Politecnico di Milano

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE
Corso di Laurea in Ingegneria Matematica

Statistical analysis of optical data for tumour diagnosis

Relatore:
Prof. Anna Maria Paganoni

Autore:
Jacopo Cotta Ramusino

29 Aprile 2015
Anno Accademico 2013 - 2014

Abstract

The main technique for the identification of breast cancer is mammography, ie the identification of areas with a high concentration by X-rays emitted to the affected area. Optical spectroscopy is a possible innovation in this field, in fact, it would enable people to find tumors non-invasively, using visible and near infrared radiations.
This approach is based on the estimation of concentrations of the main constituents of the breast tissue, these concentrations can also be used to build a classifier to identify malignant tumors and benign tumors.
Meantime you can proceed to the identification of components that can be considered dangerous, or significant risk factors for the preventive diagnosis of breast cancer. In recent work, the topic of breast density has already been studied, proving that a higher density implies a strong increase in the probability of contracting breast cancer. It was also noted that a relationship item exists between collagen and density.
It can therefore be interesting to see if there are indeed differences between these two risk factors or whether they can be considered substantially similar in the preventive analysis of breast tissue.
These are precisely the issues addressed in this work, in which we tried to get answers aiming to the quality and meaningfulness of the classifiers in the first case, and to the definition of possible compatibility or incompatibility of the two risk factors in the second case.

keywords: *collagen*, *density*, *optical mammography*, *boosting methods*, *random forest*, *malignant and benign tumours*, *classification*.

Sommario

La tecnica più utilizzata per l'identificazione di tumori al seno è la mammografia, che consiste nell'identificazione delle aree con concentrazione estremamente elevata mediante l'emissione di raggi X nell'area interessata. La spettroscopia ottica può essere una possibile innovazione in questo campo, infatti permetterebbe l'identificazione di tumori in modo non invasivo usando radiazioni nel campo del visibile e quasi infrarosse.
L'intero approccio è basato sulla stima delle concentrazioni dei principali costituenti del tessuto mammario, le quali sono successivamente utilizzate per costruire un adeguato classificatore per identificare tumori maligni e benigni.
Allo stesso tempo è fondamentale cercare di identificare quali costituenti possono essere considerati pericolosi in termini di diagnosi preventiva: studi attuali mostrano che i soggetti con alta densità mammaria hanno una maggiore probabilità di contrarre tumore al seno. Inoltre è stata identificata una forte relazione tra densità mammaria e concentrazione di collagene.
La naturale conseguenza è quindi la valutazione dell'impatto del collagene nella quantificazione del rischio di contrarre un tumore; è molto interessante anche verificare se tale costituente permette di avere un maggiore potere predittivo rispetto alle informazioni fornite dalla densità mammaria.
Questi sono esattamente i temi trattati in questo lavoro, nel quale ci si è concentrati sulle performance e sulla significatività dei classificatori per quanto riguarda la prima parte, mentre per quanto concerne la seconda parte è stata data molta attenzione alle possibili compatibilità dei due fattori di rischio prima citati (collagene e densità).

keywords: *collagene*, *densità*, *mammografia ottica*, *boosting methods*, *random forest*, *tumori maligni e benigni*, *classificazione*.

Contents

Chapter 1

Introduction

Breast cancer is a leading cause of death in women and a major health burden worldwide; an early diagnosis is a key element in order to grow up the percentage of survival.
Breast density has been recognized as a strong risk factor for breast cancer: high density involves a four to six times higher risk as compared to low density. Breast density describes the relative amount of glandular and connettive tissue present in the breast. Currently it's assessed with radiological appearance of breast tissue (mammographic density), but it's easy to understand that a tool for non-invasive estimation would allow an early identification of high-risk women.
A response to this need could be found in the optical techniques because they could provide structural information on breast tissue in a non-invasive way. These techniques have been successfully applied to the characterization of breast tissue [15]. In effect, in recent years, diffuse spectroscopy has opened the way to non-invasive optical characterization of biological tissue and has fostered the development of several pre-clinical and clinical applications like optical mammography.
Correlation between optically derived parameters and mammographic density was observed in previous experiments [16]. Time-resolved transmittance data were collected at seven red and near-infrared wavelengths using a portable clinical instrument for time-resolved optical mammography. The instrument is presently applied in a clinical study approved by the institutional review of the European Institute of Oncology.
The study has a twofold aim: the optical characterization of malignant and benign breast lesions and the non-invasive assessment of breast density. For the second aim the study referred to the BI-RAD System (Breast Imaging and Reporting Data System) mammographic density categories:

1. almost entirely fat;
2. scattered fibrogranular densities;

3. heterogeneously dense;
4. extremely dense.

In [16] this categorization has been used to define which factors are significant in determining the density of a subject, and these factors are:

- scattering parameters (amplitude and power);
- density of Collagen.

We must also consider that the BIRAD classification is subjective and it is not the only method of qualitative classification based on mammography. The non-invasive assessment of breast density with optically derived parameters is a possible solution to this problem. The next step is the assessment of malignant and benign tumours based on the concentrations of the main components in tissue composition, and this is the goal of the current work.
In the second part of the work an evaluation similar to that made for the density will be made, trying to understand if the collagen can also be considered as a significant risk factor for the detection of breast cancer.

1.1 Time-resolved diffuse optical spectroscopy

The portable clinical instrument for time-resolved optical mammography used to collect data operates in transmittance geometry on the mildly compressed breast [17]. Time-resolved transmittance data are collected at seven wavelengths in the range 635 to 1060 nm (i.e., 635, 685, 785, 905, 930, 975, 1060 nm) using picosecond pulsed diode lasers as light sources, and two photomultiplier tubes and PC boards for time-correlated single photon counting to detect the time distribution of the transmitted pulses. Data are stored every millimetre (Figure 1.1).

A single driver controls all the laser heads, and their output pulses are properly delayed by means of graded index optical fibers, and combined into a single coupler. A lens produces a collimated beam that illuminates the breast (softly compressed between parallel anti-reflection plates) and a fiber bundle collects the output light on the opposite side of the compression unit; the distal end of the bundle is bifurcated, and its two legs guide photons respectively to a photomultiplier tube for the detection of VIS or NIR wavelengths. The PC boards allow the acquisition of time-resolved transmittance curves at VIS and NIR wavelengths.
Time-resolved spectral data are interpreted with a spectrally constrained global fitting procedure to estimate tissue composition in term of water, lipid, collagen, oxy- and deoxy-hemoglobin content, as well as scattering parameters: a (amplitude) and b (power).
The time-resolved spectral constraint analysis is based on a two step procedure [7]: in

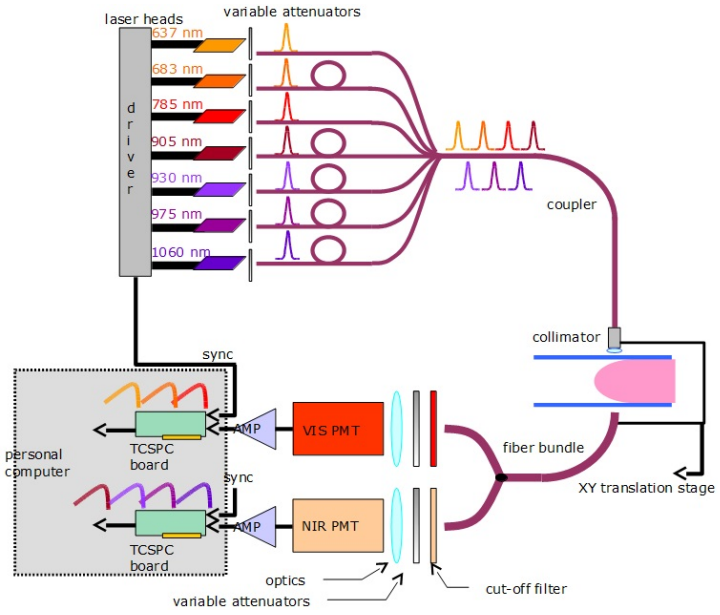

Figure 1.1: Instrument set-up: VIS visible (635, 680, 785nm); NIR near infrared (905, 930, 975, 1060nm); PMT photomultiplier tube; TCSPC time-correlated single photon counting

the first step optical parameters (μ_a and μ'_s) are calculated, they are the absorption and scattering parameters. To do calculations you can fit the time-resolved curves to an analytical solution of the diffusion approximation to the transport equation for an infinite homogeneous slab with extrapolated boundary conditions, given by:

$$T\left(t;\mu_a,\mu'_s\right) = 0.5\left(\frac{4\pi}{3\mu'_s}\right)^{-3/2} t^{-5/2} exp\left(-\mu_a \nu t\right) \times \\ \times \sum_{n=-\infty}^{+\infty}\left[z_n^+ exp\left(-\frac{3\mu'_s\left(z_n^+\right)^2}{4\nu t}\right) - z_n^- exp\left(-\frac{3\mu'_s\left(z_n^-\right)^2}{4\nu t}\right)\right] \tag{1.1}$$

where

$$z_n^+ = (1-2n)\,d - 4nz_e - z_0$$

$$z_n^- = (1-2n)\,d - (4n-2)\,z_e + z_0$$

$$z_0 = \left(9\mu'_s\right)^{-1}$$

$$z_e = \left(2A/3\mu'_s\right)$$

where ν is the speed of light in the medium, d is the thickness of the sample and A take into account the reflections at the slab surface. The theoretical curve is convolved

with instrumental response function and normalized to the area of the experimental curve. Substantially the goal is to find the values of μ_a and μ_s' that minimize the difference between the theoretical curve and experimental data (Levenberg-Marquard algorithm) [12].
In the second step the absorption coefficient μ_a is used with the Lambert-Beer law:

$$\mu_a(\lambda) = \sum_i c_i \epsilon_i(\lambda) \tag{1.2}$$

where c_i is the concentration of a constituent and $\epsilon_i(\lambda)$ is the extinction coefficient of the i-th constituent. So this law allows us to estimate tissue components concentration from the knowledge of the extinction coefficients $\epsilon_i(\lambda)$.
The choice of the value of λ depends on the technological availability and the absorption spectrum of the constituents (Figure 1.2).
We can see that several wavelengths are chosen to be near a peak: this fact allows us to

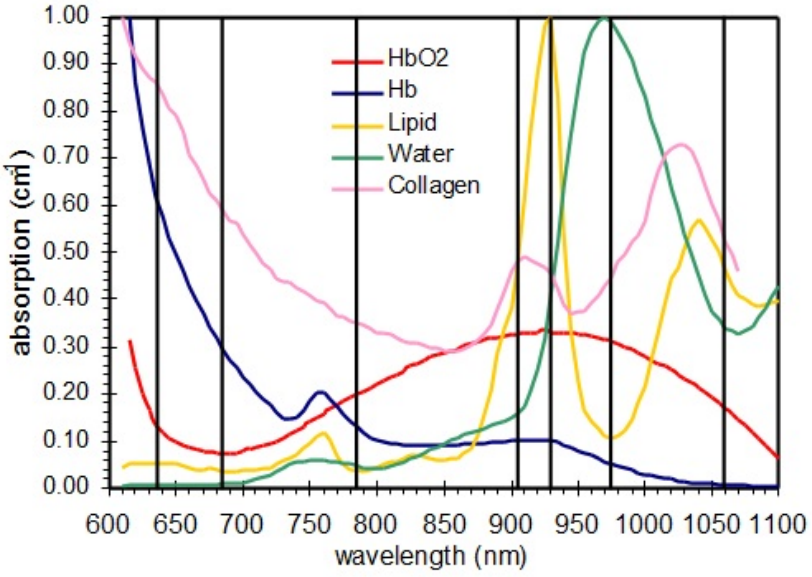

Figure 1.2: Absorption spectrum of the main constituents

have a good idea about a particular constituent, while the lowest values of λ describe the behaviour of hemoglobine and blood.

1.2 Data Collection

In the course of the analysis 4 different datasets were used; for simplicity they have been renamed as Dataset 1, Dataset 2, Dataset 3 and Dataset 4.

The first three datasets refer to a single phase of data collection: a time-domain multiwavelength (635 to 1060 nm) optical mammography was performed on 83 subjects (45 malignant and 38 benign) and average breast tissue compositions were estimated.
The compression unit can be rotated by an angle up to 90°in both clock-wise and counter-clock-wise direction, so that the images of both breast can be recorded in the cranio-caudal (CC) as well as medio-lateral or oblique (OB) views [18]. For this reason 16 benign and 23 malignant subjects have double absorption (and consequently concentration) data.
For the first part of the analysis we consider every observation independent from the others, although some of them are derived from the same subjects; obviously this is a strong approximation due to the fact that both the measurement methods and the approximation of concentrations are subject to errors and therefore lead to different results for the same subject.
Then taking into account these observations, for a first qualitative analysis, we have a total of 122 observations.
In particular, you can briefly describe the datasets as follows:

- **Dataset 1** is composed by two parts, the first one collects the absorptions at the different wavelengths, the second one the different concentrations (relating to the constituents). In each of the two parts we have data relating to the damaged area of the breast and healthy area. The concentrations were estimated in the way described in Section 1.1 and the available variables are:
  - reduced hemoglobin (HHb);
  - oxidized hemoglobin (O_2Hb);
  - total hemoglobin (tHb);
  - oxygenation (SO_2);
  - Lipid;
  - Water;
  - Collagen.
- **Dataset 2** contains the differences between lesioned area and healthy are (delta) in terms of both concentrations and absorptions (values are then calculated from Dataset 1).
  The reason of the use of the delta is mainly based on the following observation: values related to the specific type of tissue in the case of benign and malignant lesions are affected by an error caused by the error on estimation of the shape and size of the lesion that is given to various clinical data (histopathology, mammography, ultrasound, depending on the cases and the type of lesion).
  This error on the estimation of the shape/size of the lesion is the most likely cause of

some negative values of concentrations. It's been made an attempt to fix the possible errors of dimensions and in fact it could reduce the number of negative values, but the choice of the regularization factor was arbitrary. Negative concentrations are definitely wrong, but, once we have positive values, there is a whole wide range of physiologically possible values and there is no way to figure out what is the right value in the case of any specific lesion.
This is why we have decided to use the delta.

- **Dataset 3** is built in the following way: considering the Dataset 2, replacing the observations with double angle for measurements (CC and OB) with the mean value (both in terms of concentrations and absorptions).

During the work comparisons have been made between the results derived from the whole dataset, Dataset 2, and results derived from transformed data, Dataset 3 (average value of OB and CC views for every subject).
A first observation about the variables used is the following one:

- $tHB = HHb + O_2Hb$;
- $SO_2 = \frac{O_2Hb}{HHb+O_2Hb} = \frac{O_2Hb}{tHB}$.

The relations expressed are deterministic, then to make a reduction in the number of variables used we will avoid to use SO_2 and tHb.
The first step of the analysis is mainly focused on the differences between Malignant and Benign subjects' spectrums (both in the healthy and in the lesioned area) and on the identification of the causes of those differences: to do this we use the Dataset 1. We will pay attention to the differences of the covariance matrices (if any) and we will investigate which are the different correlation values in terms of absorption and concentration using Dataset 2 and 3.
The second step is mainly based on the search for a suitable classifier in order to identify the differences between subjects with malignant tumours and patients with benign tumours. These are the contents of the Chapter 2 and Chapter 3.
The second part of the work is to identify a further significant risk factor for the identifications of patients with early breast cancer. As already mentioned the first significant risk factor coincides with the density. High density significantly increases the probability of developing cancer.
The work shown in Chapter 4 is aimed at verifying whether the collagen can be identified as a significant risk factor and if the two risk factors are somehow connected to each other (if they can be considered as complementary factors to identify high probability of contracting the disease or if they are somehow correlated)
Recent research has shown that there is a probable link between cancer risk and concentration of collagen (in addition to the density of the breast as described above). For this reason we've studied data related to 107 subjects; the variables are:

- Density of the breast;
- Collagen;
- Age;
- Menopausal state;
- BMI.

All of this information constitute the **Dataset 4** and the analysis is shown in Chapter 4. The software used for the whole analysis is R [24] and the packages used are listed in Chapter 6.

Chapter 2

Differences in Variance and Mean

We can find the differences in covariance structure of data (Dataset 2) using the **Multivariate Bartlett Test** or a more stable method for non normal data (**Levene's test**), or we can try to define a different variability of the two groups (Malignant and Benign) with distance-based tests for homogeneity of multivariate dispersion.

2.1 Multivariate Bartlett Test

The univariate Bartlett Test was proposed in 1937 to test the homogeneity of variances and it has been later extended to the multivariate case. The basic hypothesis for this test is that the samples of size $n_1, n_2, \ldots, n_k$ are randomly drawn from k multivariate normally distributed populations [13].
The null hypothesis of equality of covariances matrices is given by

$$H_0 : \Sigma_1 = \Sigma_2 \tag{2.1}$$

Suppose p is the number of variables involved, so Σ_i is of size $p \times p$. To perform the test we calculate

$$M = \frac{|S_1|^{\nu_1/2}|S_2|^{\nu_2/2}}{|S_{pooled}|^{\nu_E/2}} \tag{2.2}$$

where

- $\nu_i = n_i - 1$;
- $\nu_E = n - k$;
- S_i is the covariance matrix of the ith sample;

- S_{pooled} is the pooled sample covariance matrix: $S_{pooled} = \frac{\nu_1 S_1 + \nu_2 S_2}{n-k}$;
- n is the number of observations.

We refer to a chi-square approximation for the distribution of M using the statistic u which is:

$$u = -2\,(1-c)\log M \tag{2.3}$$

where

$$c = \left[\frac{1}{\nu_1} + \frac{1}{\nu_2} - \left(\frac{1}{\nu_1+\nu_2}\right)\right]\left[\frac{2p^2+3p-1}{6\,(p+1)\,(k-1)}\right]. \tag{2.4}$$

It can be proved that u is approximately distributed as $\chi^2\left[\frac{1}{2}\,(k-1)\,p\,(p+1)\right]$.
We reject H_0 if

$$u > \chi^2_{1-alpha}\left[\tfrac{1}{2}\,(k-1)\,p\,(p+1)\right].$$

For completeness, we note that there is a further approximated distribution of M (generally underused): the F approximation.
For the F approximation, the statistic depends on two quantities, c_1 and c_1:

- $c_1 = \left[\frac{1}{\nu_1} + \frac{1}{\nu_2} - \left(\frac{1}{\nu_1+\nu_2}\right)\right]\left[\frac{2p^2+3p-1}{6(p+1)(k-1)}\right]$
- $c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\frac{1}{\nu_i^2} - \left(\frac{1}{\left(\sum_{i=1}^{k}\nu_i\right)^2}\right)\right]$

If $c_2 > c_1^2$, then

$$F = -2b_1 \log M \tag{2.5}$$

If $c_2 < c_1^2$, then

$$F = -\frac{a_2 b_2 \log M}{a_1\,(1+2b_2\log M)} \tag{2.6}$$

where

- $a_1 = \frac{1}{2}\,(k-1)\,p\,(p+1)$
- $a_2 = \frac{a_1+2}{|c_2-c_1^2|}$
- $b_1 = \frac{1-c_1-a_1/a_2}{a_1}$
- $b_2 = \frac{1-c_1-2/a_2}{a_2}$

In both cases F is approximately distributed as $F\,(a_1, a_2)$.
The meaning of the test is highly dependent on the normality assumption, that in our case is not verified (for Dataset 2). For this reason it is reasonable to use other tests that are quite robust to departures from normality (**Levene's Test**).

