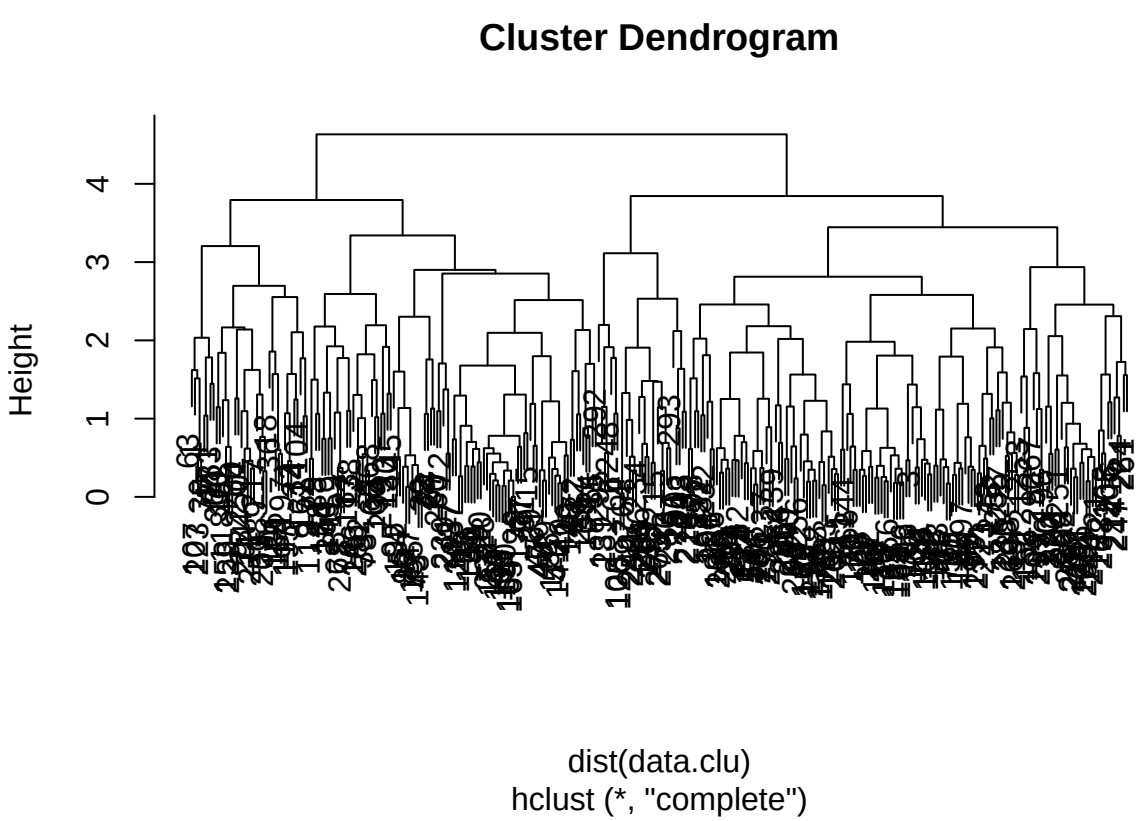
Table 13: chi squared test for caa, thall

Chi-squared	p-value
73.68904	0
88.38166	0

- The previous consideration regarding the **number of non-obstructed major vessel** are confirmed through the analysis of its relative contingency table.
- Finally, the situation for the **coronary status** appears mostly odd. fixed defect (1) appears uncorrelated. Reversible defect(3) corresponds to low level of risk, while normal (2) corresponds to higher risk; this result is as strange as the one that was observed with the depth of the ST-T curve dive. Since the coding of this variable was modified with respect to the original data set, it is likely that this situation arose from an error in the construction of the data set used.

In conclusion, output appears correlated with most of the variables in the data set, albeit in some strange ways regarding coronary status and depth of ST-T curve depression. Three variables are candidates for exclusion from the model: fbs, chol, and trtbps.

1.1 - Clustering A hierarchical clustering algorithm with complete linkage is performed in order to divide the observations into similarity groups. Numerical variables are scaled and the output is removed before performing the analysis: in this way, the resulting groups will be aggregated based only on the predictors. Group membership is then assigned to the observations and the distribution of the output variable in the different clusters is examined.



The dendrogram suggests the presence of 8 sufficiently large clusters (the cut is performed at height = 3) . Among the resulting groups, we can focus our attention on four in particular.

```
##         output
## cluster  0  1
##       1  3  5
##       2 13 17
##       3 55 52
##       4  4  3
##       5  6 21
##       6  8 58
##       7 18  4
##       8 31  4
```

cluster 5, 6, 7 and 8 split the outcome in two distinct groups; by examining the mean vectors of each group it's possible to infer what factors are mostly associated with higher (groups 5 and 6) and lower risk (groups 7 and 8).

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	membership
54.370	0.815	1.370	133.481	240.593	0.333	0.370	160.889	0.111	0.311	1.963	0.593	2.481	5
52.091	0.515	2.136	129.909	241.955	0.121	0.742	156.758	0.167	0.605	1.742	0.318	2.076	6
58.091	0.682	0.000	134.182	236.727	0.136	0.182	130.682	0.818	1.427	1.045	1.091	1.682	7
57.486	0.486	0.000	139.343	266.571	0.171	0.714	128.714	0.743	2.346	0.771	1.171	2.829	8

While categorical variables' distribution can be examined by looking directly at the data (see the appendix), from the result is mostly evident that the presence of chest pain, either typical or atypical angina, is a in

itself a strong classifier for assessing the risk of heart failure. Corollary to this finding is the distinction between justified chest pain (induced by stress) and unsolicited angina. Slope and depth of the ST-T curve are also strong indicators (once accounted for the usual consideration about the actual sign of the correlation between the depth of the depression of the ST-T curve and the risk of heart failures). High heart rates are also associated with higher risk; any heartbeat above 130 bpm at rest is considered signal of high cardiac stress. While age doesn't seem to be a particularly relevant factor alone, it becomes relevant when considered in pair with gender: as argued above, older women are at much more risk than older men. From the distribution of sex in the four groups, it can be noted how the higher risk clusters (in particular group 5) have a much more higher percentage of women with respect to the population in the sample.

Having sufficiently explored the relation among the data and the output, the next section will focus on presenting the models chosen for classifying patients to risk categories.

2 - Models

This section offers a brief presentation of the two classes of parametric models that will be used in the rest of the analysis. Although both types of models are used for classification, they differ in many aspects and is interesting to examine their mathematical formulations in order to better understand how they work. All models are presented in their multivariate specification.

2.1 - Logistic Regression Logistic regression is a discriminative method which uses a linear function of X to predict a binary response(in the binomial case); the method works by using the logarithm of the logistic function expressed as the odds, that is $\frac{p(X)}{1-p(X)}$, to model the probability that the response Y belongs to a certain category:

$$log\frac{(p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{1}$$

with $p(X) = Pr(Y = 1|X)$ and where $X = (X_1, ..., X_p)$ are the predictors. Equation 1 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{2}$$

through the inverse logit transformation: in this way, the log-odds (equation 1) can be rewritten as probabilities. the coefficients in equation 2 are estimated through classic MLE approach.Although the log-odds function in equation 1 is linear in X, the logistic function (equation 2) is not; this is why the logistic function has the shape of an s and can model probability for binary response in a [0,1] interval. This also means that β_p does not correspond to the change in $p(X)$ associated with a one unit increase in X_p. The amount that $p(X)$ changes due to a one-unit change in X_p depends on the current value of X. Regardless, if β_p is positive, then increasing X will be associated with increasing $p(X)$.

2.2 - LDA & QDA Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are two examples of generative models; they use a Bayesian approach to classification, in the sense that they estimate the posterior density, conditional on parameters estimated from the observations initially assigned to each class based on a prior distribution (which in this case is just the proportion of observations in the two classes), of the class k for all possible assignments of the observations, maximizing the posterior likelihood and then finally computing the probability that observation y was generated by the density of the k th class. In formula:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{3}$$

where $Pr(Y = k|X = x)$ is the posterior probability that an observation $X = x$ belongs to the kth class; $f_k(X) \equiv Pr(X|Y = k)$ is the density function of X for an observation that comes from the kth class and π_k is the prior distribution for class k. In both LDA and QDA, $f_k(x)$ is assumed to be a normal density with unknown parameters μ_k and Σ (for the LDA, which assumes a common variance-covariance matrix for all classes) or Σ_k (for the QDA, that instead allows Σ to vary between classes). By substituting the multivaraite gaussian density for $f_k(x)$ in equation 3, the LDA compute the posterior density as the product between the prior and the likelihood, obtaining close form estimates for μ_k and Σ and assigning the observation $X = x$ to the class for which the posterior probability

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma{-}1\mu_k + log\pi_k \tag{4}$$

is largest. The QDA works exactly the same but, given that the variance-covariance matrix varies between groups, equation 4 becomes

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}log|\Sigma_k| + log\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}log|\Sigma_k| + log\pi_k\end{aligned} \tag{5}$$

Note that equation 5 is non linear in x, as opposed to equation 4; hence the names "Linear" and "Quadratic" Discriminant Analysis.

3 - Empirical comparison and model selection

Once verified the assumptions for the statistical methods chosen, their performances are compared in this section. Several models[4] were fitted for each parametric method, their specification chosen through backward selection. In subsection 3.1 two models are presented for each method: the starting point, i.e. models containing all the predictions and no transformations, and the end model, that is the specification after which accuracy of the trained classifier and significance of the coefficients start to decrease. In order to examine further the role of variability in the performance of the classifier on this particular data set, in subsection 3.2 two final models are fitted through K-nearest neighbor algorithm, a non-parametric method whose degree of variability can be adjusted through the choice of k, performed in this case via 10-fold cross-validation. Such non-parametric approach is obviously not capable of giving an interpretation of the coefficients, thus missing one of the three focal points of this analysis, but was trained and tested nonetheless to compare its performances with the parametric methods and discuss the bias-variance trade off. Finally, subsection 3.3 pertain to model selection, where the performance of the chosen model and its accuracy are compared to that of the other candidates to explain why it was chosen.

3.1 - Parametric models Before fitting the logistic regression model, a Box-Tidwell test is performed to verify if the log odds are a linear function of the predictors; all predictors are included in the test.

```
##          MLE of lambda Score Statistic (z)  Pr(>|z|)
## age           -0.52249             -0.1684 0.8662891
## trtbps         1.09804             -0.0320 0.9744872
## chol           0.86223             -0.0623 0.9502964
## thalachh       3.37494              0.0582 0.9535787
## oldpeak        0.63318              0.3380 0.7353972
## caa           -0.92643              3.3023 0.0009589 ***
## ---
```

[4]model that resulted in low prediction accuracy are not presented

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  20
```

The test shows that all predictors but **caa** are linearly related to the logit of the outcome variable. A viable solution is to add a quadratic term for caa in the regression model to account for non linearity.

The first fitted model is the **Logistic regression with all the predictors.**

```
##
## Call:
## glm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
##     restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     family = binomial, data = d.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5171  -0.2599   0.1189   0.4745   2.6787
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.09873    0.25344  -0.390  0.69687
## age          0.20325    0.29231   0.695  0.48685
## sex         -0.88484    0.27653  -3.200  0.00138 **
## cp           0.79495    0.25522   3.115  0.00184 **
## trtbps      -0.79724    0.26100  -3.055  0.00225 **
## chol        -0.13677    0.27671  -0.494  0.62111
## fbs          0.33781    0.28200   1.198  0.23095
## restecg      0.46036    0.24583   1.873  0.06111 .
## thalachh     0.47882    0.31273   1.531  0.12574
## exng        -0.35124    0.25396  -1.383  0.16666
## oldpeak     -0.60901    0.36053  -1.689  0.09118 .
## slp          0.51089    0.27883   1.832  0.06691 .
## caa         -1.68978    0.37086  -4.556  5.2e-06 ***
## thall       -0.68952    0.24790  -2.781  0.00541 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 292.69  on 211  degrees of freedom
## Residual deviance: 128.15  on 198  degrees of freedom
## AIC: 156.15
##
## Number of Fisher Scoring iterations: 6
```

The significance of the variables' coefficients mirrors the findings of section 1: age, cholesterol and fbs stand out as the the coefficients with the highest p-values, and are candidates for being cast aside. Sex, chest pain, maximum bpm and the number of not obstructed coronaries show instead a strong impact on the output. As per the coefficients themselves, they are interpreted as the expected change in log odds of observing a higher risk of heart failure per unit change in X. So increasing the predictor by 1 unit (or going from 1 level to the next) multiplies the odds of having the outcome by e^{β}. This means that, if we take sex as a reference, results show that women in the sample have 2.41 times the odds of the men of developing higher risk of heart failure. The same considerations apply to the other predictors.

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
## 1.660003 1.434214 1.265120 1.365924 1.259338 1.152633 1.165287 1.573161
##     exng  oldpeak      slp      caa    thall
## 1.215831 1.486427 1.565599 1.548818 1.208473
```

Variance inflation factor shows no signs of pronounced multicollinearity, thus we can verify another assumption of the logistic model. The performance of the predictor is summarized via a confusion matrix; in the train set the logistic regression, a linear model, reaches an accuracy of 88%

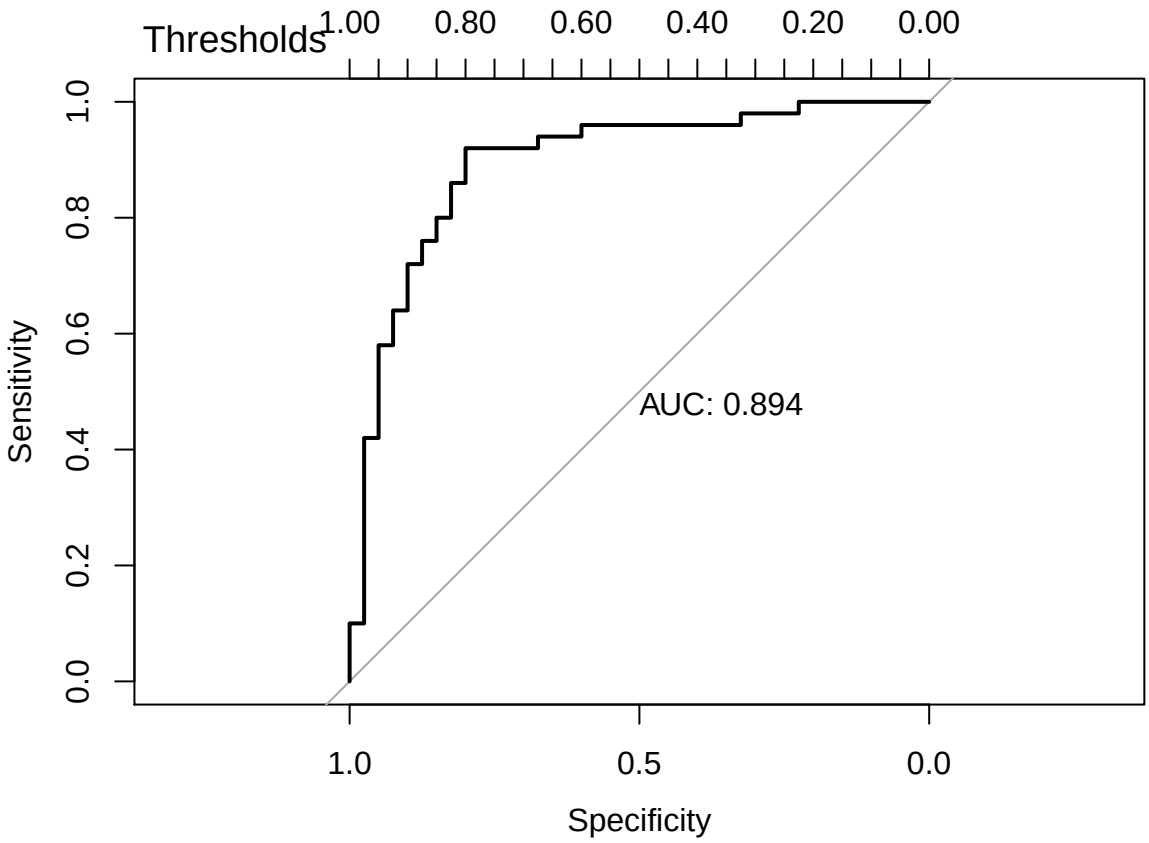
```
##          Actual
## Predicted   0   1
##         0  83  10
##         1  15 104
```

```
## [1] 0.8820755
```

We expect to see this value decreasing in the test set, but not by much: in fact,

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 32  6
##          1  8 44
##
##                Accuracy : 0.8444
##                  95% CI : (0.7528, 0.9123)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 5.419e-09
##
##                   Kappa : 0.6834
##
##  Mcnemar's Test P-Value : 0.7893
##
##             Sensitivity : 0.8800
##             Specificity : 0.8000
##          Pos Pred Value : 0.8462
##          Neg Pred Value : 0.8421
##              Prevalence : 0.5556
##          Detection Rate : 0.4889
##    Detection Prevalence : 0.5778
##       Balanced Accuracy : 0.8400
##
##        'Positive' Class : 1
##
```

the accuracy in the test set is 84%; moreover, type II error is lower than type I by a factor of 1/4.



This result is summarized by the ROC curve for this model: note that, by tweaking the threshold of the conditional probability evaluated to assign the observations to the risk class, it's possible to reach 99.9[…]% sensitivity in the test set, at the expense of a reduction in specificity of 50 percentage points; that is, in correspondence of a threshold of 0.225, we will expect no type II errors in the test set.

The logistic regression model was trained with different sets of predictors, until the final model was produced; this is a **logistic regression without chol and fbs, and a quadratic transformation of the caa variable**:

```
##
## Call:
## glm(formula = output ~ age + sex + cp + trtbps + restecg + thalachh +
##     exng + oldpeak + slp + caa + caa2 + thall, family = binomial,
##     data = d.train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7481  -0.2508   0.1052   0.4375   2.8020
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6299     0.3612  -1.744 0.081126 .
## age           0.2112     0.2793   0.756 0.449591
## sex          -0.8614     0.2677  -3.218 0.001291 **
## cp            0.9182     0.2631   3.490 0.000483 ***
## trtbps       -0.8324     0.2723  -3.057 0.002236 **
## restecg       0.5117     0.2430   2.105 0.035276 *
```

```
## thalachh      0.4941     0.3077   1.606 0.108251
## exng         -0.3242     0.2575  -1.259 0.208038
## oldpeak      -0.7320     0.3695  -1.981 0.047595 *
## slp           0.4803     0.2747   1.748 0.080408 .
## caa          -2.1157     0.4491  -4.711 2.47e-06 ***
## caa2          0.6467     0.2894   2.235 0.025423 *
## thall        -0.7035     0.2490  -2.825 0.004726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 292.69  on 211  degrees of freedom
## Residual deviance: 125.68  on 199  degrees of freedom
## AIC: 151.68
##
## Number of Fisher Scoring iterations: 6
```

the general reduction in p-values signals that this specification is better able to fit the data; however, the improvement is small: the AIC actually decreases, as the reduced penalty is not enough to offset the weight of the two missing predictors.