2.2 Levene's Test and Some Multivariate Analogues

In the univariate case, let x_{ij} be a set of $j = 1, \ldots, n_i$ observations in each of $i = 1, \ldots, g$ groups. Levene's test statistic is the ANOVA F-ratio comparing the g groups, calculated on the absolute deviations $z_{ij} = |x_{ij} - \overline{x_i}|$ from the group means $\overline{x_i}$.
For the multivariate case, let x_{ij} be the vector which denotes the jth observation of the ith group in the multivariate space of p variables. Let $\Delta(\cdot, \cdot)$ denote the Euclidean distance between two points, let c_i be the centroid for group i. One multivariate analogue to Levene's test is to perform the ANOVA on the Euclidean distances from individual points in a group to their group centroid,

$$z_{ij}^c = \Delta(x_{ij}, c_i)$$

A p-value of the F-statistic (F_c) may be obtained either by using the traditional F-distribution or by using a permutation procedure.
A most robust method is obtained by using the distances from individual points in a group to their group median [5], the point that minimizes the sum of distances to points within that group:

$$z_{ij}^m = \Delta(x_{ij}, m_i)$$

A p-value of the F-statistic (F_m) could be obtained in the same way of the centroid case.

2.3 Extension to any Dissimilarity using Principal Coordinates

The approach shown before could be extended on any distance through the use of principal coordinates. This type of analysis is helpful for us in order to have a graphical representation of the sample (if we are interested in Euclidean distances) or to repeat the test of equal covariance matrices (using a different dissimilarity measure).
Let $d_{ll'}$ be the distance between the lth and l'th observations, to obtain principal coordinates first we have to define the matrix A, where $a_{ll'} = -\frac{1}{2} d_{ll'}$.
Centering this matrix [11] we obtain:

$$G = [g_{ll'}] = [a_{ll'} - \overline{a_{l\cdot}} - \overline{a_{\cdot l'}} + \overline{a_{\cdot\cdot}}] \tag{2.7}$$

where

- $\overline{a_{l\cdot}}$ is the mean for row l;
- $\overline{a_{\cdot l'}}$ is the mean for column l';

- $\overline{a..}$ is the overall mean of the values of A.

Spectral decomposition of G yields

$$G = \sum_{l=1}^{N} \lambda_l \boldsymbol{q_l} \boldsymbol{q_l}^T \tag{2.8}$$

where $\lambda_1 \geq \cdots \geq \lambda_N$ are ordered eigenvalues of G and $\boldsymbol{q_1}, \cdots, \boldsymbol{q_N}$ are the corresponding orthonormal eigenvectors. Principal coordinate axes are then obtained by scaling each axis q_l by the square root ot its corresponding eigenvalue, $\boldsymbol{u_l} = (\lambda_l)^{1/2} \boldsymbol{q_l}$.
Matrix G may not be nonnegative definite, this is generally the case of a semimetric used as distance function. If some eigenvalues are negative, the axes of matrix Q can be split into two sets:

$$Q = [\boldsymbol{q_1} \cdots \boldsymbol{q_r} | \boldsymbol{q_{r+1}}, \cdots \boldsymbol{q_N}] \tag{2.9}$$

such that $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ and $0 > \lambda_{r+1} \geq \cdots \geq \lambda_N$. For eigenvectors corresponding to negative eigenvalues, we may scale by the square root of the absolute value of λ_l and subsequently multiply by $(-1)^{1/2}$, recognizing that these correspond to axes in imaginary space, i.e. $(-1)^{1/2} u_l = (|\lambda_l|)^{1/2} q_l$. So we can consider two different groups of principal coordinate axes.
Thus, let

$$U = \left[U^+ | U^-\right] \tag{2.10}$$

be a $N \times N$ matrix of principal coordinate axis, the row $\boldsymbol{u_{ij}}^+$ gives the coordinates along $1, \cdots, r$ real axes for jth observation in the ith group, and the row $\boldsymbol{u_{ij}}^-$ gives the coordinates along $r+1, \cdots, N$ imaginary axes. We can calculate a centroid for each group in each of the real and imaginary spaces as c_i^+ and c_i^-. Then we can define

$$z_{ij}^c = \sqrt{\Delta^2\left(\boldsymbol{u_{ij}}^+, c_i^+\right) - \Delta^2\left(\boldsymbol{u_{ij}}^-, c_i^-\right)}. \tag{2.11}$$

Similarly, using spatial medians instead of centroids we can define

$$z_{ij}^m = \sqrt{\Delta^2\left(\boldsymbol{u_{ij}}^+, m_i^+\right) - \Delta^2\left(\boldsymbol{u_{ij}}^-, m_i^-\right)}. \tag{2.12}$$

The test for homogeneity of dispersion then simply consists of doing univariate one-way ANOVA on the z's with or without the use of permutations. It's been demonstrated that if D contains Euclidean distances between the original observations, then the distances calculated with this method are unchanged [1].

2.4 Differences in mean for data with different covariance

If there are different covariance matrices, we can't apply the Hotelling test because the hypothesis of equal covariance matrices is not satisfied. We have to use the central limit

theorem even if the sample size is not very large because every test on the mean that approximates the Hotelling test assume the normality of the mean vectors.
This assumption is highly forced but necessary to use the approximation of the Hotelling test: this is based on the quadratic form

$$T^2 = \left(\overline{X_1} - \overline{X_2}\right)' \tilde{\Omega}^{-1} \left(\overline{X_1} - \overline{X_2}\right) \tag{2.13}$$

where $\tilde{\Omega}$ is an estimate of the $Cov\left(\overline{X_1} - \overline{X_2}\right) = \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}$.
Using the unbiased estimator $\tilde{S}_i = \frac{S_i}{n_i}$ for $\frac{\Sigma_i}{n_i}$ we get the statistic

$$T_u^2 = \left(\overline{X_1} - \overline{X_2}\right)' \tilde{S}^{-1} \left(\overline{X_1} - \overline{X_2}\right) \tag{2.14}$$

with $\tilde{S} = \tilde{S}_1 + \tilde{S}_2$ It's possible to demonstrate that

$$T_u^2 = \frac{\mu p}{\mu - p + 1} F_{p,\mu-p+1} \tag{2.15}$$

where

$$\mu = \frac{p+p^2}{\frac{1}{n_1-1}\left\{tr\left[\left(\tilde{S}_1\tilde{S}^{-1}\right)^2\right]+\left[tr\left(\tilde{S}_1\tilde{S}^{-1}\right)\right]^2\right\}+\frac{1}{n_2-1}\left\{tr\left[\left(\tilde{S}_2\tilde{S}^{-1}\right)^2\right]+\left[tr\left(\tilde{S}_2\tilde{S}^{-1}\right)\right]^2\right\}}$$

For a given observed value T_{u0}^2, the test rejects the null hypothesis of equal mean vectors when

$$P\left(\frac{\mu p}{\mu-p+1} F_{p,\mu-p+1} > T_{u0}^2\right) < \alpha$$

2.5 Testing the differences in absorptions

The first step of the analysis is mainly focused on the differences between the spectra of malignant and benign patients both in lesioned area and in healthy area. We use Dataset 1 to do this, and in particular the part of the data relating to absorption at the seven wavelengths. As we can see observing the matplots in Figure 2.1 and Figure 2.2 (they recreate the idea of the spectra) there's a strong difference in lesioned area (while the healthy area is very similar).

For 5 patients with benign disease and 4 patients with malignant disease the lesioned area's absorptions assume negative values (for at least one wavelength). Negative values can't be considered reasonable, they could be due to technical errors, for this reason they have been deleted from Dataset 1.
These subjects should not be considered even in the other studies, then observations relating to these subjects have been removed even in Dataset 2 and Dataset 3.
We consider from now Dataset 2, containing the differences between lesioned and healthy areas for every subject. Before starting to use the tests presented above, Mahalanobis

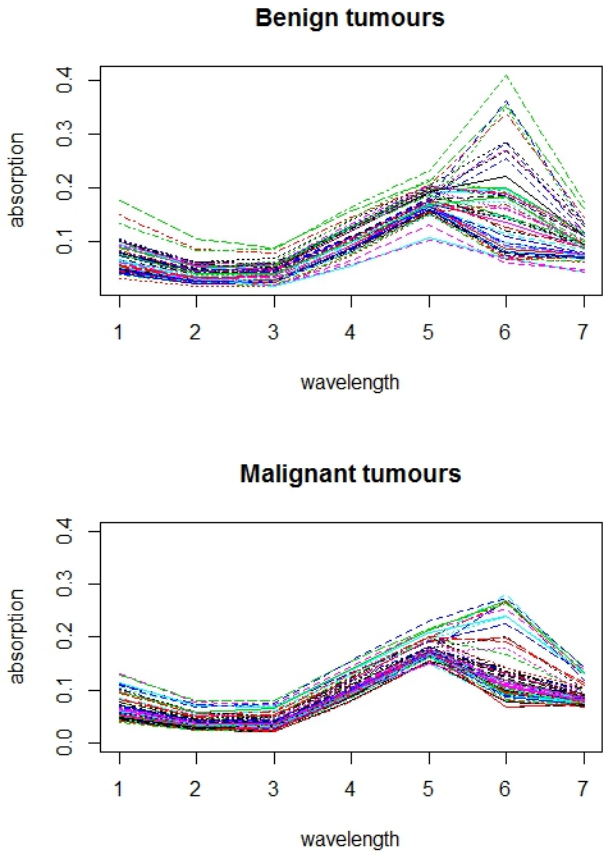

Figure 2.1: Absorption spectra in healthy area for subjects with Benign and Malignant disease, Dataset 1 has been used. The wavelengths are: $1 = 635nm$, $2 = 685nm$, $3 = 785nm$, $4 = 905nm$, $5 = 930nm$, $6 = 975nm$, $7 = 1060nm$

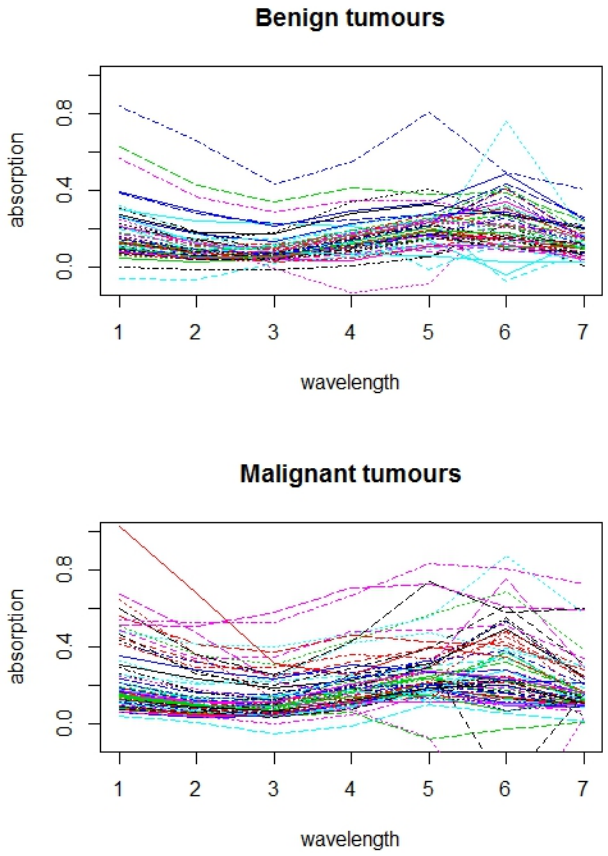

Figure 2.2: Absorption spectra in lesioned area for subjects with Benign and Malignant disease, Dataset 1 has been used. The wavelengths are: $1 = 635nm$, $2 = 685nm$, $3 = 785nm$, $4 = 905nm$, $5 = 930nm$, $6 = 975nm$, $7 = 1060nm$

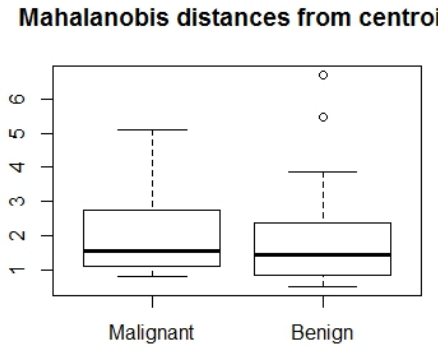

Figure 2.3: Mahalanobis distances from each point to the goup centroid

distances from centroids are calculated in order to find possible outliers (Figure 2.3).
For this reason others 1 subject with malignant disease and 2 subjects with benign disease are deleted because they're considered outliers. Naturally the observations relating to these subjects were also deleted in the Dataset 1 and Dataset 3.
With the remaining observations the hypothesis of equal covariance matrices and equal means have been tested.
The basic hypothesis for the Multivariate Bartlett Test is the normality of the populations. This hypothesis has been tested in our case for Malignant and Benign tumour subjects ($k = 2$) with the **Shapiro multivariate test**, but the p-values was very low ($\approx 10^{-10}$).
I want to remember that we are testing the data referred to the differences between healthy and lesioned area for both the subjects' groups.
A first observation is that the hypothesis of normality isn't acceptable in this case, but we have to note that the sample size is low and we're in the case of multivariate sample, so we would need a really high sample size to apply the Shapiro test with reliable results. For this reason the Bartlett Test has been used even if the assumption of normality is not satisfied. Naturally the results should be interpreted considering this fact.
The Multivariate Bartlett Test has given the following results:

- $u = 149.9745$;
- $\chi^2_{0.95}\,[28] = 41.33714$;
- $p - value = 3.08642e - 14$.

The null hypothesis could be rejected because the p-value is really low.
The second test is then the Levene's Test, more robust than the Bartlett Test to departures from normality:

- $F_c = 20.45$;
- $p - value_c = 1.57e - 05$;
- $F_m = 9.6235$;
- $p - value_m = 0.002452$.

where F_c and $p - value_c$ are referred to the test based on distances from centroids, F_m and $p - value_m$ are referred to the test based on distances from medians. Figure 2.4 shows the boxplots with the euclidean distances from group centroids and group medians. This test

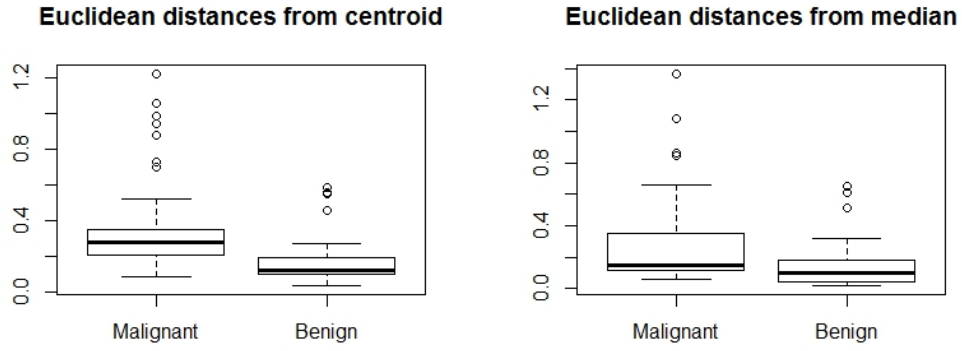

Figure 2.4: Euclidean distances from the individual points and the centroid/median

leads to the same conclusions of the Bartlett test.
To confirm the results obtained, principal coordinates are used with the test presented in section 2.3, euclidean distances are used just to obtain a graphical representation of the data spatial distribution (Figure 2.5) because the F-values would be the same than in the Levene's test (because of the invariance of distances after transformation), then other distances (i.e. Manhattan distances) could be used to have a confirmation of the results obtained until now.

The Figures 2.6 and 2.7 are the representations of the method described before using Manhattan distance.

The results for the principal coordinate method with Manhattan distances are as follows:

- $F_c = 18.452$;
- $p - value_c = 3.823e - 05$;
- $F_m = 8.522$;

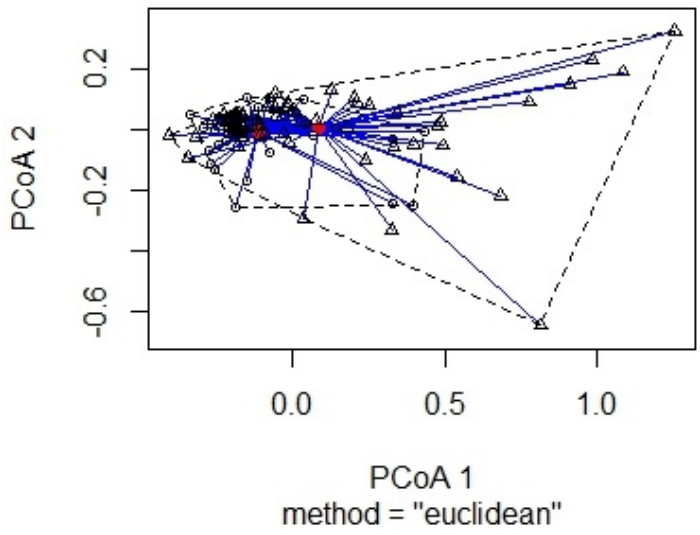

Figure 2.5: Principal coordinates ordination for the differences in absorption between subjects with Malignant and Benign diseases using euclidean distances: the triangles represent subjects with Malignant disease, the circles represent subjects with Benign disease. In this graph and in all the following relating to the method in question, only the first two principal coordinates (PCoA 1 and PCoA 2) will be reported for reasons of graphical intuition

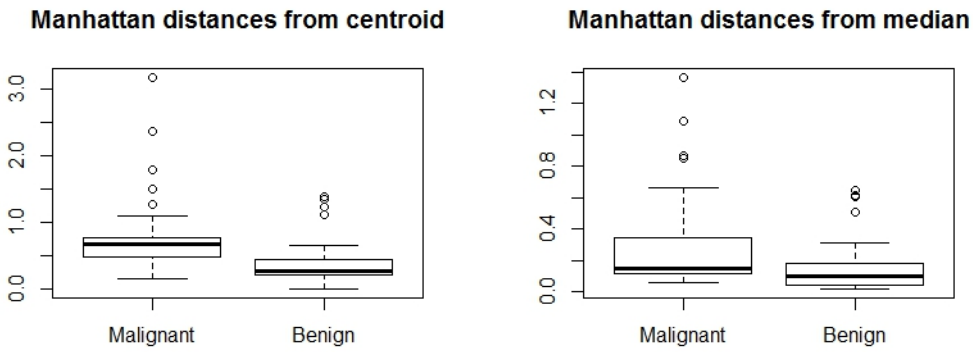

Figure 2.6: Boxplots obtained with Manhattan distances from the group centroids/medians

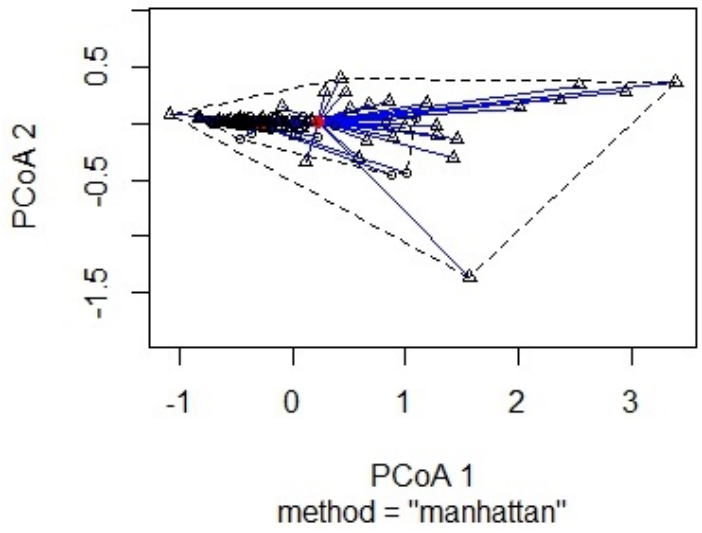

Figure 2.7: Principal coordinates ordination for the differences in absorption between subjects with Malignant and Benign diseases using Manhattan distances: the triangles represent subjects with Malignant disease, the circles represent subjects with Benign disease.