```
##           Actual
## Predicted   0   1
##         0  81  10
##         1  17 104
```

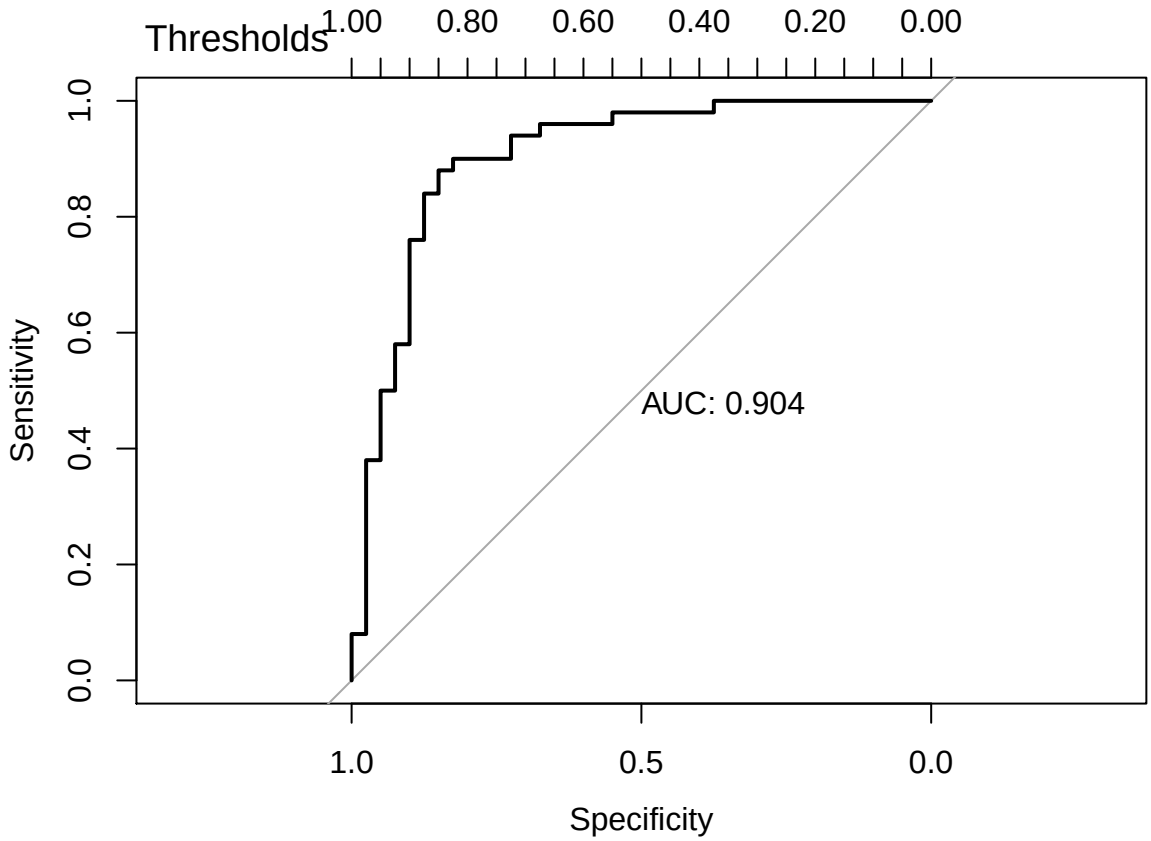
```
## [1] 0.8726415
```

the accuracy in the training set has slightly decreased; again, the quadratic transformation increases the variance, but the absence of two of the predictors, albeit not very significant, makes the difference when comparing such closer results. The test error, however, has also decreased:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 34  6
##          1  6 44
##
##                Accuracy : 0.8667
##                  95% CI : (0.7787, 0.9292)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 2.452e-10
##
##                   Kappa : 0.73
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8800
##             Specificity : 0.8500
##          Pos Pred Value : 0.8800
##          Neg Pred Value : 0.8500
```

```
##              Prevalence : 0.5556
##          Detection Rate : 0.4889
##    Detection Prevalence : 0.5556
##       Balanced Accuracy : 0.8650
##
##        'Positive' Class : 1
##
```

the significant coefficient of caa2 and the -slightly- improved test performance indicates the presence of quadratic boundaries between the classes. Because the logistic regression is a linear model, the presence of a quadratic term increases the fit in the test. Such findings can be confirmed by inspecting again the ROC curve:



Because the specificity is higher with respect to the previous model specification, the lowest possible type I error can be achieved with a threshold of 0.375, making this model a possible candidate for the final selection.

The next step in the analysis marks the passage from discriminative models, such as the logistic regression, to generative models like the Linear Discriminant Analysis and the Quadratic Discriminant analysis. Such models do not estimate directly the probability that the observation y belongs to class k through a function of the predictors, instead they use a Bayesian approach to build a classifier starting from the assumption that the data are generated from k Gaussian distributions; thus first one has to verify this assumption in order to be able to apply these methods, although, even when this assumption does not hold, this type of models can still provide a reasonably good result. To check for multivariate normality in the dependent variables, the Mardia's test for multivariate skewness and kurtosis is performed. Given that tests for multivariate normality such as this are sensitive to large sample sizes, it will be extracted a small sample of ten rows from the data set:

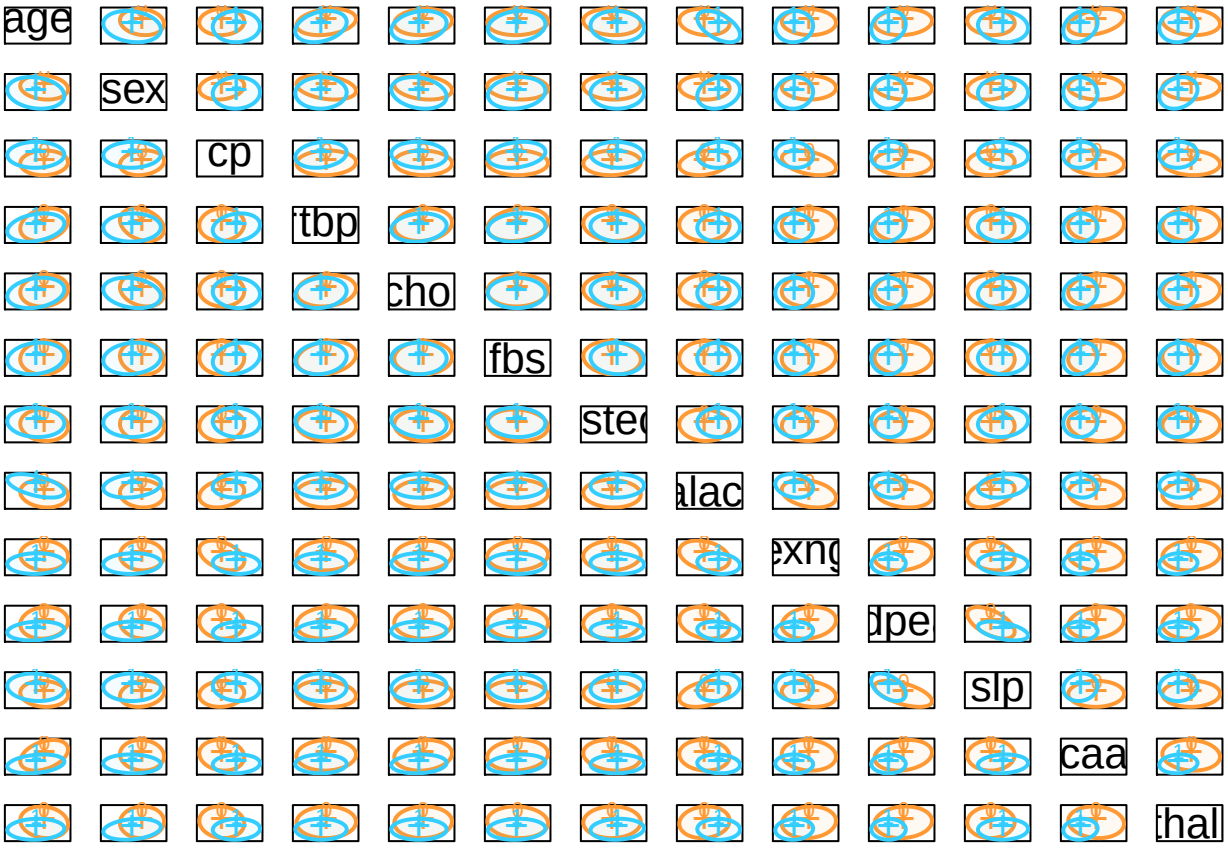
```
##           Beta-hat     kappa      p-val
```

```
## Skewness  7.930853 13.218088 0.99968929
## Kurtosis 21.800331 -2.494503 0.01261337
```

The kurtosis of the multivariate distribution is compatible with a Gaussian, but its skeweness is not. This could be caused by the presence of outliers, or indicate that the distribution has in fact different tails that the ones of a multivariate normal. The Shapiro-Wilk test, less sensitive to outliers, is carried out to verify the assumption.

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.92452, p-value = 3.157e-11
```

The distribution appears indeed not normal. As argued before, it's still possible to use discriminant analysis, but the assumption of a multivariate normal distribution, forcefully superimposed over the data, will cause an approximation error in the classification rate. As in practice this approximation is often still reasonably good, we carry on with the analysis. As mentioned in section 2, QDA differs from LDA in the assumption of the structure of the Gaussian distributions from which the data would be generated: while LDA assumes a common variance-covariance matrix, QDA allows each group to have not only its own unique vector of means, but also one specific variance-covariance matrix. Therefore, before fitting the first model, the assumption of a common variance covariance matrix will be verified through another visual inspection.



While the assumption of a common variance covariance matrix Σ seems reasonable in some cases, the variance of the predictors in the two classes of output appear different by looking at the mapping of the variances; thus the variance-covariance matrices of the resulting multivariate distributions will differ as well. This, and

the fact that there is at least one quadratic term in the function of the posterior probabilities for the classes of y (namely, caa) would suggest that QDA is a better choice wrt to LDA. However, in practice, we have few observations and a high flexible model like QDA would increase the variance significantly, enough to offset completely the gain in bias over LDA and logistic regression. Thus we expect QDA to perform worse than LDA in this context.

The first LDA model fits a classifier using all variables and no transformations:

```
## Call:
## lda(output ~ ., data = d.train)
##
## Prior probabilities of groups:
##         0         1
## 0.4622642 0.5377358
##
## Group means:
##          age        sex         cp     trtbps        chol         fbs
## 0  0.2781724  0.3253051 -0.5381680  0.1608340 -0.01612896  0.01136743
## 1 -0.2282910 -0.2855184  0.2647814 -0.2867386 -0.12519266 -0.04885999
##      restecg   thalachh       exng    oldpeak        slp        caa      thall
## 0 -0.2313173 -0.4564878  0.3879011  0.3862075 -0.3352442  0.5758682  0.3649273
## 1  0.1768676  0.3676099 -0.3613809 -0.4648222  0.3601463 -0.4303142 -0.3512311
##
## Coefficients of linear discriminants:
##                  LD1
## age       0.03322408
## sex      -0.38753388
## cp        0.37952342
## trtbps   -0.26479849
## chol      0.02950422
## fbs       0.12780954
## restecg   0.16860641
## thalachh  0.24772858
## exng     -0.19966169
## oldpeak  -0.16194247
## slp       0.21967081
## caa      -0.64318018
## thall    -0.32535177
```

The output of this model presents other information with respect to the one of the logistic regression: first of all, we can inspect the two groups mean vectors; it's easy to see that these findings are consistent with the results from the hierarchical clustering and the preliminary analysis regarding the sign and value of the predictors' means. The linear discriminant instead can be used to infer the relative importance of each predictor in determining the risk class; again, the results are consistent with what was find through logistic regression.

```
##           Actual
## Predicted   0   1
##         0  76   7
##         1  22 107
```

```
## [1] 0.8632075
```

the accuracy in the training set is 86%, while the misclassification is predominantly caused by type I error; the result is exactly the same in the test set.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 32  4
##          1  8 46
##
##                Accuracy : 0.8667
##                  95% CI : (0.7787, 0.9292)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 2.452e-10
##
##                   Kappa : 0.7273
##
##  Mcnemar's Test P-Value : 0.3865
##
##             Sensitivity : 0.9200
##             Specificity : 0.8000
##          Pos Pred Value : 0.8519
##          Neg Pred Value : 0.8889
##              Prevalence : 0.5556
##          Detection Rate : 0.5111
##    Detection Prevalence : 0.6000
##       Balanced Accuracy : 0.8600
##
##        'Positive' Class : 1
##
```

As expected, LDA results are almost the same as the logistic regression; the small decrease in accuracy in the training set is due to the approximation in the assumptions on the density and characteristic of the covariance matrix; such approximation however slightly increase the bias and thus lead to lower test error -exactly the same as the second model for the linear regression- we therefore expect to observe a corresponding behaviour if the second specification for the model with the quadratic term for caa and withouth fbs and chol is applied.

```
## Call:
## lda(output ~ age + sex + cp + trtbps + restecg + thalachh + exng +
##     oldpeak + slp + caa + caa2 + thall, data = d.train2)
##
## Prior probabilities of groups:
##         0         1
## 0.4622642 0.5377358
##
## Group means:
##          age        sex         cp     trtbps    restecg   thalachh       exng
## 0  0.2781724  0.3253051 -0.5381680  0.1608340 -0.2313173 -0.4564878  0.3879011
## 1 -0.2282910 -0.2855184  0.2647814 -0.2867386  0.1768676  0.3676099 -0.3613809
##      oldpeak        slp        caa      caa2      thall
## 0  0.3862075 -0.3352442  0.5758682 1.4975309  0.3649273
## 1 -0.4648222  0.3601463 -0.4303142 0.5674002 -0.3512311
##
## Coefficients of linear discriminants:
##                  LD1
## age       0.09365519
## sex      -0.39441796
```

```
## cp        0.38897249
## trtbps   -0.24903070
## restecg   0.18149605
## thalachh  0.23388128
## exng     -0.16669971
## oldpeak  -0.22871575
## slp       0.21699721
## caa      -0.99934662
## caa2      0.30870936
## thall    -0.31147774
```

```
##          Actual
## Predicted   0   1
##         0  79   7
##         1  19 107
```

```
## [1] 0.8773585
```

the accuracy in the training set is back to 87%, even in the restricted model; this leads to think that the results in the test set will be worse due to overfitting. In fact

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 31  5
##          1  9 45
##
##                Accuracy : 0.8444
##                  95% CI : (0.7528, 0.9123)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 5.419e-09
##
##                   Kappa : 0.6818
##
##  Mcnemar's Test P-Value : 0.4227
##
##             Sensitivity : 0.9000
##             Specificity : 0.7750
##          Pos Pred Value : 0.8333
##          Neg Pred Value : 0.8611
##              Prevalence : 0.5556
##          Detection Rate : 0.5000
##    Detection Prevalence : 0.6000
##       Balanced Accuracy : 0.8375
##
##        'Positive' Class : 1
##
```

the accuracy in the test set is the lowest among all the four models presented so far.

The quadratic discriminant analysis is the most flexible model among the parametric alternatives presented in this paper. Given what was just observed regarding the performance of the last model, it's expected to perform well in the training set, but poorly in the test. **The first model** uses all the features:

```
## Call:
## qda(output ~ ., data = d.train)
##
## Prior probabilities of groups:
##         0         1
## 0.4622642 0.5377358
##
## Group means:
##          age        sex         cp     trtbps        chol         fbs
## 0  0.2781724  0.3253051 -0.5381680  0.1608340 -0.01612896  0.01136743
## 1 -0.2282910 -0.2855184  0.2647814 -0.2867386 -0.12519266 -0.04885999
##      restecg   thalachh       exng    oldpeak        slp        caa      thall
## 0 -0.2313173 -0.4564878  0.3879011  0.3862075 -0.3352442  0.5758682  0.3649273
## 1  0.1768676  0.3676099 -0.3613809 -0.4648222  0.3601463 -0.4303142 -0.3512311
```

```
##          Actual
## Predicted   0   1
##         0  86   7
##         1  12 107
```

```
## [1] 0.9103774
```

91% accuracy in the training set is yet another clear indicator of the presence of quadratic boundaries, as the most flexible model seen so far approaches the bayesian classifier in the training. However, as noted many times throughout this paper, this will surely lead to the worst results yet in the test set.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 34 11
##          1  6 39
##
##                Accuracy : 0.8111
##                  95% CI : (0.7149, 0.8859)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 3.097e-07
##
##                   Kappa : 0.6222
##
##  Mcnemar's Test P-Value : 0.332
##
##             Sensitivity : 0.7800
##             Specificity : 0.8500
##          Pos Pred Value : 0.8667
##          Neg Pred Value : 0.7556
##              Prevalence : 0.5556
##          Detection Rate : 0.4333
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.8150
##
##        'Positive' Class : 1
##
```