- $p - value_m = 0.004271$.

The results obtained until now are a strong evidence in favour of a difference in covariance matrices structure. This fact justifies the use of the test described in section 2.4 for equality of means; the results of the test are as follows:

- $\frac{\mu-p+1}{\mu p} T_0^2 = 2.380886$;
- $F_{0.95,p,\mu-p+1} = 2.056987$;
- $p - value = 0.02340946$.

The easy consideration that results is that it is possible to say that there is a difference in means between the two populations.

2.6 Testing the differences in concentrations

After analyzing the differences in absorption between the affected areas and healthy areas, we focused on the differences in concentrations of the main constituents, checking if there are different behaviours for patients with benign and those with malignant tumours. Of

course, some differences are expected because the concentrations are closely related to the absorptions (where we checked differences both on mean values and covariance matrices). We are using Dataset 2 without the observations previously deleted because of the anomaly of measured absorption or consideration as outliers; but now we're interested in the part of the data relating to concentrations.
In fact all the tests described above lead to obtain results similar to those previously obtained:

- **Multivariate Bartlett Test**: with the same assumptions made in the case of the absorption about the non-normality of the data (Shapiro tests' p-values were very low) the results are:
  - $u = 95.3543$;
  - $\chi^2_{0.95}[15] = 24.99579$;
  - $p - value = 3.745056e - 10$.

  The really low p-value is a strong evidence in favour of difference in variance structure between subjects with Benign disease and Malignant disease.
- **Levene's Test**:
  - $F_c = 20.45$;
  - $p - value_c = 1.57e - 05$;
  - $F_m = 5.1771$;
  - $p - value_m = 0.02486$.

  even in this case we've reported the results of this test both for distances from centroid and distances from median. The values of F and p allow us to reject the null hypothesis of equal variance.
- **Principal Coordinates Analysis**: this test was performed using both euclidean and Manhattan measure; the meaning of the test is the same as in the previous case. Figure 2.8 and Figure 2.9 show the principal coordinate ordinations with Euclidean and Manhattan measures: it's possible to note that Malignant data have an higher variability compared to the variability of subjects with Benign disease.
  The results with Manhattan measure are:
  - $F_c = 12.274$;
  - $p - value_c = 0.0006695$;
  - $F_m = 6.5537$;
  - $p - value_m = 0.01185$.

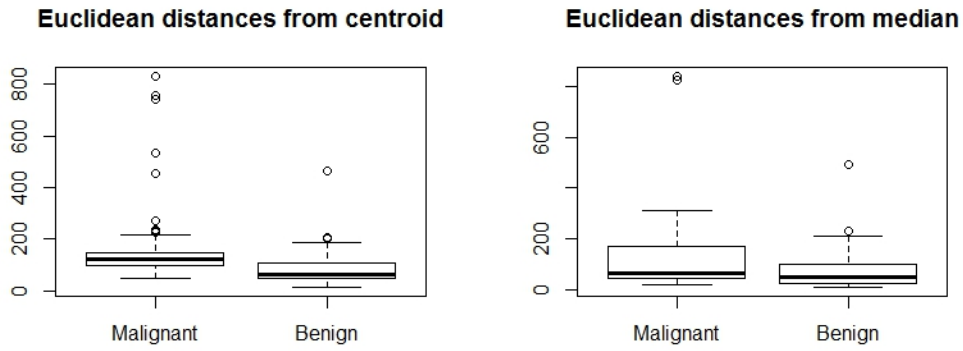

Figure 2.8: Euclidean distances from the individual points and the centroid/median

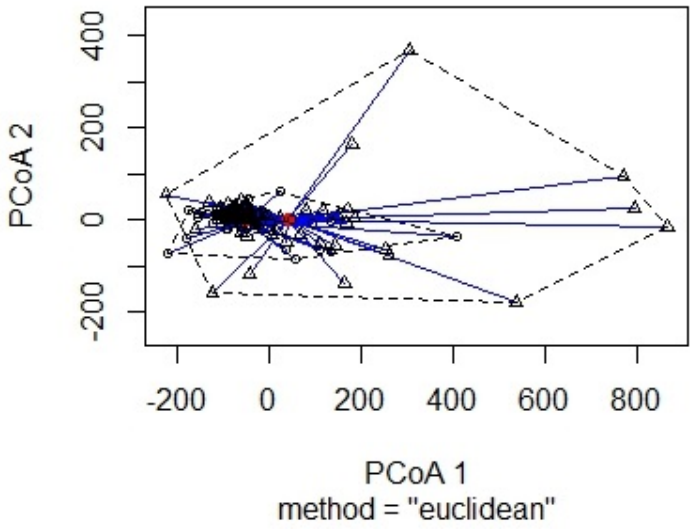

Figure 2.9: Principal coordinates ordination for the differences in absorption between subjects with Malignant and Benign diseases using euclidean distances: the triangles represent subjects with Malignant disease, the circles represent subjects with Benign disease.

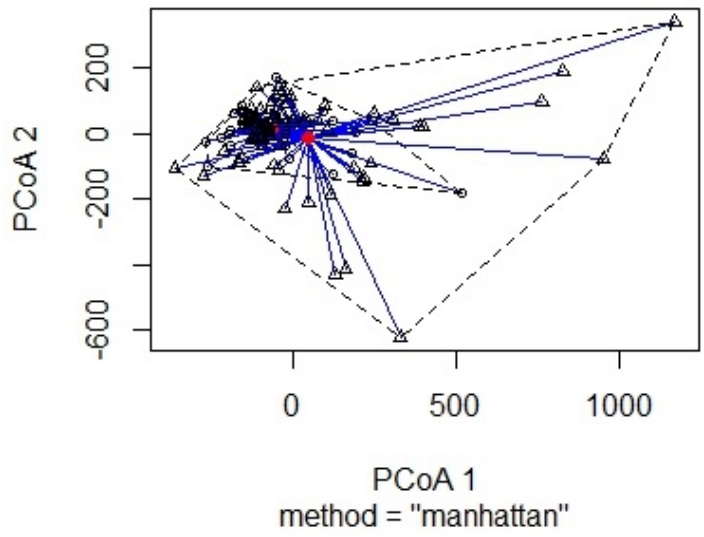

Figure 2.10: Principal coordinates ordination for the differences in absorption between subjects with Malignant and Benign diseases using euclidean distances: the triangles represent subjects with Malignant disease, the circles represent subjects with Benign disease.

- **equality of mean vectors**: previous tests allow us to confirm the hypothesis of heterogeneity of the data. For this reason it's possible to apply the test shown in section 2.5. The results are as follows:
  - $\frac{\mu-p+1}{\nu p} T_0^2 = 2.846808$;
  - $F_{0.95,p,\mu-p+1} = 2.262422$;
  - $p - value = 0.01674482$.

The conclusions of this tests are that it is possible to confirm a difference in terms of mean and variability for the two populations.
This consideration is fundamental to begin to investigate the most effective methods that allow us to build a suitable classifier.

Chapter 3

Classification

In this chapter the problem of classification is treated. The main idea is to find a good logistic regression model to explain which factors affect particularly the probability of being a subject with malignant tumour instead of a subject with a Benign one.
To do this, it's necessary to explore the correlations between the variables and then to find a significant subset of them. In the latest part of the chapter other classifiers are used in order to define a variable importance rank: this is a possible help to decide which variables to consider.
The pattern of action is the following one:

- try to build a classifier based on the concentration of Dataset 2 (we want to use observations that reduce the interpersonal effect and bring down as much as possible the effect of measurement errors, see section 1.2 for more explanations). Concentrations have better physical meaning, for this reason they are preferred in the construction of a classifier instead of absorptions;
- Repeat the analysis with the part of Dataset 2 concerning absorptions data.

Given the difference tested previously on the variability of the two samples, an initial attempt to classify was made based on PCA (Principal Components).

3.1 PCA

The PCA has been applied on Dataset 2 with the following results: considering the firsts 3 components, a total of over 98% of variability is explained.
Figure 3.1 shows the boxplots of the scores on the original directions and on the directions identified by the Principal Component Analysis (PCA).

```
Importance of components:
                          Comp.1      Comp.2      Comp.3       Comp.4       Comp.5
Standard deviation     182.0360645 56.03447768 38.22933960 19.316174722 11.981468188
Proportion of Variance   0.8662137  0.08207688  0.03820355  0.009753319  0.003752582
Cumulative Proportion    0.8662137  0.94829055  0.98649410  0.996247418  1.000000000
```

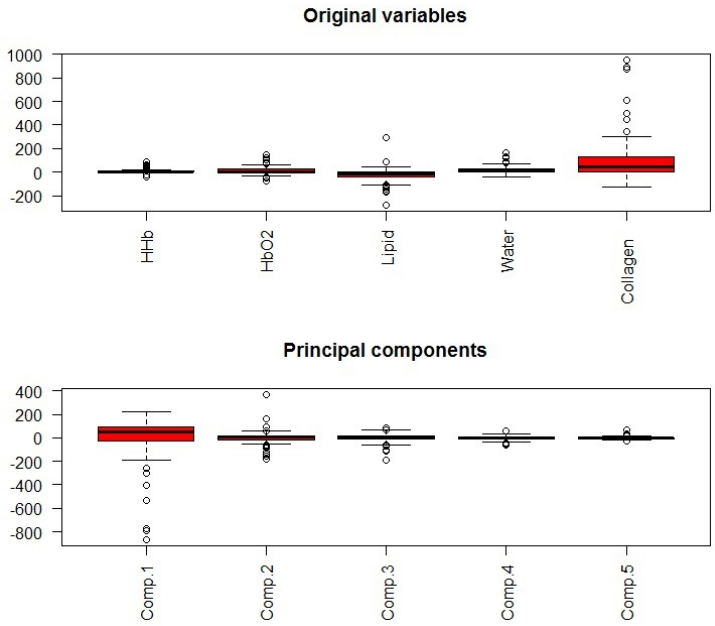

Figure 3.1: Boxplot with original variables and PCA directions of the whole sample

It's important to note that Collagen and Lipid are very meaningful for the whole variability because they're present in the first two principal components directions (Figure 3.2).

The main cause of the difference in variability between benign and malignant tumours

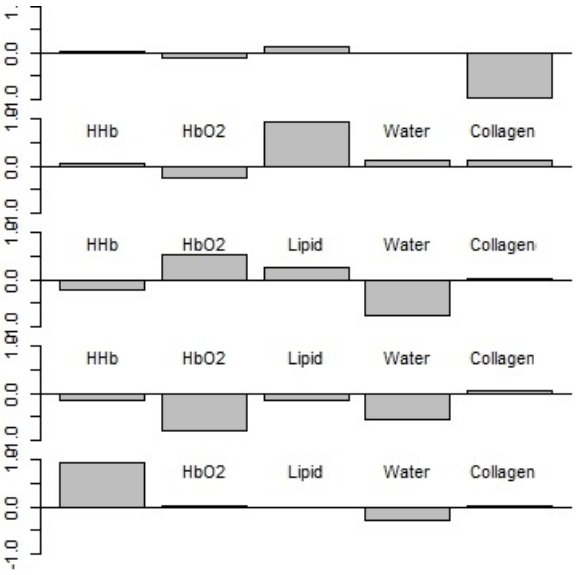

Figure 3.2: Boxplot with original variables and PCA directions of Dataset 2

should be likely found in the variables involved in loadings of the 3 components.
As previously affirmed, a logistic regression model was fitted with the scores along the PCA directions, and a stepwise selection has been made (in particular a backward elimination considering p-values of the coefficients). The resulting model is then:

$$logit\,(p_i) = \alpha_0 + \alpha_1 \cdot [Comp.1]_i + \alpha_2 \cdot [Comp.4]_i \tag{3.1}$$

where $[Comp.1]_i$ and $[Comp.4]_i$ are the scores of the i-th observation on the directions 1 and 4. Surprisingly $[Comp.2]_i$ and $[Comp.3]_i$ were not significant in this model.
The performance are as follows:

- Sensitivity=0.6666667;
- Specificity=0.5531915;
- $1 - TotalErrorRate = 0.6181818$.

The performance are defined in the following way: calculate the predicted probability (according to the fitted model) of being a subject with Malignant disease. If this probability is more than 0.5, the subject is classified as Malignant (subjects with $p = 1$ are considered with malignant disease), otherwise he's classified as Benign. Using these classifications you

can calculate the performances of the classifier.
A summary description of model 3.1 is now reported:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)    0.387955   0.213357   1.818   0.0690.
Comp.1        -0.004292   0.001776  -2.417   0.0157*
Comp.4        -0.020035   0.011592  -1.728   0.0839.
---
  Null deviance  : 150.16 on 109 degrees of freedom
Residual deviance: 138.54 on 107 degrees of freedom
AIC: 144.54
```

It is important to underline that the first principal component is substantially identified by Collagen while the fourth component is identified by Water and HBO_2 (with loading with same verse, that these two components have the same behaviour in the calculation of the probability of being in the case of malignant tumour).
This approach has not lead to good performances. A similar attempt was done with absorption data but with similar results in terms of performance of classifiers. It's important to note that it's preferable to classify with optical derived parameters because they're more physically meaningful. It was decided to try using more variables in an attempt to boost the performance of a hypothetical classifier.

3.2 Classification with more variables

The next step is the attempt to classify with data contained in Dataset 2 using other available variables for every subject in order to increase the power of the classifier. The most important available variables on the basis of the hypothetical relationship with the type of tumour (at the level of general medical knowledge) are:

- age;
- weight;
- height;
- menopausal state;
- use of oral contraceptives (OC);
- number of children;
- recent or present use of hormone replacement therapy (HRT);

- recent or present use of Tamoxifen (TAM).

Attention is paid in particular on age and menopausal state because a link between them and constituents concentration has been verified in previous studies (Figure 3.3). The

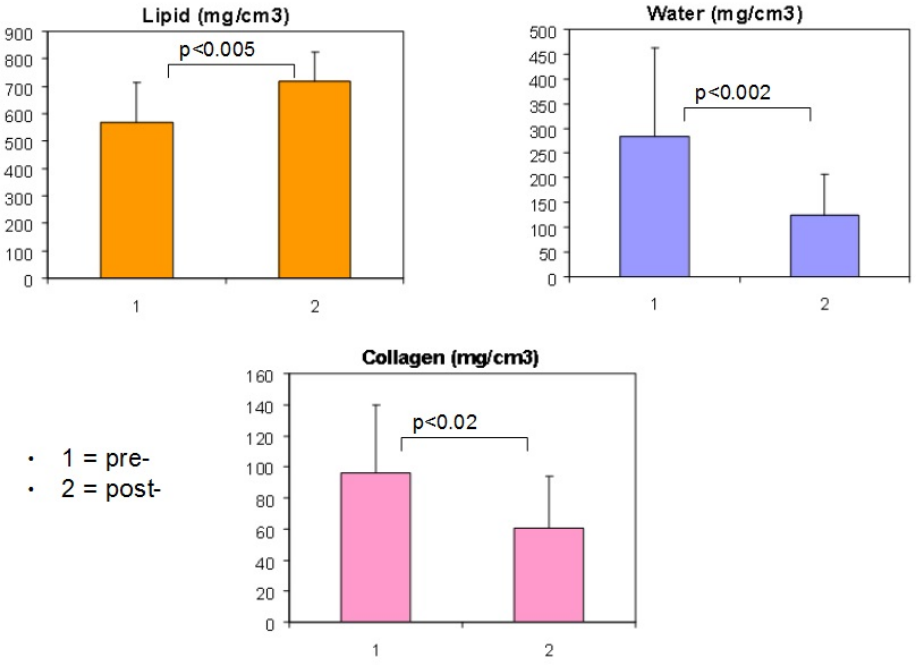

Figure 3.3: Boxplots with Lipid/Water/Collagen in pre- and post- menopausal state in subjects available in previous studies. This studies take into account 74 subjects with Benign and they've been subjected to optical spectroscopy in order to fit the concentrations of the main constituents. The p-values shown are referred to the Wilcoxon-tests.

model (3.2) has been fitted considering only the concentrations, the age and the menopausal state:

$$logit\,(p_i) = \alpha_0 + \alpha_1\,[HbO2]_i + \alpha_2\,[Water]_i + \alpha_3\,[Age]_i \tag{3.2}$$

where p_i is the probability for the i-th subject to belong to the Malignant class.
The menopausal state wasn't significant in this model (p-value very high) probably because of the strong correlation with Age. Given that the age seems to have greater significance then it was decided to use this variable instead of menopausal state (substantially it's been applied a backward selection in which the only excluded variable is the menopausal state). The values of the coefficients are reported below:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -6.330783   1.403007  -4.512  6.41e-06***
HbO2           0.020636   0.008857   2.330    0.0198*
Water          0.017175   0.009076   1.892    0.0584.
Age            0.121003   0.027281   4.435  9.19e-06.
---
Null deviance  : 144.52 on 105 degrees of freedom
Residual deviance: 105.95 on 102 degrees of freedom
AIC: 113.95
```

The performance of this model is very better than previous ones:

- Sensitivity= 0.852459;
- Specificity= 0.8;
- $1 - TotalErrorRate = 0.8301887$.