As envisioned, the classifier is badly overfitting and what's worse is that the sensitivity is lower than the specificity. At this point, one could ask how a QDA model with less parameters, and therefore less variance, would perform: **the second QDA model** answer this question, by removing fbs and chol from the predictor space. There is no need to add a quadratic term for caa, as this method already estimates the posterior probabilities through a function that is quadratic in X.

```
## Call:
## qda(output ~ age + sex + cp + trtbps + restecg + thalachh + exng +
##     oldpeak + slp + caa + thall, data = d.train)
##
## Prior probabilities of groups:
##         0         1
## 0.4622642 0.5377358
##
## Group means:
##          age        sex         cp     trtbps    restecg   thalachh       exng
## 0  0.2781724  0.3253051 -0.5381680  0.1608340 -0.2313173 -0.4564878  0.3879011
## 1 -0.2282910 -0.2855184  0.2647814 -0.2867386  0.1768676  0.3676099 -0.3613809
##      oldpeak        slp        caa      thall
## 0  0.3862075 -0.3352442  0.5758682  0.3649273
## 1 -0.4648222  0.3601463 -0.4303142 -0.3512311
```

```
##          Actual
## Predicted   0   1
##         0  84   8
##         1  14 106
```

```
## [1] 0.8962264
```

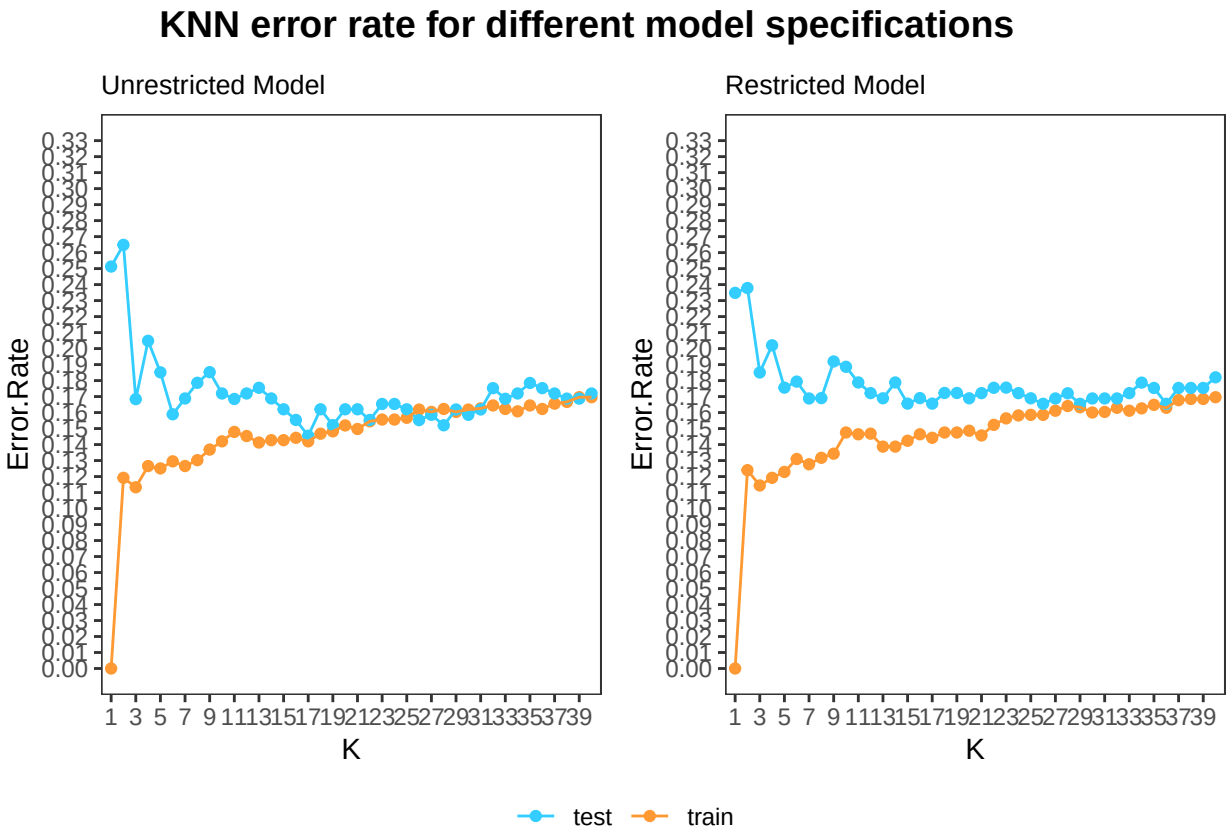
the training error has increased, resulting in a lower accuracy; however it is still quite high.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 34 10
##          1  6 40
##
##                Accuracy : 0.8222
##                  95% CI : (0.7274, 0.8948)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 8.658e-08
##
##                   Kappa : 0.6436
##
##  Mcnemar's Test P-Value : 0.4533
##
##             Sensitivity : 0.8000
##             Specificity : 0.8500
##          Pos Pred Value : 0.8696
##          Neg Pred Value : 0.7727
##              Prevalence : 0.5556
##          Detection Rate : 0.4444
```

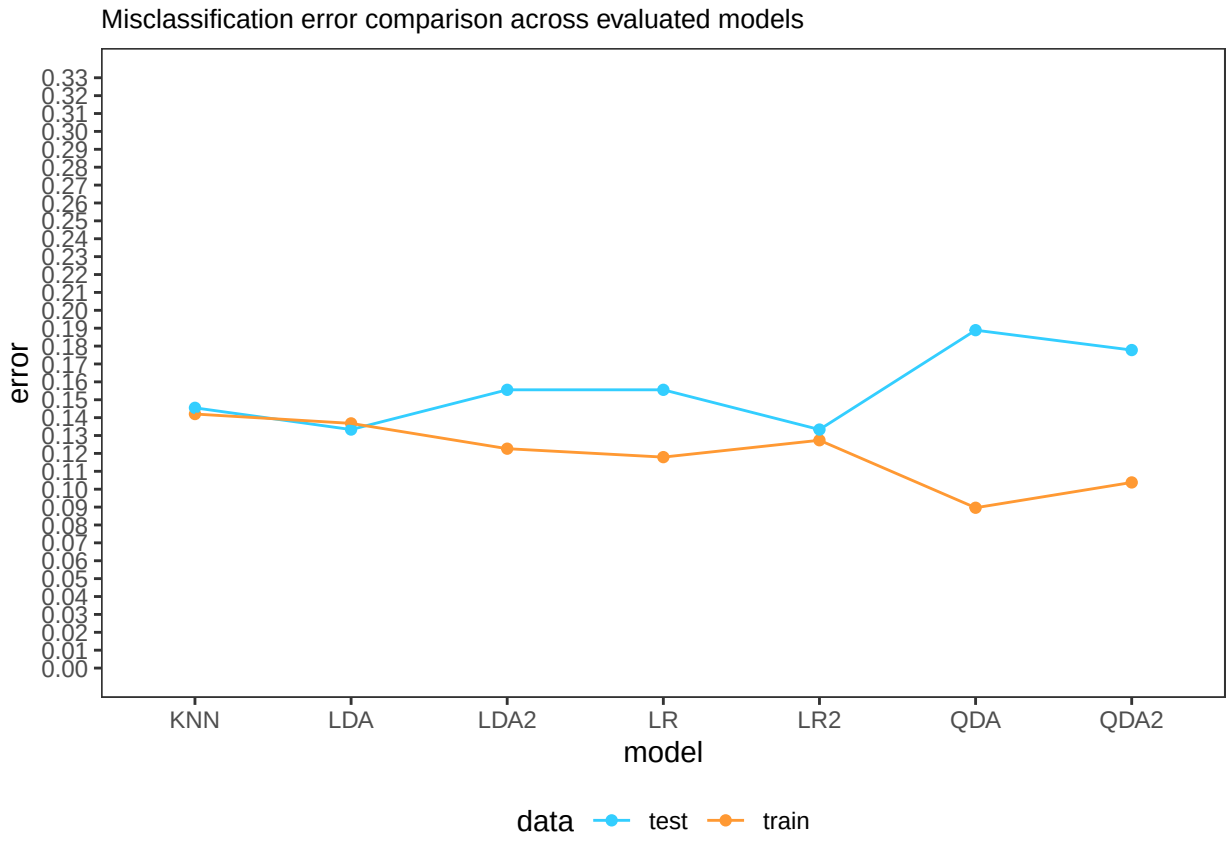
```
##    Detection Prevalence : 0.5111
##       Balanced Accuracy : 0.8250
##
##        'Positive' Class : 1
##
```

sensitivity and test accuracy have slightly increased but, given that the predictors omitted were only weakly correlated with the output, their absence is not making much of a difference in this highly flexible setting: the method still overfit.

3.2 - Non-parametric approach In the models presented so far, a part from the insights gained from interpretation of the coefficients', which lead to the same conclusion across all models, the constant phenomenon has been the bias-variance trade-off in training and test accuracy: from linear and less flexible methods to quadratic discriminant analysis, it was possible to see how models with more variance are performing worse even when bias-inducing assumptions are superimposed over the data and noise coming from redundant predictors is omitted. To complete the model showcase and offer a final insight into this dynamic, a final non-parametric method will now be introduced: **K-nearest neighbor.** This classification algorithm uses distance between observations to search the k-closest point in a neighborhood of an observation to be classified to infer its probable label, based on that of its neighbors. To select the optimal k and inspect the bias-variance trade off dynamic graphically, cross-validated training and test errors are evaluated and plotted. Along with the complete model, another classifier is trained and tested following the usual variable selection, which eliminates chol and fbs from the predictor space. The overfitting region is immediately visible looking at the plot: in correspondence of values of k equal to 1 and 2, the training error is the lowest (in fact, k-nn will have exactly 0 training error when k=1) while the test error is extremely high; as k increases, the curves start to get closer until the lowest possible test error is reached in correspondence of $k = 17$ for the unrestricted model, and $k = 27$ for the restricted one. After this values, the test error starts to slowly rise again, as the bias increases and the classifier gets closer and closer to underfitting. The difference in the two data sets used to train these two models, albeit small, is sufficient to see the other dynamic encountered throughout the analysis with parametric models, still related to the variability in the data. In a noisier environment, including less relevant predictors, the overly flexible knn with $k = 2$ will achieve a lower training error, but a higher test error with respect to its counterpart trained in a less noisy data set; this is due to the fact the the algorithm has to work harder to fit all the training points, and outputs a classifier characterized by higher variance as the predictor space is more crowded. However, as the number of neighbours increases, so does the bias and it does so more quickly and more decisevely in the noisy environment than in the less crowded space; that is why the unrestricted model achieve better performance with a smaller value for k : in a neighborhood of seventeen points around the target there is enough noise to avoid overfitting and output a less flexible classifier with respect to its counterpart in the quieter environment. This double dynamic was encountered throughout the whole analysis, and is the reason why linear methods such as the linear regression and the LDA performed better than quadratic models such as QDA. At the same time, a classifier capable of accounting for the variability in the dataset will perform better in absence of overfitting, that is if enough bias is introduced; this phenomenon was evident when comparing the performance of the restricted QDA with that of its unrestricted counterpart, but is also the reason why LDA, whose assumptions were forcefully introduced thus raising the bias, still outperformed the naive logistic regression; because the function of x in the linear discriminant analysis is accounting also for the prior distribution of the observations, it is a more flexible method with respect to the logistic regression; however, in a less noisy environment, such as the one in which the restricted LDA was trained, the higher flexibility of the method led to worse performances.



3.3 - Model selection In the light of all the previous findings, two models emerge as the best solutions to the problem posed at the beginning of this paper: **Naive LDA** and **Restricted Logistic regression with a quadratic transformation of caa.** Both these models are equally capable of assigning the correct label to the observations in the test set with 87% accuracy, but in truth there is a reason to prefer one over the other. We can visually compare the results of each model (a part for the restricted k-nn, that was already discussed before and served more as a base for the discussion rather than a real alternative) with a final plot.



Here all the considerations presented in the previous sections find a graphical representation, that can instantly summarize the model selection. As for the final choice between LDA and logistic regression, because restricted logistic regression has higher specificity with respect to its competitor, its preferable to choose it over LDA. This seems counter intuitive, but if we were to change the threshold value to increase sensitivity, we would obtain better overall performance from a method with higher specificity, where the increase in the false positives rate would be less troublesome. Of course, if changing the threshold is of no interest, one could prefer the LDA; but from the point of view of a hospital that needs to achieve the lowest possible type II error while still avoid to spend too much resources on patients that don't really need them, logistic regression with a lower threshold value makes more sense.

4 - Conclusions

The scope of this analysis was triple: explore the data at hand to gain insights about the predictors distribution in the population represented by the data set - with no claims on external validity- to find relevant common factors correlated with higher risk of heart failure; compare the performance of several models highlighting each one's faults and merits, with a particular attention on the dynamic of the bias-variance trade-off; select the best possible model for prediction and evaluation of the underlying factors leading to higher risk of heart failure. All this tasks were accomplished and the selected model results presented earlier in the discussion. However, this analysis was biased from the start due to the wrong coding of the variables in the data set. Sadly, the original data set was unretrievable, thus the classifier trained on this data will never be able to work efficiently in another scenario. Not all is lost, as the procedures outlined in this paper for model selection and the conclusion about the models can still be of use for further analysis; moreover, even if on faulty data, the best model was able to achieve 87% accuracy in the test set with high sensitivity and specificity, which leaves room for improvement over the type II error as discussed before.

5- Appendix

```
library(dplyr)
library(knitr)
library(ggplot2)
library(MASS)
library(Gifi)
library(BBmisc)
library(car)
library(caret)
library(arm)
library(pROC)
library(mvnormtest)
library(gridExtra)
library(heplots)
library(class)
library(cowplot)
library(dplyr)
library(boot)
library(QuantPsyc)

data.all<-read.csv("heart.csv", sep = ",")
data.all[c(93,252,159,164,165),12] <- 0
data.all[49,13] <- 2
data.all[282,13]<-3
data.all<-data.all[-86,]
data.unscaled<-data.all
data<-data.all[,-14]
###

#correlations to output of numerical variables (categorical have to be inspected singularly
#since correlation doesn't make sense for categorical variables)
cor(data.all[ , c(1,4,5,8,10,12,14)])[,7]


box_variables <- colnames(data.all)
plot.<-NULL
for(i in box_variables) {
  plot.[[i]] <- ggplot(data.all,
                       aes_string(x = "output",
                                  y = i,
                                  group="output",
                                  fill = "output")) +
    geom_boxplot(alpha = 0.2) +
    theme(legend.position = "none") +
    scale_fill_viridis_c("output")
}
do.call(grid.arrange, c(plot.[c(1,4,5,8,10)], nrow = 1))

#categorical variables: sex, exang,cp,fbs,restecg,slp,thal,caa
#inspect each bivariate contingency table and chi square test to verify if they are independent (thus u
#fbs
contfbs<-table(data.all$output,data.all$fbs,dnn = c("Risk","High Sugar"))
```

```
contfbs #high sugar alone seem indeed uncorrelated in the sample
chisugar<-chisq.test(contfbs)
chisugar #the p-value is not significant and the correlation is low: sugar is independent wrt output
#repeat the same for other predictors:
#exng
contexng<-table(data.all$output,data.all$exng,dnn=c("Risk","Exercise Angina"))
contexng
chiexng<-chisq.test(contexng)
chiexng #if chest pain is induced by exercise, it is less likeable to observe higher risk
#ecg
contecg<-table(data.all$output,data.all$restecg,dnn = c("Risk","Electrocardio"))
contecg #here we see an increase in risk wrt to abnormity in ecg
chicg<-chisq.test(contecg)
chicg #indeed a positive dependence exists, the p value is significant.
#sex
contsex<-table(data.all$output,data.all$sex,dnn = c("Risk","Sex"))
contsex #incidence of high risk in women seems statistically significant
chisexf<-chisq.test(contsex[,1])
chisexf
chisexm<-chisq.test(contsex[,2])
chisexm #sex is  related to high risk, but only because higher risk in women is much higher.
#holding sex constant, there is an increased risk for women wrt men, that experience equally both level
#holding risk constant, the proportion of women experiencing high risk is much higher wrt men. (75% of

#cp
contcp <-table(data.all$output,data.all$cp,dnn=c("Risk","Chest Pain"))
contcp #chest pain appears strongly (positively) correlated with risk
#note how anginal and asymptomatic are more strongly correlated to the response
#wrt to non anginal pain. in particular atypical anginal seems a little more indicative.
#while non anginal chi square valued is not significant at the 5% level
chicp0<-chisq.test(contcp[,1])
chicp0
chicp1<-chisq.test(contcp[,2])
chicp1
chicp2<-chisq.test(contcp[,3])
chicp2
chicp3<-chisq.test(contcp[,4])
chicp3
#slp
contslp <-table(data.all$output,data.all$slp,dnn=c("Risk","Slope Ecg Exercise"))
contslp #slope of the ecg under stress show dependence with risk prediction
chislp<-chisq.test(contslp)
chislp
#thal
contthal<-table(data.all$output,data.all$thall,dnn=c("Risk","Defect"))
contthal #fixed defect (1)appears uncorrelated. Reversible defect(3) corresponds to low level of risk
#normal defect (2) corresponds to higher risk
chital<-chisq.test(contthal)
chital
chithalfix<-chisq.test(contthal[,1])
chithalfix
chithalrev<-chisq.test(contthal[,3])
chithalrev
```

```
chithalnorm<-chisq.test(contthal[,2])
chithalnorm
#caa
caacon<-table(data.all$output,data.all$caa,dnn=c("Risk","Defect"))
caacon
caachi<-chisq.test(caacon)
caachi
#conclusion: output appear dependent from all categorical variables

#hierarchical clustering (all dimensions)
set.seed(13)
num <-normalize(data[ , c(1,4,5,8,10,12)], method = "range", range = c(0,1))
data.clu<-data
data.clu[ , c(1,4,5,8,10,12)] <- num
hclu<-hclust(dist(data.clu), method = "complete")
plot(hclu) #dendrogram suggests 8 clusters
labels<- cutree(hclu,8) #split the tree, return cluster labels
table(labels)
#assign membership
dtclu<-data.all
dtclu$membership <- labels
p<-table(dtclu$membership,dtclu$output, dnn = c("cluster","output"))

#comment on results:
#clusters 5,6,7 and 8 appear to identify combinations of variables that split
#the outcome: verify this with chi square test; then if confirmed, they will identify
#particularly favorable(clusters 7 and 8) and risky(clusters 5 and 6) sets of features
a <- dtclu %>%
  filter(dtclu$membership == 5)
b<- dtclu %>%
  filter(dtclu$membership == 6)

c<- dtclu %>%
  filter(dtclu$membership == 7)
d<- dtclu %>%
  filter(dtclu$membership == 8)
clu.df<-rbind(colMeans(a[,-14]),colMeans(b[,-14]),colMeans(c[,-14]),colMeans(d[,-14]))
b
c
d
#logistic regression:
#test if the log odds are a linear function of the predictors; all predictors are
#included in the test (box-tidwell test)
conti<-data.all[,c(1,4,5,8,10,12)]
conti<-data.matrix(normalize(conti, method = "range", range = c(0.1,1)))
cate<-data.matrix(data.all[,c(2,3,6,7,9,11,13)])
boxTidwell(data.all$output, conti, cate)
#the test shows that all predictors but caa are linearly related to the logit of the
#outcome variable. we can add a quadratic term for caa in the regression model to account
#for non linearity.