It's important to note that the values of the coefficients (α_i) are positive, so we can argue that probably the three factors considered in this model are risk factors in favour of the malignancy of the tumour in a subject.
In order to fit a model with interaction, it's crucial to study the correlation matrix of numerical ones (Figure 3.4).
It's possible to underline the following strong correlations:

	HHb	HbO2	Lipid	Water	Collagen	age	BMI
HHb	1.000000000	-0.4579506	0.1423634	0.53613904	-0.23826531	-0.006117052	0.0285604
HbO2	-0.457950601	1.0000000	-0.4452307	-0.49435734	0.48227042	-0.175299255	-0.1993622
Lipid	0.142363424	-0.4452307	1.0000000	0.04363180	-0.31491405	0.128268852	0.1628370
Water	0.536139041	-0.4943573	0.0436318	1.00000000	0.03722506	0.208902060	0.1005209
Collagen	-0.238265310	0.4822704	-0.3149141	0.03722506	1.00000000	0.034936256	0.1105714
age	-0.006117052	-0.1752993	0.1282689	0.20890206	0.03493626	1.000000000	0.3106935
BMI	0.028560399	-0.1993622	0.1628370	0.10052094	0.11057141	0.310693529	1.0000000

Figure 3.4: Correlation matrix for the numerical variables

- **HHb-Water** and **HbO_2-Water**: Figure 3.5 shows the plot of these correlated variables. The regression line of the first graphic is constituted by the intercept and slope coefficients shown below synthetically:

  ```
  Coefficients: Estimate   Std.Error   t value  Pr(>|t|)
  (Intercept)   2.9600     1.54226     1.921    0.0527.
  Water         0.25896    0.04119     6.288    8.99e-09 ***
  ---
  ```

  The intercept and slope for the second graphic are:

```
Coefficients: Estimate   Std.Error   t value  Pr(>|t|)
(Intercept)   22.85877    3.10572     7.36    5.72e-11 ***
HbO2         -0.46269    0.08218    -5.63    1.72e-07 ***
---
```

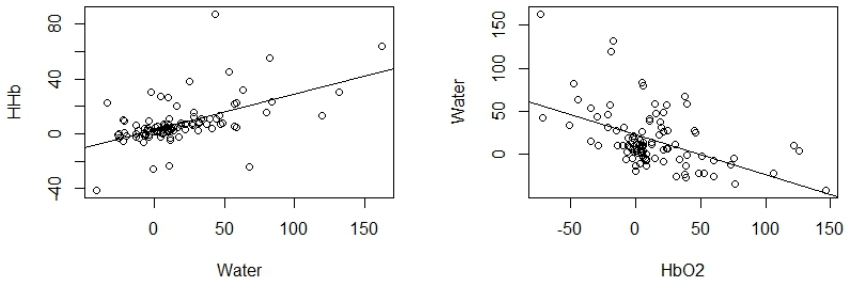

Figure 3.5: Plot of Water vs. HHb and Water vs. HbO_2 with regression lines

- **HbO2-Collagen**: Figure 3.6 shows the plot of these correlated variables. The inter-

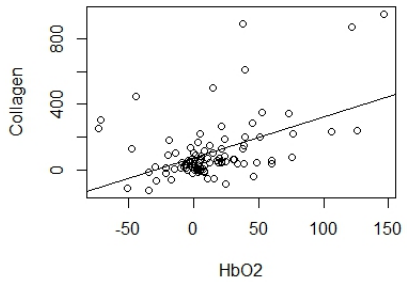

Figure 3.6: Plot of Collagen vs. HbO_2 with regression line

cept and slope for the graphic are:

```
Coefficients: Estimate   Std.Error   t value  Pr(>|t|)
(Intercept)   71.5981    17.4283     4.108    8.28e-05 ***
HbO2           2.5134     0.4612     5.450    3.76e-07 ***
---
```

The next objective was to test whether other dichotomous variables could have a relationship with the differentiation between benign and malignant tumours:

- present or recent use of **hormone replacement therapy (HRT)**: the studies con-

ducted in [6] assume that the incidence of breast cancer, all histologic types combined, was increased by 60% to 85% in recent long-term users of HRT, whether estrogen alone or estrogen plus progestin. Longer use of HRT (odds ratio [OR], 3.07 for 57 months or more; 95% confidence interval [CI], 1.55-6.06) and current use of combination therapy (OR, 3.91; 95% CI, 2.05-7.44) were associated with increased risk of lobular breast cancer. Long-term HRT use was associated with a 50% increase in nonlobular cancer (OR, 1.52 for 57 months or more; 95% CI, 1.01-2.29).
Now what we have to check is the dependence between two dichotomous variables: one that shows me the presence of a malignant tumour and one that shows me the presence of HRT. For this reason it is useful to calculate the OR with its confidence interval.
The calculation of confidence intervals uses the approach expressed by Woolf [22]:

$$IC\left(lnOR\right) = \left[ln\hat{O}R \pm z_{\alpha/2}\sqrt{Var\left(lnOR\right)}\right] \tag{3.3}$$

where:

- $ln\hat{O}R = \frac{a \cdot d}{b \cdot c}$
- $Var\left(lnOR\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$
- a=Benign subjects with use of HRT;
- b=Malignant subjects with use of HRT;
- c=Benign subjects without use of HRT;
- d=Malignant subjects without use of HRT.

In case of null values for a, b, c or d is possible to use the Gart approximation for the variance and the logarithm of the OR:

- $Var\left(lnOR\right) = \frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}$
- $ln\hat{O}R = \frac{(a+0.5)\cdot(d+0.5)}{(b+0.5)\cdot(c+0.5)}$

The estimated IC is then (considering $\alpha = 0.05$)

$$IC\left(lnOR\right) = [-2.734248; 1.124546]$$

Also no relationship with constituents was found.

- Present or recent use of **oral contraceptives (OC)**: most of the current knowledge on the risk of breast cancer associated with the use of OC results from a collaborative reanalysis of 54 epidemiological studies including a total of 53 297 cases of breast cancer [23]. This re-analysis has shown that the use of estrogen-progestin OC, in current users and those who have quit for 10 years or less from the suspension is

associated with a small increased risk of breast cancer (relative risk, RR= 1.24). The duration of use, therefore, as the dosage and formulation type seem to have little effect on the risk.
The estimated IC for this variable (in relation to the presence of malignant or benign tumour) is

$$IC\,(lnOR) = [0.05545144; 1.41887669] \tag{3.4}$$

Also there was a significant difference between the distributions of Collagen (expressed in terms of the difference between healthy tissue and diseased) between subjects with and subjects without taking OC (Figure 3.7).
The p-value of the relative t-test was < 0.001.

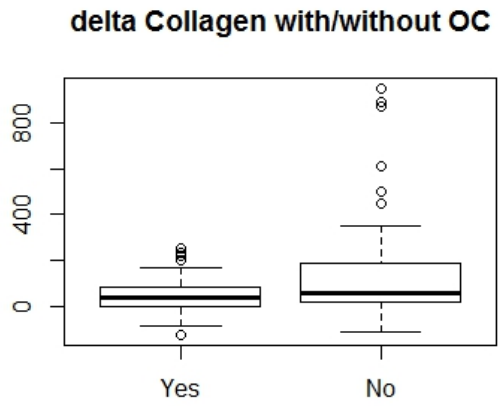

Figure 3.7: Values of Collagen in subjects with/without OC

- **Use of tamoxifen**: tamoxifen is a cancer medication taken orally, and belonging to the family of selective estrogen receptor modulators. This medicine inhibits the effects of estrogen, the female hormone and thus nullifying the effects of estrogen-receptor binding to DNA. This is useful because, often, the cancer cells of the breast cancer do benefit from these hormones.
  The estimated IC for this variable (in relation to the presence of malignant or benign tumour) is

  $$IC\,(lnOR) = [0.1399434; 5.0912428] \tag{3.5}$$

  But no relationship with constituents was found.

- **familiarity**: previous studies treated this topic [14]. Compared with women without a family history of breast cancer, women who had an affected first-degree relative had

a relative risk of 2.3; women with an affected second-degree relative had a relative risk of 1.5; and women with both an affected mother and sister had a relative risk of 14. The risk of breast cancer for a woman was higher if her first-degree relative had unilateral rather than bilateral breast cancer or had breast cancer detected at a younger rather than older age.
The estimated IC for this variable (in relation to the presence of malignant or benign tumour) is

$$IC\,(lnOR) = [-0.1985094; 1.2956413] \tag{3.6}$$

But a relationship is noted between this variable and lipids and water (Figure 3.8) Both with a p-value < 0.1.

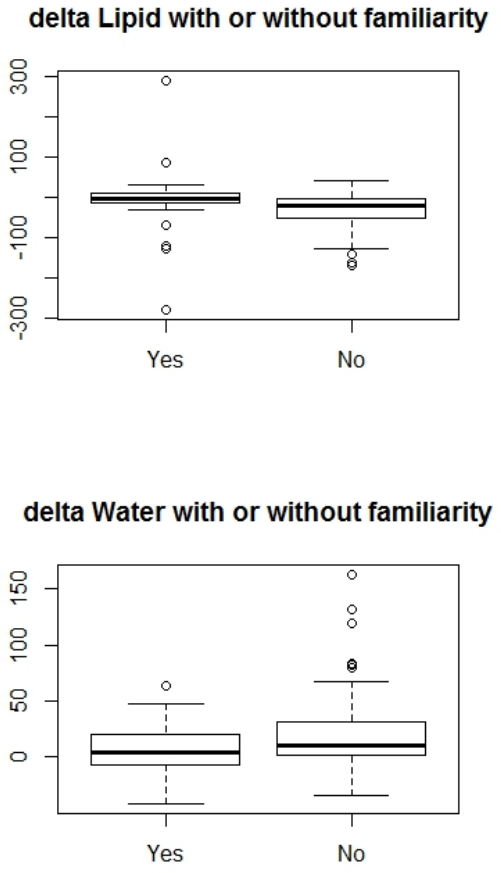

Figure 3.8: Plots of Lipid and Water with and without the familiarity factor

Subsequently the contingency tables were analysed, and one relationship have been highlighted based on the confidence intervals of the odds ratio:

- **HRT-TAMOXIFEN**: the confidence interval of the logarithm of the OR is

$$IC\left(lnOR\right) = [-4.2760129; -0.5997883] \tag{3.7}$$

Considering these statements, the next step is to find a model that considers all the variables (including the BMI as a combination of Weight and height) and the interactions between them. To do this we have exploited the VIFs as indices of collinearity during the work of simplification of the model and choice of the significant variables.
One way to determine if there is multicollinearity is to calculate the so-called variance inflation factors (VIF). In fact, when there is multicollinearity, the estimated variance of the j-th regression coefficient can be written as:

$$\hat{Var}(\beta_j) = \frac{S^2}{(n-1)\,S_j^2}\frac{1}{1-R_j^2} \tag{3.8}$$

where

- S^2 is the variance of the error;
- S_j^2 is the variance of x_j;
- R_j^2 is the coefficient of determination calculated by the regression of x_j over the other variables.

The quantity

$$VIF_j = \frac{1}{1-R_j^2} \tag{3.9}$$

is called inflation factor of variance. The VIF are used as measures of multicollinearity, because the square root of the VIF indicates how much the confidence interval, built on each of the regression coefficients β_j is larger compared to the situation of uncorrelated data. In particular, therefore, the variables that are most suspects to cause the phenomenon of multicollinearity are those that present the highest VIF. At the same time we call Tolerance the following expression:

$$Tolerance_j = 1 - R_j^2 \tag{3.10}$$

and its meaning is easily derivable from the previous definition.
With this consideration the several models were obtained: the model with the best performances in terms of total error rate is the following one:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1\left[Collagen\right]_i + \alpha_2\left[Age\right]_i + \alpha_3\left[Familiarity\right]_i \tag{3.11}$$

The performance of this model is:

- Sensitivity= 0.8448276;
- Specificity= 0.7317073;
- $1 - TotalErrorRate = 0.79$.

The summary of the logistic regression model is reported below:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -6.522609   1.496842  -4.358   1.32e-05 ***
Collagen       0.003892   0.001904   2.045   0.0409 *
Age            0.135838   0.030994   4.383   1.17e-05 ***
Familiarity   -1.256729   0.616182  -2.040   0.0414 *
---
Null deviance  : 136.058 on 98 degrees of freedom
Residual deviance: 93.863 on 96 degrees of freedom
AIC: 101.86
```

This is a particularly interesting model because it shows us that familiarity is an important factor for the classification but its meaning is the opposite of what we expected: the negative sign indicates that familiarity implies a reduction in the likelihood of a malignant tumour in favour of a benign one.
The positive sign of Collagen and Age was predictable because you can easily reconnect to the model previously calculated; Age was already with a positive sign, while Collagen is positively correlated with HbO_2.
One last valid model was identified using principal coordinates (also considering possible interactions):

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[Comp.1\right]_i + \alpha_2 \left[Comp.4\right]_i + \alpha_3 \left[Age\right]_i + \alpha_4 \left[Familiarity\right]_i \qquad (3.12)$$

The performance of this model is:

- Sensitivity= 0.8103448;
- Specificity= 0.7804878;
- $1 - TotalErrorRate = 0.8$.

The summary of the logistic regression model is:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -6.436505   1.542061  -4.174   2.99e-05 ***
Comp.1        -0.004287   0.001982  -2.163   0.0305 *
Comp.4        -0.024216   0.014000  -1.730   0.0837 .
Age            0.142012   0.031888   4.453   8.45e-06 ***
Familiarity   -1.024660   0.614078  -1.669   0.0952 .
---
Null deviance  : 136.058 on 99 degrees of freedom
Residual deviance: 90.539 on 95 degrees of freedom
AIC: 100.54
```

This model is really similar to the model 3.11, because the first component is substantially the Collagen. The difference coincides with the fourth main component. The anomaly of this model is the negative sign of the coefficient on the first principal component, in fact, as said before we would expect a positive sign. Probably this is due to the fact that the directions are "contaminated" by the other optically derived variables, and this can result in a change of the value of the coefficient which in this case also leads to a change of sign. In general, however, this model can be considered acceptable and meaningful even if in case of equal performances, the models previously shown are preferable.

3.3 Further observations about classification

After the previous considerations, it's important to verify the robustness of the classifiers described before and to treat the problem of double observations for some patients. For the latter problem a solution has been proposed: the concentration values obtained at the two angles are mediated so that we consider the average estimated concentration. It's important to see the differences with previous results and look for possible reasons.
For this reason we've repeated the previous analysis with Dataset 3: PCA for constituent concentrations and logistic regression. The results obtained trying to classify with PCA directions wasn't good. The best model in terms of performances (Total Error Rate) was again model (3.11), with similar values of coefficients and performances. But a second model was relevant:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[HbO2\right]_i + \alpha_2 \left[Age\right]_i + \alpha_3 \left[Familiarity\right]_i + \alpha_4 \left[HRT\right]_i + \alpha_5 \left[HbO2\right]_i \left[Water\right]_i \tag{3.13}$$