#logistic regression fit(all predictors, no quadratic terms) and prediction evaluation
#note that dropping fbs the fit improves as suggested by preliminary findings
```

```r
#training and test data partition
set.seed(55)
data.all[,1:13]<-data.frame(scale(data.all[,1:13]))
split<- createDataPartition(data.all$output,p=0.7,list=FALSE)
d.train<- data.all[split,]
d.test<-  data.all[-split,]
logfit<-glm(output~age+sex+cp+trtbps+chol+fbs+restecg+thalachh+exng+oldpeak+slp+caa+thall,
            data = d.train,family = binomial)
summary(logfit)
vif(logfit) #variance inflation factor shows no signs of pronounced multicollinearity

#evaluation model prediction accuracy #training
lr.probtr <- predict(logfit, d.train, type="response")
lr.predtr <- ifelse(lr.probtr > 0.5,"1","0")
res.tr<-table(Predicted = lr.predtr, Actual = d.train$output)
res.tr
accuracy.tr <- (res.tr[1,1] + res.tr[2,2])/sum(res.tr)
accuracy.tr #the logistic model with all predictors and no transformation of the variables has
#an accuracy of 88% in the training set

#evaluation model prediction accuracy #test
lr.probte <- predict(logfit, d.test, type="response")
lr.predte <- ifelse(lr.probte > 0.5,"1","0")
res.te<-table(Predicted = lr.predte, Actual = d.test$output)
res.te
accuracy.te <- (res.te[1,1] + res.te[2,2])/sum(res.te)
accuracy.te #in the test set, the accuracy is 84%.


#confusion matrix and ROC curve
confusionMatrix(as.factor(lr.predte),
                as.factor(d.test$output),
                positive = "1" )

lr.prob <- predict(logfit, d.test, type="response")
t.roc = roc(d.test$output ~ lr.prob, plot = TRUE, print.auc = TRUE)
lab.thre<-round(t.roc$thresholds, digits = 2)
par(mar = c(5.1, 4.1, 2.1, 2.1))
axis(3, at = seq(from= lab.thre[2], to = lab.thre[length(lab.thre)-1], by =.05))
title("Thresholds", adj=0.05, font.main = 1)

### MODEL 2: NO FBS, NO CHOL, CAA^2
d.train2 <-d.train
d.test2<-d.test
d.train2$caa2<-d.train$caa^2
d.test2$caa2<-d.test$caa^2

logfit2<-glm(output~age+sex+cp+trtbps+restecg+thalachh+exng+oldpeak+slp+caa+caa2+thall,
            data = d.train2,family = binomial)
summary(logfit2)
vif(logfit2)

#evaluation model prediction accuracy #training
```

```
lr.probtr2 <- predict(logfit2, d.train2, type="response")
lr.predtr2 <- ifelse(lr.probtr2 > 0.5,"1","0")
res.tr2<-table(Predicted = lr.predtr2, Actual = d.train$output)
res.tr2
accuracy.tr2 <- (res.tr2[1,1] + res.tr2[2,2])/sum(res.tr2)
accuracy.tr2 #the logistic model withouth chol, fbs and with caa^2 has
#an accuracy of 87% in the training set

#evaluation model prediction accuracy #test
lr.probte2 <- predict(logfit2, d.test2, type="response")
lr.predte2 <- ifelse(lr.probte2 > 0.5,"1","0")
res.te2<-table(Predicted = lr.predte2, Actual = d.test2$output)
res.te2
accuracy.te2 <- (res.te2[1,1] + res.te2[2,2])/sum(res.te2)
accuracy.te2 #in the test set, the accuracy of the adjusted model is 86%.
lr.prob2 <- predict(logfit2, d.test2, type="response")
t.roc2 = roc(d.test2$output ~ lr.prob2, plot = TRUE, print.auc = TRUE)
lab.thre2<-round(t.roc2$thresholds, digits = 2)
par(mar = c(5.1, 4.1, 2.1, 2.1))
axis(3, at = seq(from= lab.thre2[2], to = lab.thre2[length(lab.thre)-1], by =.05))
title("Thresholds", adj=0.05, font.main = 1)
#odds as probabilities


#QDA/LDA
#assumptions: observations are normally distributed (the two classes have class
#specific mean vectors and common[different]LDA[QDA] covariance matrix, while the distribution is a (mu
#we check for multivariate normality in the dependent variables using the Mardia's test
#for multivariate skewness and kurtosis. Given that tests for multivariate normality are sensitive to l
#sample sizes, it will be extracted a small sample of ten rows from the data set:
set.seed(14)
mult.norm(data.unscaled[sample(nrow(data.unscaled), 10),c(1,4,5,8,10)])$mult.test
#kurtosis of the multivariate distribution is compatible with a gaussian, but skeweness is
#not. This could be caused by the presence of outliers, or indicate that the distribution
#has in fact different tails that ones of a multivariate normal.
#we use the shapiro test, less sensitive to outliers:
test<-(data.all[,c(1,4,5,8,10)])
mvnormtest::mshapiro.test(t(test))
#the distribution appears indeed not normal. We can still use discriminant analysis,
#but the assumption of a multivariate normal distribution, forcefully superimposed over the
#data, will cause an approximation error in the classification rate. As in practice this assumption
#is still reasonably good, we carry on with the analysis.

#variance-covariance testing:
do.call(grid.arrange, c(plot.[c(1,4,5,8,10)], nrow = 1))
covEllipses(data.all[,-14],data.all$output, fill=T, pooled = F, col = c("#FF9933", "#33CEFF"),
            variables=colnames(data.all[,-14]),fill.alpha = 0.05)
#the variance of the predictors in the two classes of output appear different from the box plots and
#the mapping of the variances, thus the variance covariance matrix of the multivariate normal
#for the two classes will differ as well.
#this, and the fact that there is at least one quadratic term in the function of the log odds of the po
#probabilities for the classes of y (namely, caa) would suggest that QDA is a better choice wrt to LDA.
#However, in practice, we have few observations and a high flexible model like QDA would increase
```

```r
#the variance significantly, enough to offset completely the gain in bias over LDA and logistic regress
#Thus we expect QDA to perform worse than LDA in this context.

#fitting
#LDA #model with all predictors and no transformations
#training
set.seed(11)
lda.1<-lda(output~.,data = d.train)
lda.1
lda.prob.tr1 <- predict(lda.1, d.train, type="response")
lda.pred.tr1 <- ifelse(lda.prob.tr1$posterior[,2] > 0.5,"1","0")
lda.res.tr1<-table(Predicted = lda.pred.tr1, Actual = d.train$output)
lda.res.tr1
accuracy.lda.tr.1 <- (lda.res.tr1[1,1] + lda.res.tr1[2,2])/sum(lda.res.tr1)
accuracy.lda.tr.1 #86%
#test
lda.prob.ts1 <- predict(lda.1, d.test, type="response")
lda.pred.ts1 <- ifelse(lda.prob.ts1$posterior[,2] > 0.5,"1","0")
lda.res.ts1<-table(Predicted = lda.pred.ts1, Actual = d.test$output)
lda.res.ts1
accuracy.lda.ts.1 <- (lda.res.ts1[1,1] + lda.res.ts1[2,2])/sum(lda.res.ts1)
accuracy.lda.ts.1 #86%
#as expected, LDA results are almost the same as the logistic regression; the small decrease in accurac
#is due to the approximation in the assumptions on the density and characteristic of the covariance mat
#we therefore expect to observe the same behaviour if we use the second specification for the model
#with the quadratic term for caa and withouth fbs and chol
lda.2<-lda(output~age+sex+cp+trtbps+restecg+thalachh+exng+oldpeak+slp+caa+caa2+thall,data = d.train2)
lda.2
lda.prob.tr2 <- predict(lda.2, d.train2, type="response")
lda.pred.tr2 <- ifelse(lda.prob.tr2$posterior[,2] > 0.5,"1","0")
lda.res.tr2<-table(Predicted = lda.pred.tr2, Actual = d.train2$output)
lda.res.tr2
accuracy.lda.tr.2 <- (lda.res.tr2[1,1] + lda.res.tr2[2,2])/sum(lda.res.tr2)
accuracy.lda.tr.2 #87%
#test
lda.prob.ts2 <- predict(lda.2, d.test2, type="response")
lda.pred.ts2 <- ifelse(lda.prob.ts2$posterior[,2] > 0.5,"1","0")
lda.res.ts2<-table(Predicted = lda.pred.ts2, Actual = d.test2$output)
lda.res.ts2
accuracy.lda.ts.2 <- (lda.res.ts2[1,1] + lda.res.ts2[2,2])/sum(lda.res.ts2)
accuracy.lda.ts.2 #84%
#exactly as expected

#QDA
#model 1 (all predictors)
qda.1<-qda(output~.,data = d.train)
qda.1
qda.prob.tr1 <- predict(qda.1, d.train, type="response")
qda.pred.tr1 <- ifelse(qda.prob.tr1$posterior[,2] > 0.5,"1","0")
qda.res.tr1<-table(Predicted = qda.pred.tr1, Actual = d.train$output)
qda.res.tr1
accuracy.qda.tr.1 <- (qda.res.tr1[1,1] + qda.res.tr1[2,2])/sum(qda.res.tr1)
accuracy.qda.tr.1 #91% the decrease in training error reflects the flexibility of the method
```

```r
#test
qda.prob.ts1 <- predict(qda.1, d.test, type="response")
qda.pred.ts1 <- ifelse(qda.prob.ts1$posterior[,2] > 0.5,"1","0")
qda.res.ts1<-table(Predicted = qda.pred.ts1, Actual = d.test$output)
qda.res.ts1
accuracy.qda.ts.1 <- (qda.res.ts1[1,1] + qda.res.ts1[2,2])/sum(qda.res.ts1)
accuracy.qda.ts.1 #81% as expected, the test error is higher wrt to the less flexible methods

#model 2 (restricted: drop chol and fbs)
qda.2<-qda(output~age+sex+cp+trtbps+restecg+thalachh+exng+oldpeak+slp+caa+thall,data = d.train)
qda.2
qda.prob.tr2 <- predict(qda.2, d.train, type="response")
qda.pred.tr2 <- ifelse(qda.prob.tr2$posterior[,2] > 0.5,"1","0")
qda.res.tr2<-table(Predicted = qda.pred.tr2, Actual = d.train$output)
qda.res.tr2
accuracy.qda.tr.2 <- (qda.res.tr2[1,1] + qda.res.tr2[2,2])/sum(qda.res.tr2)
accuracy.qda.tr.2 #89% having dropped some predictors, the variability has decreased
qda.prob.ts2 <- predict(qda.2, d.test, type="response")
qda.pred.ts2 <- ifelse(qda.prob.ts2$posterior[,2] > 0.5,"1","0")
qda.res.ts2<-table(Predicted = qda.pred.ts2, Actual = d.test$output)
qda.res.ts2
accuracy.qda.ts.2 <- (qda.res.ts2[1,1] + qda.res.ts2[2,2])/sum(qda.res.ts2)
accuracy.qda.ts.2 #82% even tho we dropped some predictors, given that they were not significant
#to build the decision boundary, the method still overfit wrt to the others

#KNN
set.seed(17)
data.all<-read.csv("heart.csv", sep = ",")
data.all[c(93,252,159,164,165),12] <- 0
data.all[49,13] <- 2
data.all[282,13]<-3
data.all<-data.all[-86,]
data.all<-as.data.frame(scale(data.all))

calc_error_rate <- function(predicted.value, true.value){
  return(mean(true.value!=predicted.value)) }

#model 1
#use 10 fold cv to select k
cv.ertest.1<-rep(1:40)
cv.ertrai.1<-rep(1:40)
x<-rep(1:10)
y<-rep(1:10)
data.all<-data.all[sample(nrow(data.all)),]
folds <- cut(seq(1,nrow(data.all)),breaks=10,labels=FALSE)

for (i in 1:40){

  for(j in 1:10){
    testIndexes <- which(folds==j,arr.ind=TRUE)
    testData <- data.all[testIndexes, ]
    trainData <- data.all[-testIndexes, ]
```

```
    knn.pred1.1<-knn(trainData[-14],trainData[-14], trainData$output, k = i) #train
    as.numeric(as.character(knn.pred1.1))
    x[j]<-calc_error_rate(knn.pred1.1, trainData$output)

    knn.pred1.2<-knn(trainData[-14],testData[-14], trainData$output, k = i) #test
    as.numeric(as.character(knn.pred1.2))
    y[j]<-calc_error_rate(knn.pred1.2, testData$output)
  }
  cv.ertrai.1[i] = sum(x)/length(x)
  cv.ertest.1[i] = sum(y)/length(y)

}
cv.ertrai.1
cv.ertest.1

cvknn1.1<-data.frame(cv.ertrai.1, seq(1,40),rep("train",40))
colnames(cvknn1.1)<-c("Error.Rate","K","data")
cvknn1.2<-data.frame(cv.ertest.1, seq(1,40),rep("test",40))
colnames(cvknn1.2)<-c("Error.Rate","K","data")
df1<-rbind(cvknn1.1,cvknn1.2)

error.plot.knn<-ggplot(data = df1, aes(x = K, y = Error.Rate, color=data)) +
  geom_point()+ #arancio training
  geom_line() +
  scale_color_manual(name= "",values=c("#33CEFF", "#FF9933"))+
  scale_x_continuous(breaks = seq(min(cvknn1.1$K), max(cvknn1.1$K), by = 2),expand = c(.02,.02))+
  theme_test()+
  scale_y_continuous(breaks = round(seq(0, 0.33, by = 0.01),2), limits = c(0,0.33))+
  labs(title = "Unrestricted Model")+
  theme(plot.title = element_text(size = 10), legend.position = "none")

error.plot.knn
#### model without chol and fbs
cv.ertest.2<-rep(1:40)
cv.ertrai.2<-rep(1:40)
x<-rep(1:10)
y<-rep(1:10)
data.all<-data.all[sample(nrow(data.all)),]
folds <- cut(seq(1,nrow(data.all)),breaks=10,labels=FALSE)

for (i in 1:40){

  for(j in 1:10){
    testIndexes <- which(folds==j,arr.ind=TRUE)
    testData <- data.all[testIndexes, ]
    trainData <- data.all[-testIndexes, ]

    knn.pred2.1<-knn(trainData[-14],trainData[-14], trainData$output, k = i) #train
    as.numeric(as.character(knn.pred2.1))
    x[j]<-calc_error_rate(knn.pred2.1, trainData$output)

    knn.pred2.2<-knn(trainData[-14],testData[-14], trainData$output, k = i) #test
    as.numeric(as.character(knn.pred2.2))
```

```
    y[j]<-calc_error_rate(knn.pred2.2, testData$output)
  }
  cv.ertrai.2[i] = sum(x)/length(x)
  cv.ertest.2[i] = sum(y)/length(y)

}
cv.ertrai.2
cv.ertest.2

cvknn2.1<-data.frame(cv.ertrai.2, seq(1,40),rep("train",40))
colnames(cvknn2.1)<-c("Error.Rate","K","data")
cvknn2.2<-data.frame(cv.ertest.2, seq(1,40),rep("test",40))
colnames(cvknn2.2)<-c("Error.Rate","K","data")
df2<-rbind(cvknn2.1,cvknn2.2)

error.plot.knn2<-ggplot(data = df2, aes(x = K, y = Error.Rate, color=data)) +
  geom_point()+ #arancio training
  geom_line() +
  scale_color_manual(values=c("#33CEFF", "#FF9933"))+
  scale_x_continuous(breaks = seq(min(cvknn2.1$K), max(cvknn2.1$K), by = 2),expand = c(.02,.02))+
  theme_test()+
  scale_y_continuous(breaks = round(seq(0, 0.33, by = 0.01),2), limits = c(0,0.33))+
  labs(title = "Restricted Model")+
  theme(plot.title = element_text(size = 10), legend.position = "none")

error.plot.knn2

knn.error.plot <-plot_grid(error.plot.knn,error.plot.knn2)
title <- ggdraw() +
  draw_label("KNN error rate for different model specifications",
             x = 0,hjust = 0,fontface = "bold") +
  theme(plot.margin = margin(0, 0, 0, 63))

legend <- get_legend(
  error.plot.knn +
    guides(color = guide_legend(nrow = 1)) +
    theme(legend.position = "bottom"))

plot_grid(title,knn.error.plot,ncol = 1,rel_heights = c(0.1, 1),legend)

####
#model comparison: train and error test for all considered model, plotted

train.error<-c(1-accuracy.tr,1-accuracy.tr2,1-accuracy.lda.tr.1,1-accuracy.lda.tr.2,1-accuracy.qda.tr.1
test.error<-c(1-accuracy.te,1-accuracy.te2,1-accuracy.lda.ts.1,1-accuracy.lda.ts.2,1-accuracy.qda.ts.1,1
error<-c(train.error,test.error)
error<-data.frame(error)
error$data<-NA
error$data[1:7]<-"train"
error$data[8:14]<-"test"
df3<-error
df3$model<-rep(c("LR","LR2","LDA","LDA2","QDA","QDA2","KNN"),2)
train_test<-ggplot(data = df3, aes(x = model
```

```
                                        , y = error, color=data)) +
  geom_point() + #arancio training
  geom_line(aes(group=data)) +
  scale_color_manual(values=c("#33CEFF", "#FF9933"))+
  theme_test()+
  scale_y_continuous(breaks = round(seq(0, 0.33, by = 0.01),2), limits = c(0,0.33))+
  labs(title = "Misclassification error comparison across evaluated models")+
  theme(plot.title = element_text(size = 10), legend.position = "bottom")
train_test
```