and the following summary shows the main descriptors of model 3.13 fitted to data:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -9.162809   2.368992  -3.868   1.10e-04 ***
HbO2           0.024093   0.011460   2.102   0.0355 *
Age            0.191564   0.049586   3.863   1.12e-04 ***
Familiarity   -1.523680   0.841273  -1.811   0.0701 .
HRT           -2.721635   1.628698  -1.671   0.0947 .
HbO2:Water     0.000691   0.000357  -1.935   0.0530 .
---
Null deviance  : 97.283 on 70 degrees of freedom
Residual deviance: 62.435 on 65 degrees of freedom
AIC: 74.435
```

The measured performances are:

- Sensitivity= 0.8;
- Specificity= 0.8064516;
- $1 - TotalErrorRate$ is about 0.8.

Also in this case the familiarity is a negative factor in relation to the probability of malignant tumour.
The most obvious difference with the models studied above is the variable HRT (hormone replacement therapy), which is significant (with negative sign) in this model.
In general, however, we can say that there are no substantial differences between the results obtained with Dataset 2 and Dataset 3 in terms of the significance of the models obtained.
A useful alternative to the logistic regression models is to alter the fitting process itself so that potential overfitting of a given model comes at a price. A penalty can be introduced into the loss function to be optimized. In particular these methods are particularly useful when there is high collinearity between the regressors. In part because this approach has wide applicability, it is worth our attention now.

3.4 LASSO, Ridge and elastic net Regression

Suppose that we have data $\left(\mathbf{x}^i, y_i\right)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^T$ are the predictor variables and y_i are the response. As in the usual regression set-up, we assume that the observations are independent. We assume now (without loss of generality) that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0$ and $\sum_i x_{ij}^2/N = 1$.
Let $\hat{\boldsymbol{\beta}}$ the linear lasso estimate of the vector of the parameters, it is defined by [19]:

$$\hat{\boldsymbol{\beta}} = argmin \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 \right\}$$
$$\text{subject to } \sum_j |\beta_j| \leq t$$

Here $t \geq 0$ is a tuning parameter that controls the amount of shrinkage that is applied to the estimates.
The LASSO can be applied to various models. Consider any model indexed by a vector parameter $\boldsymbol{\beta}$, for which estimation is carried out by maximization of a function $l(\boldsymbol{\beta})$; this may be a log-likelihood function or some other measure of fit. To apply the LASSO, we maximize $l(\boldsymbol{\beta})$ under the constraint $\sum |\beta_j| \leq t$. One of the possible application is then the logistic regression, and for this reason we want to apply the LASSO to our previous model. Instead Ridge regression penalizes the size of the regression coefficients in L_2-norm: specifically, the ridge regression estimate $\hat{\boldsymbol{\beta}}$ is defined as the value of $\boldsymbol{\beta}$ that minimizes

$$\sum_i \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.14}$$

Applying the ridge regression penalty has the effect of shrinking the estimates toward zero, it's been demonstrate that the benefits of ridge regression are most striking in the presence of multicollinearity.

- ridge regression achieves its better prediction performance through a bias-variance trade-off. However it can't produce a parsimonious model because it always keeps all the predictors for the model;
- for usual N>p situations, if there are high correlations between predictors, it has been empirically observed that the predictions of the LASSO is dominated by ridge regression.

For this reason it's usual to recur to the Elastic Net Problem [9]: we use the standardization previously presented (for the LASSO) for solving the next problem:

$$\min_{(\beta_0, \boldsymbol{\beta})} \left[\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \boldsymbol{\beta} \right)^2 + \lambda P_\alpha (\boldsymbol{\beta}) \right] \tag{3.15}$$

where

$$P_\alpha (\boldsymbol{\beta}) = (1-\alpha) \tfrac{1}{2} ||\boldsymbol{\beta}||_{l_2}^2 + \alpha ||\boldsymbol{\beta}||_{l_1}$$
$$= \sum_{j=1}^{p} \left[\tfrac{1}{2} (1-\alpha) \beta_j + \alpha |\beta_j| \right]$$

is the elastic-net penalty. P_α is a compromise between the ridge regression penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$). This penalty is particularly useful in the $p >> N$ situation, or any situation where there are many correlated predictor variables.

A similar argument is made when you need to apply a logistic regression: in this case denote by G the response variable, taking values in 0,1, here we fit the model by regularized maximum (binomial) likelihood. Let $p(x_i) = Pr(G = 1|x_i)$ be the probability for the i-th observation at a particular value for the parameter $(\beta_0, \boldsymbol{\beta})$, it's possible to state that

- $p(x_i) = Pr(G = 1|x_i) = \frac{1}{1+\exp\left(-\left(\beta_0 + x_i^T\boldsymbol{\beta}\right)\right)}$;
- $1 - p(x_i) = Pr(G = 0|x_i) = \frac{1}{1+\exp\left(\beta_0 + x_i^T\boldsymbol{\beta}\right)}$.

Alternatively it implies

- $\log\left(\frac{p}{1-p}\right) = \beta_0 + x^T\boldsymbol{\beta}$

Then we maximize the penalized log-likelihood

- $\max_{(\beta_0,\boldsymbol{\beta})}\left[\frac{1}{N}\sum\left\{I(g_i = 1)\log p(x_i) + I(g_i = 0)\log\left(1 - p(x_i)\right)\right\} - \lambda P_\alpha(\boldsymbol{\beta})\right]$

Denoting $y_i = I(g_i = 1)$, the logarithmic part of the penalized log-likelihood can be written in the more explicit form

$$l(\beta_0, \boldsymbol{\beta}) = \frac{1}{N}\sum y_i \cdot \left(\beta_0 + x_i^T\boldsymbol{\beta}\right) - \log\left(1 + \exp^{\left(\beta_0 + x_i^T\boldsymbol{\beta}\right)}\right) \tag{3.16}$$

We form a quadratic approximation to the log-likelihood (Taylor expansion about current estimates $\hat{\beta}_0, \hat{\boldsymbol{\beta}}$), which is

$$l_Q(\beta_0, \boldsymbol{\beta}) = -\frac{1}{2N}\sum_{i=1}^{N} w_i\left(z_i - \beta_0 - x_i^T\boldsymbol{\beta}\right)^2 + C\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^2 \tag{3.17}$$

where

- $z_i = \hat{\beta}_0 + x_i^T\boldsymbol{\beta} + \frac{y_i - \hat{p}(x_i)}{\hat{p}(x_i)(1-\hat{p}(x_i))}$
- $w_i = \hat{p}(x_i)\left(1 - \hat{p}(x_i)\right)$

The we use coordinate descendent [9] to solve the penalized weighted least-squares problem

$$\min_{(\beta_0,\boldsymbol{\beta})}\left\{-l_Q(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})\right\} \tag{3.18}$$

This type of analysis has been applied on Dataset 2 and 3 different values of α have been used:

- $\alpha = 1$ or LASSO regression;
- $\alpha = 0.5$ or elastic net regression;
- $\alpha = 0$ or ridge regression.

The comparison of the results of these 3 analysis allows us to determine the most important variable for the logistic regression model: in particular we pay attention to the first result, the LASSO is a sort of variable selection algorithm.
It is natural to understand that with the decrease of λ, also the effect of the penalty decreases getting closer to the least square regression. It's crucial to point out that in the case of LASSO regression the decrease of λ causes an increase of variables whose coefficient becomes significant ($\beta_i \neq 0$). Now the partial results of the LASSO regression are used to make a selection of significant variables, in particular we use the first 8 coefficients with significant value:

- HbO_2;
- Collagen;
- Age;
- Familiarity;
- Oral Contraceptive;
- HRT;
- BMI;
- HHb*Water.

These variables have been used in a stepwise logistic regression model with backward selection based on the evaluation of collinearity between variables (with VIFs) and significance of the coefficients (regression p-values). The result is the following model:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[HbO_2\right]_i + \alpha_2 \left[Age\right]_i + \alpha_3 \left[HHb\right]_i \left[Water\right]_i \tag{3.19}$$

Then we tried to improve the performances of this model (in terms of Total Error Rate) by adding other variables that previously were significant. The model found was:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[HbO_2\right]_i + \alpha_2 \left[Age\right]_i + \alpha_3 \left[HHb\right]_i \left[Water\right]_i + \alpha_4 \left[Age\right]_i \left[Familiarity\right]_i \tag{3.20}$$

The summary of the logistic regression model (3.20) is reported below:

```
Coefficients:   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)    -8.376137   1.776073  -4.716   2.40e-06 ***
HbO2            0.029665   0.011217   2.645   0.00818 **
Age             0.166732   0.035702   4.670   3.01e-06 ***
HHb:Water       0.001011   0.000405   2.494   0.0126 *
Age:Familiarity-0.022136   0.011940  -1.854   0.06375 .
---
Null deviance  : 136.058 on 99 degrees of freedom
Residual deviance: 85.393 on 95 degrees of freedom
AIC: 95.393
```

The sensitivity was 0.8275862, the specificity was 0.8095238 and $1-TotalErrorRate = 0.82$. The performances are in general similar to those of logistic regression models, but it is preferred for convenience to use the previous models , in particular models (3.2), (3.11) and (3.12), because they allow us to hypothesize in a very intuitive way which are the significant variables and their effect on the probability of having malignant disease or benign disease.

3.5 Other methods

In addition to logistic regression, you can also use other types of classifiers. In particular the two classifiers that are less susceptible to non-normality of the data are the **Fisher Linear Discriminant Analysis** and **CARTs**.
The trees could be used in more developed classifiers, the random forest and the AdaBoost algorithms, in which the trees are basic classifiers and they are used to built a more complex model.

3.5.1 Fisher Linear Discriminant Analysis

Firstly we describe the Fisher Linear Discriminant Analysis: suppose two classes of observations have means μ_0, μ_1 and covariances Σ_0 and Σ_1.
Then the linear combination of features $w \cdot x$ will have means $w \cdot \mu_i$ and covariances $w^T \Sigma_i w$.
It can be shown that the maximum separation between the classes occurs when

$$w \propto (\Sigma_0 + \Sigma_1)^{-1} (\mu_0 - \mu_1) \tag{3.21}$$

Generally the data points to be discriminated are projected onto w, then the threshold that best separates the data is chosen from analysis of the one-dimensional distribution.

In our case the continuous variables have been selected (we are referring to Dataset 2 for th whole sections 3.5)and the discriminant direction was:

Collagen	0.0304761917
HHb	0.0304761917
HbO_2	0.0184723287
Lipid	-0.0007622592
Water	0.0073178564
Age	0.1432901696
BMI	-0.0463709823

Figure 3.9 shows the projections of the points on the discriminant direction.
The observations with projected values smaller than the reference value are classified as

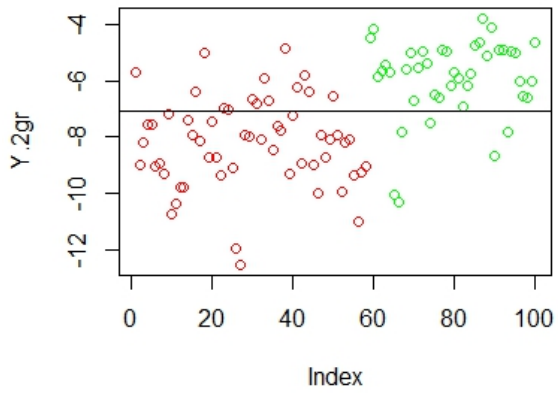

Figure 3.9: Projection of Malignant and Benign points on the discriminant direction

malignant, in the opposite case are classified as benign. The sensitivity of this classifier is 0.7586207 , the specificity is 0.8571429 and $1 - TotalErrorRate = 0.81$.

3.5.2 CART

The second alternative method used were the **prediction trees**: these have two varieties, regression trees and classification trees [21].
We want to predict a response or class Y from inputs $X_1, X_2, \ldots, X_p$. We do this by growing a binary tree. At each internal node in the tree, we apply a test to one of the

inputs. Depending on the outcome of the test, we go to either the left or the right sub-branch of the tree. Eventually we come to a leaf node, where we make a prediction. There are several advantages to this:

- Making predictions is fast (no complicated calculations);
- it's easy to understand which variables are important in making the prediction (look at the tree);
- there are fast, reliable algorithms to learn these trees.

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for measuring "best". These generally measure the homogeneity of the target variable within the subsets, they are called impurity index. Chosen the index of impurity, split will naturally be better defined as that, of all the possible, will generate the maximum decrease of the index.
The most important indices are:

- **Gini impurity**: Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

$$GINI = p - p^2 \tag{3.22}$$

where p is the probability of being a Malignant case. Figure 3.10 reports the equation (3.22). In two-class problems the Gini index can also be interpreted as a kind of

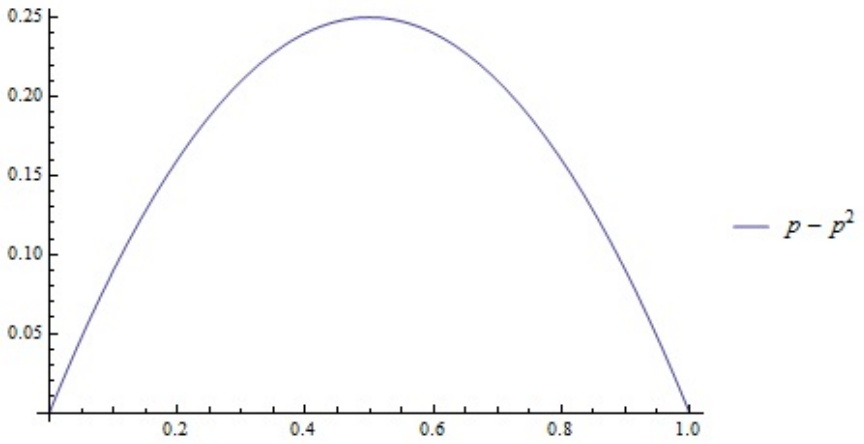

Figure 3.10: Gini index value for the variation of p, where $p = p(A|t)$

variance of the node. In fact, if we interpret the node t as a sample of N (t) realizations

of a Bernoulli process with mean and variance unknown and assign the value 1 for class A and 0 to B, we have that the average of this process will coincide with $p(A|t)$ and the variance with $p(A|t)(1-p(A|t)) = p(B|t)(1-p(B|t))$
Split according to the Gini index means then divide into two parts the node looking to minimize the average variance of child nodes.

- **Information gain**: information gain is based on the concept of entropy from information theory:

$$INF = -p\log p - (1-p)\log(1-p) \tag{3.23}$$

This method arises from an estimate of maximum likelihood data of the sample reached the node t: each leaf is interpreted as the result of a binomial process, with N(t) extractions, with parameters p_A, p_B to estimate and realizations $[N_A, N_B]$.

$$P(n_A(t) = N_A(t), n_B(t) = N_B(t)) = \frac{N(t)!}{N_A(t)!N_B(t)!} p_A^{N_A(t)} p_B^{N_B(t)} \tag{3.24}$$

The maximum is:

$$P = \left(\frac{N_A(t)}{N(t)}\right)^{N_A(t)} \left(\frac{N_B(t)}{N(t)}\right)^{N_B(t)} \tag{3.25}$$

To make it more easy to handle it is usual to calculate the logarithm. To make it an index of impurity just change the sign and make it independent from the size N(t) of the node by dividing by N(t). It is obtained in this way an index of impurity whose theoretical value coincides exactly with the definition of Entropy:

$$INF \approx -\log\left[\left(\frac{N_A(t)}{N(t)}\right)^{\frac{N_A(t)}{N(t)}} \left(\frac{N_B(t)}{N(t)}\right)^{\frac{N_B(t)}{N(t)}}\right] \tag{3.26}$$

$$INF = -\log\left[p(A|t)^{p(A|t)}\, p(B|t)^{p(B|t)}\right] \tag{3.27}$$

The function expressed in equation (3.27) is reported in Figure 3.11. This index is still used, has performance similar to that of Gini and also generates trees very similar. The latter, however, is usually preferred being more simple, computationally slightly less expensive and numerically more stable.

The indices just described could be extended in the multi-case (with more than 2 classes). It is easy to verify that these three indices verify properties that guarantee the quality of a generic index of impurity denoted $\phi(p)$ [3]:

- $\phi(p) \geq 0$: the index is always positive or zero;
- $\min_p \phi(p) = \phi(0) = \phi(1)$: the index reaches its minimum in the case where the population is totally pure, that is when there is only one class (the best condition for classification);

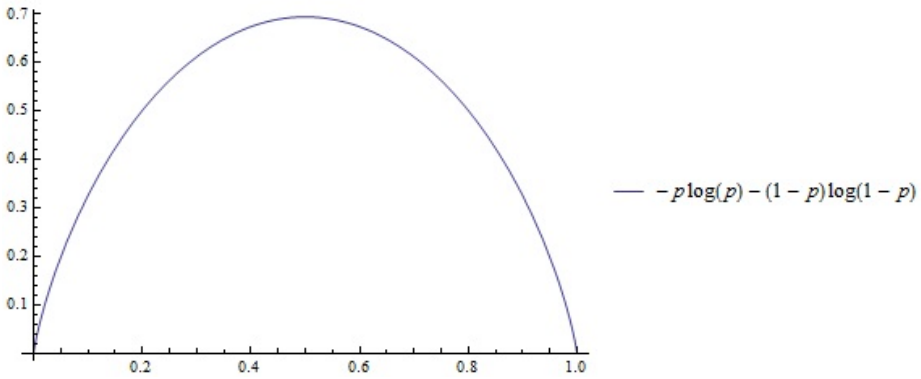

Figure 3.11: Information gain for the variation of p, where $p = p(A|t)$

- $\max_p \phi(p) = \phi(1/2)$, the index reaches its maximum when the two classes are equally divided (the worst condition for classification);
- $\phi(p) = \phi(1-p)$: the index is symmetrical with respect to the exchange of the probabilities of the two classes;
- $\phi''(p) < 0$: the index is a convex function. Please note that in general, when it fails the convexity, even the non-growth is not guaranteed, so that the existence of non-convex indices allows some splits that lead to an increase of the impurity.

Now we describe briefly the problem of the optimal split at the nodes: we introduce the probability of a single data of belonging to the one of the child nodes conditioned to the membership at the node t.

- $p_L = p(t_L|t) = p(t_L)/p(t)$;
- $p_R = p(t_R|t) = p(t_R)/p(t)$.

We define $i(t)$ the impurity index of a node t, it's possible to define the average impurity of child nodes:

- $i(t_R, t_L) = p_L i(t_L) + p_R i(t_R)$

The decrease of impurity is then:

- $\Delta i(t, t_R, t_L) = i(t) - i(t_L, t_R)$

At this point we will choose the split that minimizes the impurity of child nodes, or equivalently, but more properly in view of the construction of the tree, one that will maximize the decrease impurity.
We denote by:

- N=number of observations;
- N_j=number of observations in j-th class;
- $N(t)$=number of data in the node;
- $N_j(t)$=number of data in the node belonging to the j-th class.

Initial estimates that may come to mind of $p(j|t)$ and $p(t)$ are obviously as follow:

- $p(j|t) = \frac{N_j(t)}{N(t)}$
- $p(t) = \frac{N(t)}{N}$

These estimates are often used unwisely, in fact they are correct only when the distribution of the population from which the data set to coincide with the distribution of the real population. But this fact is not always true, so it is more correct to use estimates whose correctness is independent verification of this hypothesis or not:

- $p(j|t) \approx \frac{\frac{N_j(t)}{N_j}p(j)}{\sum_j \frac{N_j(t)}{N_j}p(j)}$
- $p(t) \approx \sum_j \frac{N_j(t)}{N_j}p(j)$

When the relative frequencies $\frac{N_j(t)}{N}$ within the sample reflect well the probability $p(j)$, i.e. $p(j) \approx \frac{N_j}{N}$ the two estimates are coincident.
Obviously, in the absence of further information than that given to us by the data, the hypothesis $p(j) \approx \frac{N_j}{N}$ is the most plausible and then what will be generally applied, and this is our case: we couldn't have further informations about the probabilities of malignant and benign tumours.
Now some of developed CARTs are reported, based on different assumption.

Figure 3.12 and Figure 3.13 show the CARTs developed using all the variables and Gini/Information impurity index.
Since the age seems to be a particularly important factor even with CART, naturally the first split concerns precisely this variable. Figures 3.14 and 3.15 show the CARTs fitted without the age: the idea is to find the other important variables that could affect the classification.
An important observation is about the BMI: this variable was not significant in any model built previously with logistic regression models, but now it is the variable considered in the first split.
The second observation is more technical: the two types of impurity index lead to similar CARTs, and in the second case even lead to the same results. The biggest problem with

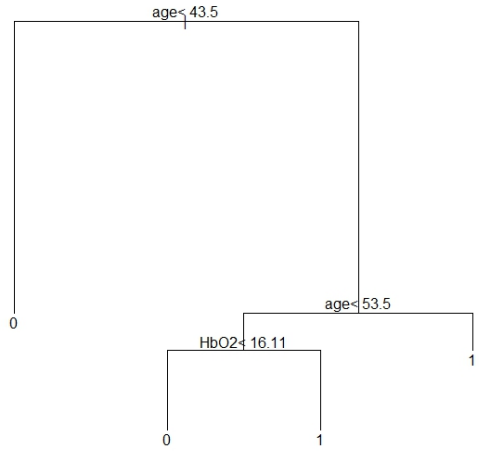

Figure 3.12: CART built considering all variables and Gini impurity

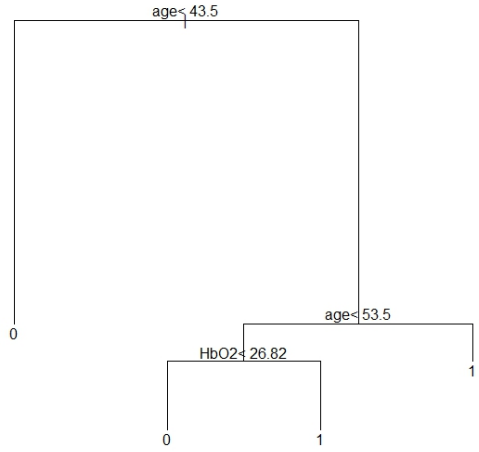

Figure 3.13: CART built considering all variables and Information gain

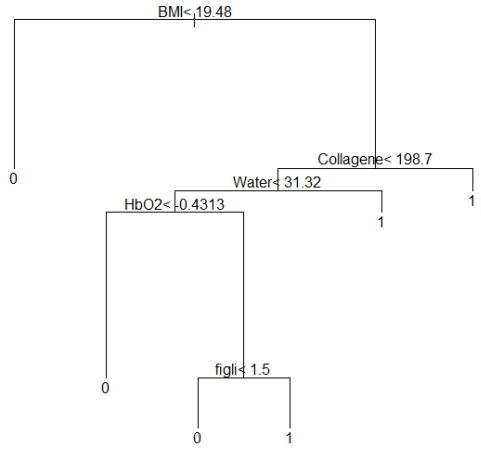

Figure 3.14: CART built considering all variables less the age and Gini impurity

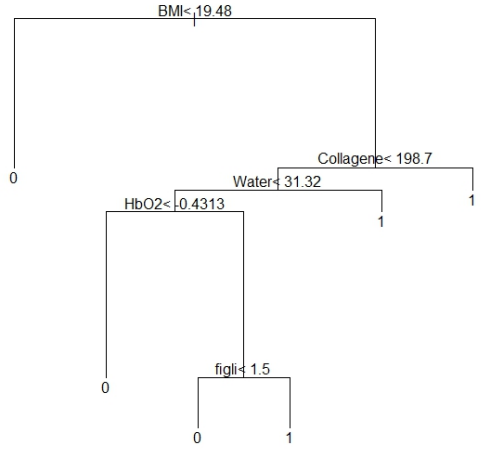

Figure 3.15: CART built considering all variables less the age and Information gain

this type of classifier consists in the performances, in fact, the global misclassification error is always below 30%. You may notice that the performance is much lower than the logistic regression models, for this reason you can try to improve this type of classifier with other techniques, such as random forest method or Boosting classifiers.

3.5.3 Random Forest

The forest structure is slightly different between classification and regression. Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class, that is the mode or the average of the classes of individual trees.
The introduction of random forests proper was first made in a paper by Leo Breiman.[4] This paper describes a method of building a forest of uncorrelated trees using a CART like procedure, combined with randomized node optimization and bagging.
Given a standard training set D of size n, bagging generates m new training sets D_i,each of size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i, this kind of sample is known as bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by averaging the output. Each tree is grown to the largest extent possible, there is no pruning. Then the training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.
Given a training set, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples, ie for every iteration:

- Sample, with replacement, n' training examples;
- Train a decision or regression tree.

After training, predictions can be made by averaging the predictions from all the individual regression trees or by taking the majority vote in the case of decision trees. The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging".
In the original paper on random forests, it was shown that the forest error rate depends on two things:

- The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate;
- The strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.

In addition, this paper combines several ingredients, which form the basis of the modern practice of random forests, in particular:

- using out-of-bag error as an estimate of the generalization error. Assume a method for constructing a classifier from any training set. Put each case left out in the construction of the k-th tree down the k-th tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case an observation k was out of bagging. The proportion of times that j is not equal to the true class of k averaged over all cases is the out of bag error estimate. This has proven to be unbiased in many tests. The use of this out of bag error is the reason because there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run;

- measuring variable importance through permutation. In every tree grown in the forest, put down the out of bag cases and count the number of votes cast for the correct class. Now randomly permute the values of variable m in the out of bag cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-m-permuted out of bag data from the number of votes for the correct class in the untouched out of bag data. The average of this number over all trees in the forest is the raw importance score for variable m. If the number of variables is very large, forests can be run once with all the variables, then run again using only the most important variables from the first run.
  Every time a split of a node is made on variable m, the Gini impurity criterion for the two descendent nodes is less than the parent node. Adding up the Gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

In our case it's been decided to use 500 trees for the random forest model, the number of variables randomly sampled as candidates at each split is 4 (the suggestion is to use about a third of the total available variables). The measured performances aren't very high, because the sensitivity is 0.8448276 and the specificity is 0.7380952. This method is more complex than those used up to this point, but it does not lead to a substantial increase of the performances.
We are very interested in the importance of the variables (Figure 3.16).
The most important variables are:

- Age;
- BMI;
- Collagen;
- HbO_2;

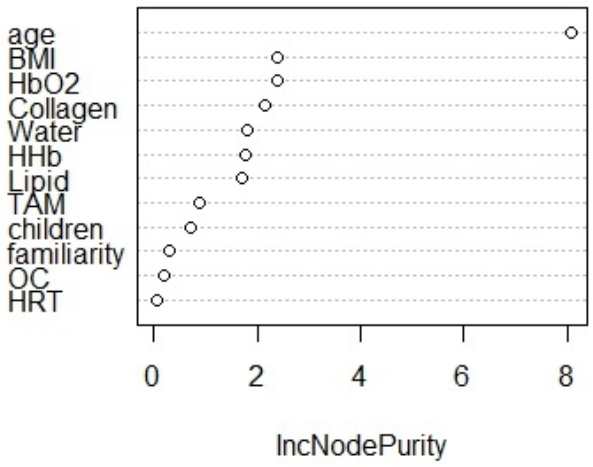

Figure 3.16: Variables Importance measured with random forest principle

- Water;
- HHb;
- Lipid.

The analysis is repeated with the most important variables, in order to improve the results for the most significant variables (for the random forest principles) and the results are shown in Figure 3.17.

A new logistic regression model has been fitted using the most important variables with all the possible interactions applying a backward variable selection and using the VIFs. The result is the following model:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[Age\right]_i + \alpha_2 \left[HbO_2\right]_i + \alpha_3 \left[Water\right]_i + \alpha_1 \left[HbO_2\right]_i \left[Water\right]_i \qquad (3.28)$$

The summary of model (3.28) is reported below:

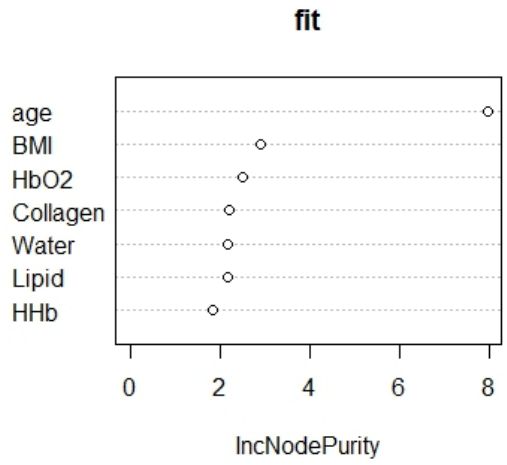

Figure 3.17: Variables Importance measured with random forest principle

```
Coefficients:   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)    -7.769485   1.693552  -4.588   4.48e-06 ***
HbO2            0.029256   0.010533   2.778   0.00548 **
Age             0.145910   0.032652   4.469   7.87e-06 ***
Water           0.019393   0.010449   1.856   0.06345 .
HbO2:Water     -0.000484   0.000281  -1.721   0.08518 .
---
Null deviance  : 136.058 on 99 degrees of freedom
Residual deviance: 90.926 on 95 degrees of freedom
AIC: 95.393
```

The sensitivity is 0.8448276, the specificity is 0.7380952 and $1 - TotalErrorRate = 0.79$, then the performance is similar to the models until now described.

3.5.4 Boosting

Another classification method recently developed is the boosting method. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data and then taking a weighted majority vote for the classifiers thus produced [10].
Boosting is a way of combining the performances of many weak classifiers to produce a powerful one. The most commonly used version of the AdaBoost procedure is the Discrete

AdaBoost [8].
We define $F(x) = \sum_{m=1}^{M} c_m f_m(x)$ where each $f_m(x)$ is a classifier producing values plus or minus 1 and c_m are constants; the corresponding prediction is $sign(F(x))$. The Adaboost procedure trains the classifiers $f_m(x)$ on weighted versions of the training sample, giving higher weight to cases that are currently misclassified. This is done for a sequence of weighted samples, and then the final classifier is defined to be a combination of the classifiers from each stage.
Much has been written about the success of AdaBoost in producing accurate classifiers. Many authors have explored the use of a tree-based classifier for $f_m(x)$ and have demonstrate that it consistently produces significantly lower error rates than a single decision tree. The algorithm could be described as reported in Algorithm 1.

Algorithm 1 Discrete AdaBoost

- Start with weights $w_i = 1/N$;
- Repeat for $m = 1, 2, \ldots, M$:
  - Fit the classifier $f_m(x) \in \{-1, 1\}$ using weights w_i on the training data;
  - Compute $err_m = E_w\left[1_{y \neq f_m(x)}\right]$, $c_m = \log\left((1 - err_m)/err_m\right)$;
  - Set $w_i \leftarrow w_i exp\left[c_m 1_{y \neq f_m(x_i)}\right]$ and renormalize so that $\sum_i w_i = 1$.
- Output the classifier $sign\left[\sum_{m=1}^{M} c_m f_m(x)\right]$.

E_w represent expectation over the data with weights $w = (w_1, w_2, \ldots, w_N)$. At each iteration, the algorithm increases the weights of the observations misclassified by $f_m(x)$ by a factor that depends on the weighted error.
This method has been subjected to a random sampling strategy at each iteration (bagging) like in the random forest method. This solution allows us to find an alternative to cross validation using only 50% of the entire sample for each iteration. For this reason the results are subject to stochasticity.
Of course, this percentage is very high but avoids to build a classifier extremely adherent to the reference sample. Also given that the average results (repeating several times the algorithm) are quite similar, with no significant variations, it was decided to use that percentage of the sample out of bag.
Then the algorithm was applied 20 times and the average performances are as follow: sensitivity of about 0.9 and specificity of about 0.86. As we can see the performances are very better than in the other cases because this classifier is more complex. The problem of this method is that the application of this classifier is particularly difficult in the medical field and it is extremely unintuitive.
For this reason (like with the random forest algorithm) a variables importance rank based on Gini impurity has been calculated (Figure 3.18).

A generalization of Discrete AdaBoost is the Real AdaBoost, in which the weak learner returns a class probability estimate $p_m(x) = \hat{P}_w(y = 1|x) \in [0, 1]$. The base classifier in

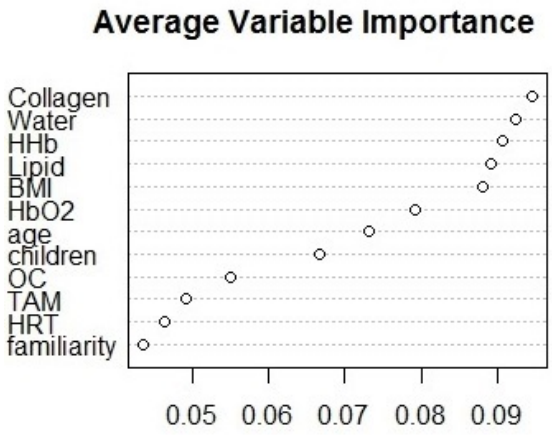

Figure 3.18: Average Variables Importance measured with Discrete AdaBoost method

Dscrete AdaBoost produces a classification rule $f_m(x) : D \to \{-1, 1\}$, where D is the domain of the predictive features x. The weak learner for the generalized AdaBoost produces a mapping $f_m(x)\, D \to \mathbb{R}$; the sign of $f_m(x)$ gives the classification, and $|f_m(x)|$ a measure of the "confidence" in the prediction. The contribution to the final classifier is half the logit-transform of this probability estimate. In this paper the AdaBoost is used in the special case where the weak learner is a decision tree. The real AdaBoost algorithm is reported in Algorithm 2.

The performances are: sensitivity of about 0.93 and specificity of about 0.87. Like in the

Algorithm 2 Real AdaBoost

- Start with weights $w_i = 1/N$;
- Repeat for $m = 1, 2, \ldots, M$:
  - Fit the classifier to obtain a class probability estimate $p_m(x) = \hat{P}_w(y = 1|x) \in [0, 1]$, using weights w_i;
  - Set $f_m(x) \leftarrow \frac{1}{2} \log p_m(x)/(1 - p_m(x))$;
  - Set $w_i \leftarrow w_i exp[-y_i f_m(x_i)]$ and renormalize so that $\sum_i w_i = 1$.
- Output the classifier $sign\left[\sum_{m=1}^{M} f_m(x)\right]$.

previous case it's possible to apply a bagging strategy to add stochasticity to the procedure. The variables importance is reported in Figure 3.19.

A latest algorithm belonging to Boosting methods is the so-called Gentle AdaBoost (Algorithm 3).

The main difference between Gentle and Real AdaBoost algorithm is how it uses its estimates of the weighted class probabilities to update the functions.

The average performances of the Gentle AdaBoost are: sensitivity of about 0.93 and speci-

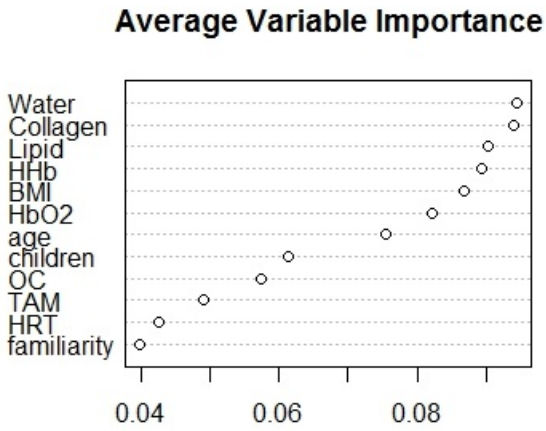

Figure 3.19: Average Variables Importance measured with Real AdaBoost method

Algorithm 3 Gentle AdaBoost

- Start with weights $w_i = 1/N$, $F(x) = 0$;
- Repeat for $m = 1, 2, \ldots, M$:
  - Fit the regression function $f_m(x)$ by weighted least-squares of y_i to x_i with weights w_i;
  - Update $F(x) \leftarrow F(x) + f_m(x)$;
  - Update $w_i \leftarrow w_i exp[-y_i f_m(x_i)]$ and renormalize.
- Output the classifier $sign\left[\sum_{m=1}^{M} f_m(x)\right]$.

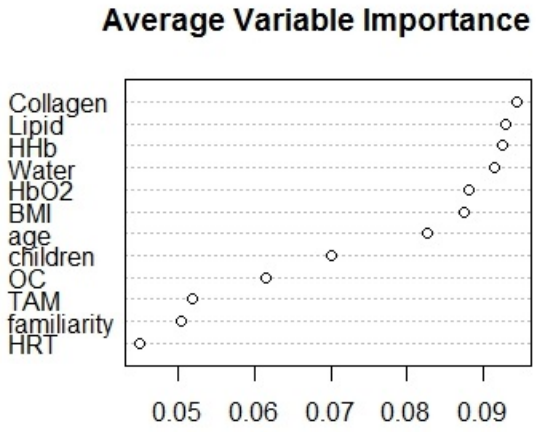

Figure 3.20: Average Variables Importance measured with Real Gentle method

ficity of about 0.88.
It's possible to demonstrate that the AdaBoost algorithms (Discrete and Real) can be interpreted as stagewise estimation procedures for fitting an additive logistic regression model. This type of regression models are characterized by the following idea: the classifier is a weighted sum of simpler classifiers.
For a two-class problem, an additive logistic model has the form

$$\log \frac{P(y=1|x)}{1-P(y=1|x)} = \sum_{m=1}^{M} f_m(x) \tag{3.29}$$

The monotone logit transformation on the left guarantees that for any values of $F(x) = \sum_{m=1}^{M} f_m(x) \in \mathbb{R}$, the probability estimates lie in $[0,1]$.
These models optimize an exponential criterion which to second order is equivalent to the binomial log-likelihood criterion.
Consider minimizing the criterion

$$J(F) = E\left(e^{-yF(x)}\right) \tag{3.30}$$

for estimation of $F(x)$. Here E represent expectation; depending on the context, this may be a population expectation or a sample average. E_w indicates the weighted expectation. It's possible to show that $E\left(e^{-yF(x)}\right)$ is minimized at

$$F(x) = \frac{1}{2}\log \frac{P(y=1|x)}{P(y=-1|x)} \tag{3.31}$$

Hence:

- $P(y=1|x) = \frac{e^{F(x)}}{e^{-F(x)}+e^{F(x)}}$
- $P(y=-1|x) = \frac{e^{-F(x)}}{e^{-F(x)}+e^{F(x)}}$

The usual logistic transform does not have the factor 1/2. By multiplying the numerator and denominator by $e^{F(x)}$, we get the usual logistic model

$$p(x) = \frac{e^{2F(x)}}{1+e^{2F(x)}} \tag{3.32}$$

Hence the usual logistic model and the new model are equivalent up to a factor 2.
The following results are demonstrated in [8]:

- The Discrete AdaBoost algorithm builds an additive logistic regression model via Newton-like updates for minimizing $E\left(e^{-yF(x)}\right)$
- The Real AdaBoost algorithm fits an additive logistic regression model by stagewise and approximate approximate optimization of $E\left(e^{-yF(x)}\right)$.

These results show that both Discrete and Real AdaBoost can be motivated as iterative algorithms for optimizing the exponential criterion. The Gentle AdaBoost optimizes $E\left(e^{-yF(x)}\right)$ by Newton stepping.
A further observation is about the different outputs of random forests and Boosting: the age is the most important variable for the first method, but it seems to be not very important for the second ones. This fact allow us to fit a classifier giving importance to optical derived parameters, that is our goal.

3.6 Classification with absorptions

In this section the results of the analysis and classification related to absorption are briefly reported.
The data belong to Dataset 2, in particular to the part of data referred to absorptions. The way to act is substantially parallel to that used for concentrations, firstly the PCA directions has been used trying to find the directions that maximizes the differences in variability. The results weren't good in terms of performances, so we focused on logistic regression models considering all the available variables and possible interactions.
The correlation matrix is reported (Figure 3.21):
It's possible to note that several variables are strictly correlated (Figure 3.22).
In a second step dichotomous variables were analyzed in a similar manner to the case of concentrations, in particular it is estimated the confidence interval for the logarithm of the

	X635	X685	X785	X905	X930	X975	X1060	Age	BMI
X635	1.00000000	0.97726369	0.826182091	0.64858655	0.5913052	0.5691796	0.6737640	0.017606798	0.030056564
X685	0.97726369	1.00000000	0.900985708	0.73254734	0.6546097	0.5508589	0.7502447	0.032508493	0.032587779
X785	0.82618209	0.90098571	1.000000000	0.92829491	0.7808805	0.5106544	0.8663496	-0.001157315	-0.007025022
X905	0.64858655	0.73254734	0.928294912	1.00000000	0.8998348	0.4516330	0.8694537	0.049905751	0.027252968
X930	0.59130524	0.65460970	0.780880462	0.89983481	1.0000000	0.5387345	0.8570223	0.194310350	0.114809754
X975	0.56917963	0.55085885	0.510654421	0.45163297	0.5387345	1.0000000	0.7035381	0.248096472	0.192554685
X1060	0.67376403	0.75024467	0.866349570	0.86945372	0.8570223	0.7035381	1.0000000	0.163545144	0.108345562
Age	0.01760680	0.03250849	-0.001157315	0.04990575	0.1943103	0.2480965	0.1635451	1.000000000	0.370187095
BMI	0.03005656	0.03258778	-0.007025022	0.02725297	0.1148098	0.1925547	0.1083456	0.370187095	1.000000000

Figure 3.21: Correlation matrix with all the numerical variables

OR and any links with the different wavelengths were studied.
The significant correlations are now reported:

- **HRT**: A correlation between presence or recent use of HRT with absorption at 975nm was found (p-value$\approx$0.01);
- **OC**: A correlation between OC with absorptions at 905nm,930nm,975nm,1060nm was found (p-value$<$0.02).

A new model has been found developing the latest considerations; like in the previous sections, the backward variable selection has been applied and the resulting model is:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[x905\right]_i \left[OC\right]_i + \alpha_2 \left[x1060\right]_i \left[OC\right]_i + \alpha_3 \left[Age\right]_i + \alpha_4 \left[x905\right]_i \qquad (3.33)$$

The summary is as follows:

```
Coefficients:   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)     -7.25679   1.60465   -4.522   6.12e-06 ***
x905             7.20375   3.50130    2.057   0.0396 *
Age              0.13631   0.03027    4.503   6.69e-06 ***
x905:OC        -31.43713  16.71910   -1.880   0.0601 .
OC:x1060        63.42972  28.59415    2.218   0.0265 *
---
Null deviance  : 134.309 on 99 degrees of freedom
Residual deviance: 84.817 on 95 degrees of freedom
AIC: 94.817
```

The sensitivity is 0.8448276, the specificity is 0.7317073 and $1 - TotalErrorRate = 0.79$. Like with concentrations, the strong correlations between variable have been treated with log-penalized methods (LASSO and Ridge regression). In particular LASSO regression was really helpful in order to do variable selection. Now the partial results of the LASSO regression are used to make a selection of significant variables (the value of λ was chosen in a way that a certain number of coefficients has assumed non-zero value), we use the first 9 significant coefficients (taking care of the interaction):

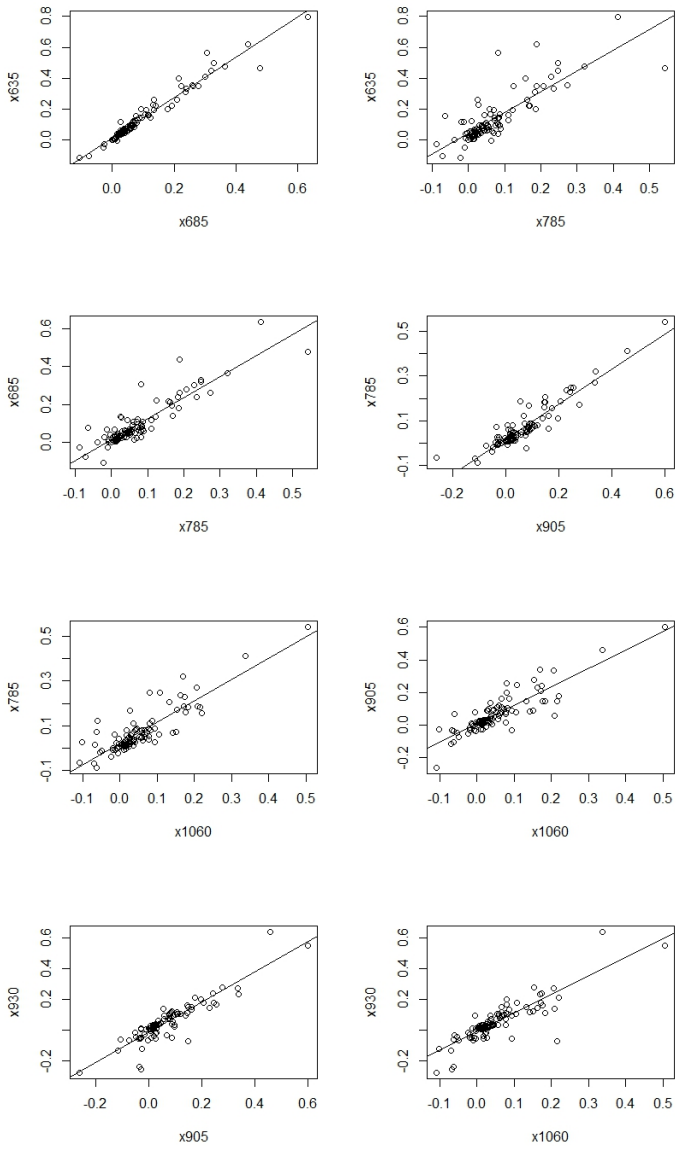

Figure 3.22: Plots of the most correlated variables. The regression lines have the following equation: 1)x635=1.378549268*x685+0.008603848; 2)x635=1.43618502 *x785+0.03518893; 3)x685=1.12245321*x785+0.01359344; 4)x785=0.83477683*x905+0.01108264; 5)x785=0.80429741*x1060+0.02444635; 6)x905=0.9783806*x1060+0.0151546; 7)x930=0.922694102*x905-0.005661969; 8)x930=1.0509828822*x1060-0.0001805148.

- x785;
- Age;
- Familiarity;
- HRT;
- BMI;
- TAMOXIFENE;
- OC*x905;
- OC*x1060;
- x1060*x785.

These variables have been used in a logistic regression model and a backward variable selection based on collinearity between variables has been made (mainly using the VIFs), the result is the following model:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[x785\right]_i + \alpha_2 \left[Age\right]_i + \alpha_3 \left[OC\right]_i \left[x905\right]_i + \alpha_4 \left[x1060\right]_i \left[OC\right]_i \qquad (3.34)$$

The summary of model (3.34) is now reported:

```
Coefficients:   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)     -7.47384   1.62987   -4.586   4.53e-06 ***
x785             8.96508   3.90185    2.298   0.0216 *
Age              0.13794   0.03042    4.535   5.76e-06 ***
x905:OC        -29.40037  15.93792   -1.845   0.0651 .
OC:x1060        62.44036  28.10551    2.222   0.0263 *
---
Null deviance  : 134.309 on 99 degrees of freedom
Residual deviance: 83.256 on 95 degrees of freedom
AIC: 93.256
```

The performances are: sensitivity is 0.8474576, specificity is 0.775 and $1-TotalErrorRate = 0.81$.
Fisher discriminant analysis and CARTs haven't lead to good results, for this reasons they haven't been reported.
The random forest algorithm was useful to try to create a ranking of the most significant variables (Figure 3.23). The technical characteristics are similar to the case of concentrations: the number of iterations is equal to 500 and the number of random chosen variables

for each split is 5 (approximately a third of the whole variables).
The most important variables are:

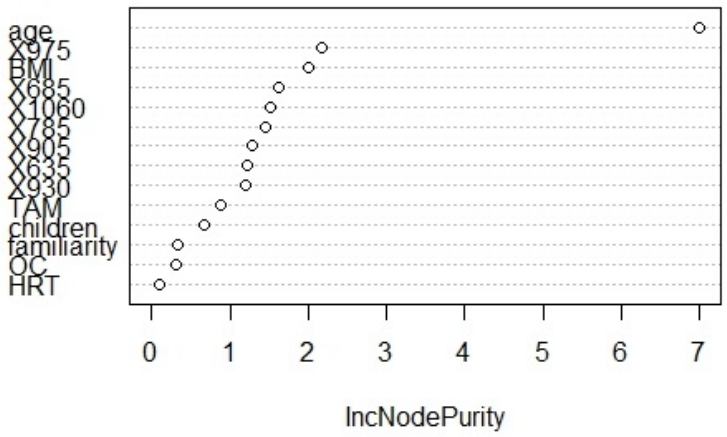

Figure 3.23: Variables Importance measured with random forest principle

- Age;
- x975;
- BMI;
- x685;
- x1060.

As usually the random forest importance measure has been calculated for the reduced model the take into account the most important variables (Figure 3.24). Performing a logistic regression on the variables reported in Figure 3.24 with all the possible interactions and applying a backward selection, we obtained the following model:

$$logit\left(p_i\right) = \alpha_0 + \alpha_1 \left[Age\right]_i + \alpha_2 \left[Age\right]_i \left[x685\right]_i \tag{3.35}$$

The summary of model (3.25) is reported below:

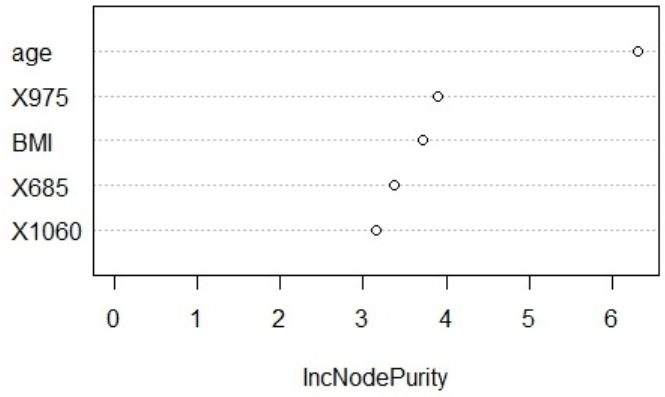

Figure 3.24: Variables Importance of reduced model measured with random forest principle

```
Coefficients:   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)     -6.70077   1.45791   -4.596   4.30e-06 ***
Age:x685         0.18752   0.07448    2.518   0.0118 *
Age              0.12750   0.02831    4.503   6.69e-06 ***
---
Null deviance  : 134.309 on 99 degrees of freedom
Residual deviance: 93.177 on 97 degrees of freedom
AIC: 99.177
```

The sensitivity is 0.8103448, the specificity is 0.7804878, $1 - TotalErrorrate = 0.7979798$. As far as AdaBoost methods are concerned, we can provide the following average importance rankings (Figures 3.25, 3.26, 3.27):

The performances are similar to those relating to concentrations: for the Discrete AdaBoost the average measured sensitivity is 0.88, the specificity is 0.90; for the Real AdaBoost the sensitivity is 0.86 and the specificity is 0.90; for the Gentle AdaBoost the sensitivity is 0.88 and the specificity is 0.90.

It is important to note that even in this case we have similar behaviours to the case of concentrations: age appears to be a key variable for classification in the case of logistic regression models and in the case of random forests, while loses importance for the methods of Boosting. Confirming this similar behaviour, in contrast to what one can imagine, the loss of importance of age also leads in this case to an improvement in the performance of the classifier.

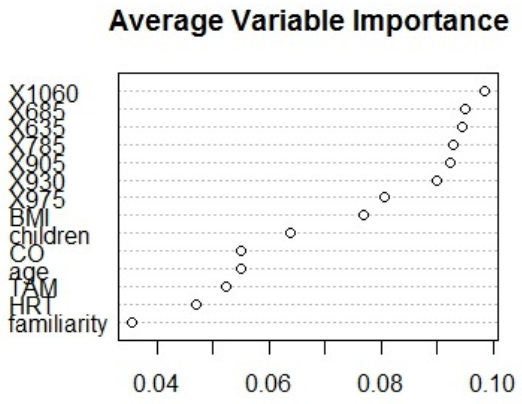

Figure 3.25: Average Variables Importance measured with Discrete AdaBoost method

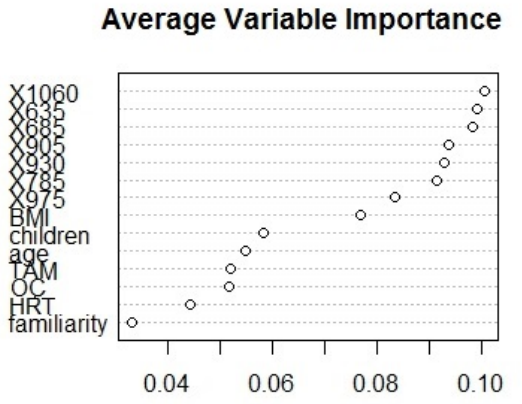

Figure 3.26: Average Variables Importance measured with Real AdaBoost method

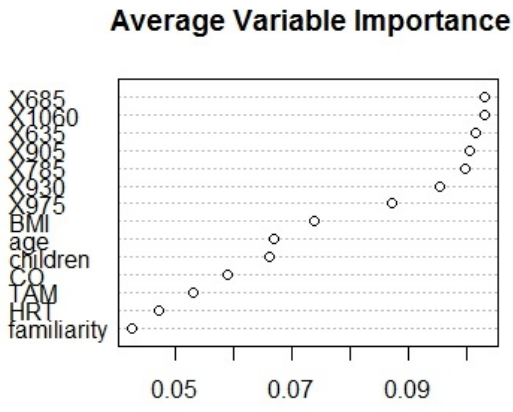

Figure 3.27: Average Variables Importance measured with Gentle AdaBoost method

3.7 Summary of classification

The most important logistic regression models developed in chapter 3 using concentrations are:

- $logit\,(p_i) = \alpha_0 + \alpha_1\,[Collagen]_i + \alpha_2\,[Age]_i + \alpha_3\,[Age]_i\,[Familiarity]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[Collagen]_i + \alpha_2\,[Age]_i + \alpha_3\,[Familiarity]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[Comp.1]_i + \alpha_2\,[Comp.4]_i + \alpha_3\,[Age]_i + \alpha_4\,[Familiarity]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[HbO_2]_i + \alpha_2\,[Age]_i + \alpha_3\,[HHb]_i\,[Water]_i + \alpha_4\,[Age]_i\,[Familiarity]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[Age]_i + \alpha_2\,[HbO_2]_i + \alpha_3\,[Water]_i + \alpha_1\,[HbO_2]_i\,[Water]_i$

The performances aren't very high, because in every case the sensitivity is less than 0.85 and the specificity is less than 0.8.
In the other hand we can classify with absorption data, the most important logistic regression models found were:

- $logit\,(p_i) = \alpha_0 + \alpha_1\,[x905]_i\,[OC]_i + \alpha_2\,[x1060]_i\,[OC]_i + \alpha_3\,[Age]_i + \alpha_4\,[x905]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[x785]_i + \alpha_2\,[Age]_i + \alpha_3\,[OC]_i\,[x905]_i + \alpha_4\,[x1060]_i$
- $logit\,(p_i) = \alpha_0 + \alpha_1\,[age]_i + \alpha_2\,[Age]_i\,[x685]_i$

Even in this case the performances aren't very high, and the values of sensitivity and specificity are almost comparable with those of the logistic regression models calculated with concentrations.
You can improve the performance of the classifiers using the boosting methods, but these methods are more complex and less intuitable than logistic regression models.
A big pro in favour of these methods are the lack of importance of the variable Age, that allows us to built a classifier mainly based on spectrally derived variables.
Since the results using the concentrations and absorptions are similar, it is preferable to use a classifier constructed with concentrations as such magnitudes have greater applicability and physical meaning.

Chapter 4

Direct estimate of risk associated with collagen

The second aim of the work is to try to identify a further significant risk factor for the identifications of patients with early breast cancer. As already mentioned in the introduction the first significant risk factor coincides with the density. High density significantly increases the probability of developing cancer.
The work shown in this chapter is aimed at verifying whether the collagen can be identified as a significant risk factor and if the two risk factors are somehow connected to each other (if they can be considered as complementary factors to identify high probability of contracting the disease or if they are somehow correlated)
Recent research has shown that there is a probable link between cancer risk and concentration of collagen (in addition to the density of the breast as described above). For this reason we've studied data related to 107 subjects (Dataset 4): the available variables for every subject are

- Density of the breast;
- Collagen;
- Age;
- Menopausal state;
- BMI.

The breast tissue is substantially constituted by:

- glandular tissue, stoma, epithelial tissue, representing the dense area, which appears white in the pictures, because it corresponds to the material that attenuates x-rays;

- adipose tissue, which appears dark (translucent) in the images, it flattens little x-rays.

The mammographic density is a percentage value because it is defined as the ratio between the area of "dense" and the total area in the images.

4.1 Differences in terms of mean and variability

As occurred previously in the case of the classification between benign and malignant tumours, the first step is intended to detect any differences between the two classes, which in this case are those of healthy subjects and diseased subjects. A possible difference in terms of variability can also be useful to explain the differences between these two populations. For this reason we've applied the tests previously described:

- Bartlett Test;
- Levene's Test;
- Analysis based on principal coordinate.

The plot of Collagen vs. Density doesn't allow us to see any sort of difference or classification criterion (Figure 4.1).
The multivariate Bartlett test has been used after having tested the hypothesis of normality

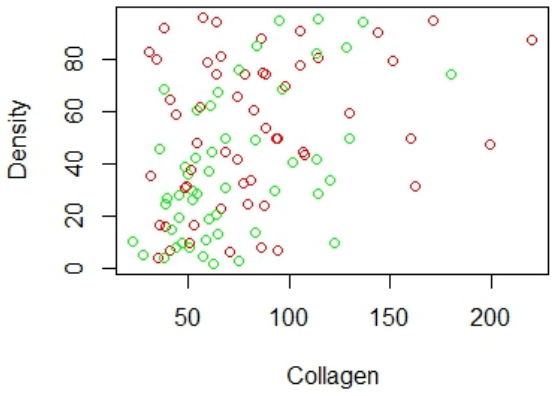

Figure 4.1: Collagen vs. Density

with the Shapiro Test (we refuses the null hypothesis of normality because the p-value was very low). The summary of the Bartlett test is:

- $u = 21.33989$;
- $\chi^2_{0.95}[15] = 24.99579$;
- $p - value = 0.1263166$.

This test is in fact testifying that it's not possible to refuse the null hypothesis of equal variances (but on the contrary there seems to be evidence that the null hypothesis is true because the p-value is very high).
The second test is the Levene's test:

- $F_c = 1.543$;
- $p - value_c = 0.217$;
- $F_m = 1.2411$;
- $p - value_m = 0.2678$.

The conclusion of this test is analogue to what we've told about the Bartlett test because the p-values are high. Figures 4.2 and 4.3 report the Euclidean distances from centroid/-

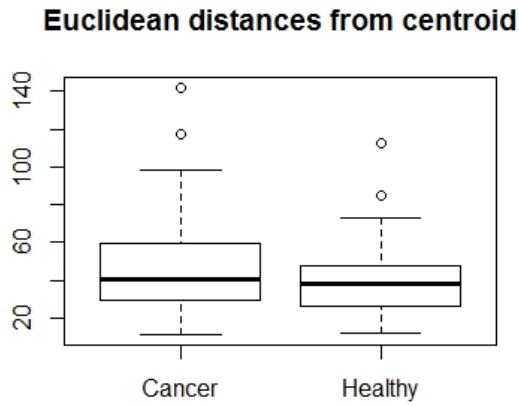

Figure 4.2: Euclidean distances from the centroid

median.
The representation with principal coordinate allows (Figure 4.4) us to note that the two populations have similar variability.

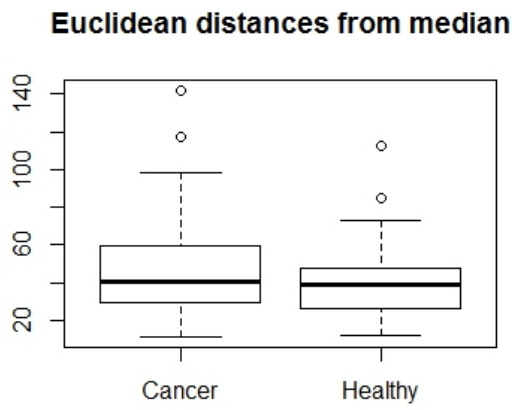

Figure 4.3: Euclidean distances from the median

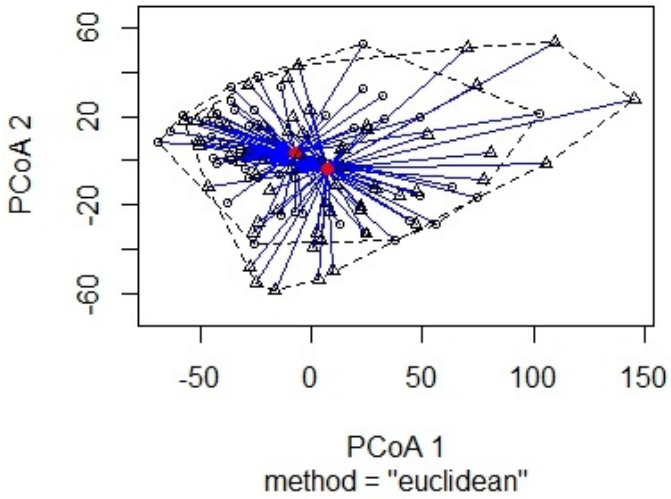

Figure 4.4: Principal coordinates ordination for Diseased and Healthy subjects: the triangles represent Diseased subjects, the circles represent Healthy subjects. The measure used is the Euclidean measure.

The last proof for equality of variability is the Tukey test for the distances from centroid. Tukey's test, also known as the Tukey range test, Tukey method, Tukey's honest significance test, Tukey's HSD (honest significant difference) test, or the Tukey-Kramer method, is a single-step multiple comparison procedure and statistical test. In our case we have only one comparison because the classes are only 2.
The Tukey method uses the studentized range distribution. The studentized range computed from a list $x_1, \ldots, x_n$ is

$$q_{n,\nu} = \frac{max[x_1 \ldots, x_n] - min[x_1, \ldots, x_n]}{s} \tag{4.1}$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \tag{4.2}$$

The critical value of q is based on three factors:

- α, the probability of rejecting a true null hypothesis;
- n, the number of observations;
- ν, the degrees of freedom used to estimate the sample variance.

Suppose that we take a sample of size n from each of k populations with the same normal distribution, let $\bar{y}_{min}$ be the smallest of the sample means and let $\bar{y}_{max}$ be the largest of the sample means. Suppose S^2 is the pooled sample variance from these samples. Then the following random variable has a Studentized range distribution:

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{S\sqrt{2/n}}$$

The Tukey confidence limits for all pairwise comparisons with confidence coefficient of at least $1 - \alpha$ are:

$$\bar{y}_i - \bar{y}_j \pm \frac{q_{\alpha;k;N-k}}{\sqrt{2}} \hat{\sigma}_\epsilon \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where n_i and n_j are the sizes of groups i and j respectively, $\hat{\sigma}_\epsilon$ is the standard deviation of the entire design, not just that of the two groups being compared.
Now we report the Tukey confidence intervals for the distances from centroid with euclidean distances. The value 0 is internal to the Tukey interval (Figure 4.5), so we can confirm the previous statement about the impossibility to reject the null hypothesis of equal variability with $\alpha = 0.05$.
The results of these tests allow us to perform the Hotelling test for the equality of the mean vectors.
The null hypothesis is:

$$H_0 : \mu_X = \mu_Y$$

Figure 4.5: Tukey confidence intervals for the distances from centroid with euclidean distances

The Two sample Hotelling's T-square test statistic is:

$$T^2 = \left(\bar{X} - \bar{Y}\right)^T \left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1} \left(\bar{X} - \bar{Y}\right) \tag{4.3}$$

where S is the pooled sample covariance matrix of X and Y, namely

$$S = \frac{(n_x - 1)\, S_X + (n_y - 1)\, S_Y}{(n_x - 1) + (n_y - 1)} \tag{4.4}$$

where S_X is the covariance matrix of the sample for X, $\bar{X}$ is the mean of the sample and n_x the number of elements in X; S_Y is the covariance matrix of the sample for Y, $\bar{Y}$ is the mean of the sample and n_y the number of elements in Y.
For n_x and n_y sufficiently large,

$$T^2 \sim \chi^2(k) \tag{4.5}$$

where k is the number of variables considered.
In our case the results are as follow:

- $T_0^2 = 0.7105255$;
- $\chi^2_{0.95}(5) = 11.0705$;
- $p - value = 0.982386$.

It is possible to synthesize the output of tests carried out so far by saying that it is not possible to say that there is a substantial difference in mean and variability of the data concerning healthy subjects and data relating to individuals with cancer.
Moreover exploratory analysis of the correlation matrix of the numeric variables hasn't shown any strong correlation (Figure 4.6).

	Collagen	Density	Type	Age	BMI
Collagen	1.0000000	0.3973235	0.13383149	-0.45360393	-0.54608519
Density	0.3973235	1.0000000	0.23534052	-0.44200837	-0.43000238
Type	0.1338315	0.2353405	1.00000000	0.01989051	-0.01112792
Age	-0.4536039	-0.4420084	0.01989051	1.00000000	0.36761463
BMI	-0.5460852	-0.4300024	-0.01112792	0.36761463	1.00000000

Figure 4.6: Correlation matrix of the numeric variables

4.2 Classification and evaluation of risk correlated to Collagen

Despite previous considerations, we tried to verify the degree of incidence of the main variables related to the probability of being a sick person. To do this we have fitted several logistic regression models considering all the variables. The only interaction considered is that between BMI and Collagen (Figure 4.7).
The idel model was the following one:

$$logit\,(p_i) = \alpha_0 + \alpha_1[age]_i + \alpha_2[Density]_i + \alpha_3[Menopause]_i \tag{4.6}$$

The values of the coefficients of model (4.6) are reported below:

```
Coefficients:  Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -2.618042   1.326848  -1.973   0.0485 *
Age            0.066367   0.027688   2.397   0.0165 *
Density        0.022467   0.008498   2.644   0.0082 **
Menopause   -1.190623   0.564995  -2.107   0.0351 *
---
Null deviance  : 148.25 on 106 degrees of freedom
Residual deviance: 135.17 on 103 degrees of freedom
AIC: 143.17
```

The performances are not very good: the sensitivity is 0.7272727, the specificity is 0.5384615 and $1 - TotalErrorRate = 0.635514$.

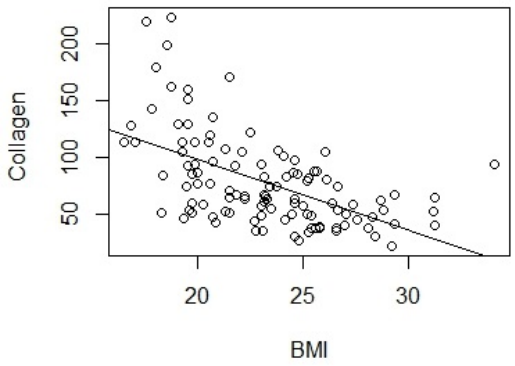

Figure 4.7: Plot of Collagen vs. BMI. The regression line has the following equation: Collagen=-5.816763*BMI+213.396344

But now we are interested in to understand if the collagen is a risk factor in addition to the density.
To do this, we have fitted two models:

- $logit(p_i) = \alpha_0 + \alpha_1[Collagen]_i$;
- $logit(p_i) = \alpha_0 + \alpha_1[Density]_i$.

and the probability of being sick were predicted by previous models ($p_{c,i}$ for the first model and $p_{d,i}$ for the second one). We can see these probabilities plotted in Figure 4.8.

From the fitted models it's possible to confirm that both the variables are significant (p-value$<$0.1) and it's obvious to say that if they were risk factors in agreement to each others, they should provide probabilities that lie on the bisector line. From Figure 4.8 is possible to say that it's not the case.
Like suggested in [2], it's possible to see the lack of agreement between two measurements with the plot shown in Figure 4.9.

Provided differences within $\overline{d} \pm 2s$ would not be statistically important, we should refer to these "limits of agreement":

- $\overline{d} + 2s = 0.22$
- $\overline{d} + 2s = -0.21$

Thus, the observations with differences less than 0.22 and higher than -0.21 would be considered acceptable. This is unacceptable because these differences are significantly relevant to the calculation of the probability of a patient to be a person with high risk of developing

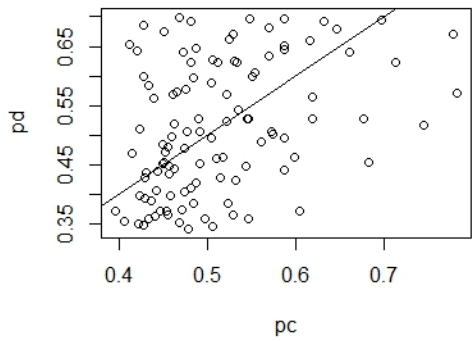

Figure 4.8: Plot of $p_{c,i}$ vs. $p_{d,i}$, the reference line is the bisector line.

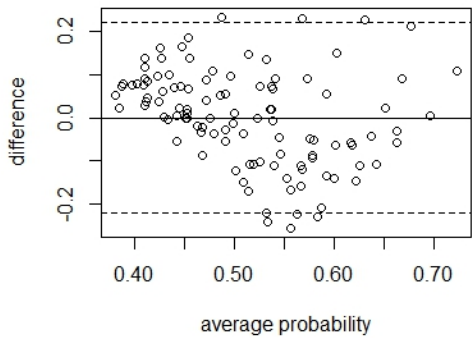

Figure 4.9: Plot of average probability vs. difference

cancer or not. This lack of agreement could not be obvious from Figure 4.8.
The same argument can be applied to models that also consider other variables (models 4.7 and 4.8):

$$logit(p_i) = \alpha_0 + \alpha_1[Collagen]_i + \alpha_2[Age]_i + \alpha_3[Menopause]_i + \alpha_4[BMI]_i \tag{4.7}$$

$$logit(p_i) = \alpha_0 + \alpha_1[Density]_i + \alpha_2[Age]_i + \alpha_3[Menopause]_i + \alpha_4[BMI]_i \tag{4.8}$$

As already described before we can consider the graphs of Figure 4.10.
You can come to the same conclusions drawn from the models analyzed before.

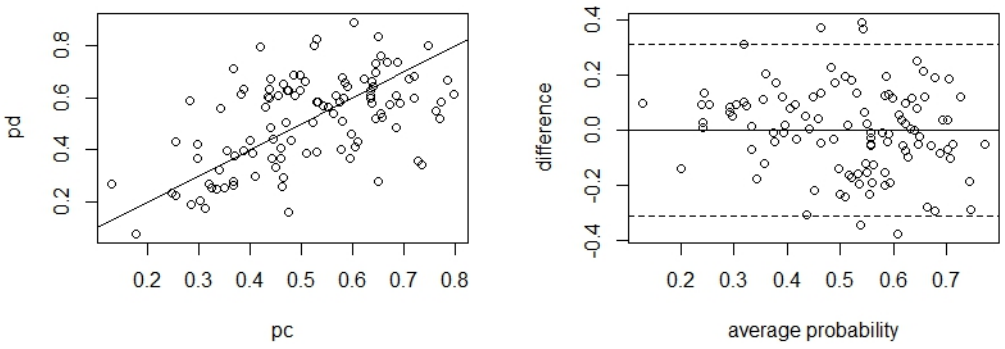

Figure 4.10: Plot $p_{c,i}$ vs. $p_{d,i}$ and probability vs. difference

In general, so it is possible to state that these methods suggest that there could be a difference between the two risk factors used.
With these considerations is thus possible to try to identify which are the real differences of subjects classified by the two variables in question as healthy / sick.
So we tried to classify according to collagen and density (using also Age, BMI, Menopausal state) using models (4.7) and (4.8), resulting in three possible classes of evaluation:

- subjects classified as sick according to both methods (therefore with high values of collagen and density);
- subjects classified sick for only one of the parameters used before;
- subjects classified as healthy according to both methods.

The results are as it follows (over the 107 subjects af the sample):

- 34 subjects are considered sick for both the methods, 22 of them are really cancer subjects, the other 12 are healthy (64.7% rightly classified);

- 20 subjects are classified sick for classification with density and healthy for the classification with collagen: 13 of them are sick and 7 are healthy;
- 17 subjects are classified sick for classification with density and healthy for the classification with collagen: 9 of them are sick and 8 are healthy;
- 36 subjects are considered sane for both the methods, 11 of them are really cancer subjects, the other 25 are healthy (69.4% rightly classified).

The summaries of the 2 models are now reported:

```
Coefficients:
Estimate  Std. Error z value Pr(>|z|)
(Intercept) -4.47808    2.28495  -1.960   0.0500 .
Age          0.04920    0.02518   1.954   0.0508 .
Collagen     0.01105    0.00684   1.616   0.1062
Menopause   -1.22282    0.56414  -2.168   0.0302 *
BMI          0.07568    0.07015   1.079   0.2807
---

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 148.25  on 106  degrees of freedom
Residual deviance: 139.85  on 102  degrees of freedom
AIC: 149.85
```

And for the second model:

```
Coefficients:
Estimate   Std. Error z value Pr(>|z|)
(Intercept) -6.008828   2.272334  -2.644  0.00819 **
Age          0.065802   0.027755   2.371  0.01775 *
Density      0.026425   0.009199   2.872  0.00407 **
Menopause   -1.370301   0.585808  -2.339  0.01933 *
BMI          0.092190   0.069225   1.332  0.18294
---

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 148.25  on 106  degrees of freedom
Residual deviance: 133.35  on 102  degrees of freedom
AIC: 143.35
```

As already predictable from what we saw in the correlation matrix, we can say that there are differences in Age and BMI in groups previously created: they both decrease in average passing from observations classified always as healthy as those classified differently by the two risk factors, and the same can be said finally passing the observations always classified as diseased.
You can build a classifier in the following way: it ranks as sick when at least one of the two basic classifiers ranks as sick and ranks as healthy in the other case. The performance of this classifier are the following: the sensitivity is 0.8 and the specificity is 0.48.

What we can say is that empirically seems to be a difference in the characteristics of the two predictive risk factors. The answer to the original problem placed at the beginning of the chapter can be summarized as follows: both variables in question are singly significant in establishing a degree of risk of having a breast cancer or not, but there are differences in relation to the prediction made by such variables. Density appears to be a stronger risk factor than the collagen (from what emerges from the logistic regression models), but also the collagen is a correction factor which often helps to rank exactly certain subjects.
Other studies have been done trying to consider other transformations of variables, such as standardization of collagen and density compared to the value predicted by a linear regression model built in reference to the age of the patients, in particular we considered the following models in which the parameters are calculated only with the training set constituted by healthy subjects:

- $[Collagene]_i = \alpha_0 + \alpha_1[Age]_i$
- $[Density]_i = \beta_0 + \beta_1[Age]_i$

and then

- $[Collagene_{std}]_i = \frac{\alpha_0+\alpha_1[Age]_i-[Collagene]_i}{[Collagene]_i}$
- $[Density_{std}]_i = \frac{\beta_0+\beta_1[Age]_i-[Density]_i}{[Density]_i}$

These new variables allow us to consider the anomaly of the data compared to the hypothetical behaviour of an healthy subject.
It's important to note that the use of these variables has not led to good results, starting from the significance of variables in the simple models reported below:

- $logit(p_i) = \alpha_0 + \alpha_1[Collagen_{std}]_i$
- $logit(p_i) = \alpha_0 + \alpha_1[Density_{std}]_i$

In fact it can be seen from the summary that standardised collagen is not meaningful in relation to the probability of being a sick person:

```
Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1092     0.1990   0.549    0.583
Collagen_std -0.5460     0.4180  -1.306    0.192

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 148.25  on 106  degrees of freedom
 Residual deviance: 146.50  on 105  degrees of freedom
 AIC: 150.5
```

And the second summary is:

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1911     0.2034   0.940   0.3474
vtd          -0.3776     0.1658  -2.277   0.0228 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 148.25  on 106  degrees of freedom
Residual deviance: 141.11  on 105  degrees of freedom
AIC: 145.11
```

Taking into account these considerations and the fact that no logistic regression model has led to better performance than the model (4.6), it's been decided not to consider further as credible these transformations of the variables in question.
A natural consequence of these results is the search to improve the performance of the classifier. The idea is to use the Boosting methods in such a way to have a ranking of the variables.
In particular these methods have been used with the following technical indications:

- The number of iteration is 50 for every classifier;
- the percentage of observation left out of bag is 50% for every iteration;
- The variables ranking is built based on the average result of 20 classifiers.

The variables rankings are reported in Figure 4.11, Figure 4.12 and Figure 4.13.
The average performances are:

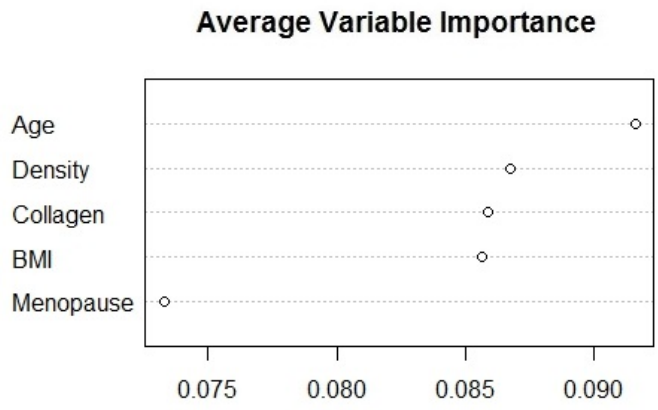

Figure 4.11: Average variable ranking with Discrete AdaBoost method.

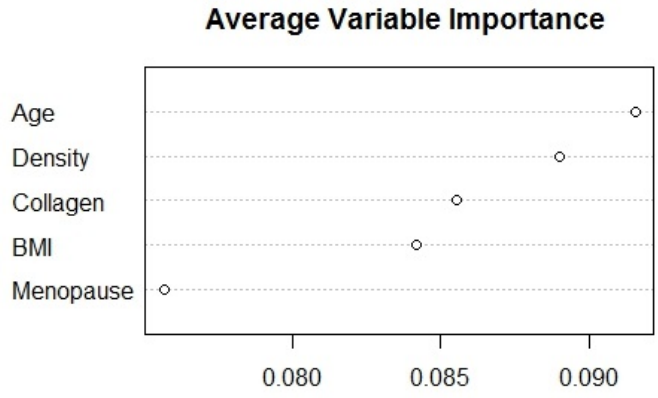

Figure 4.12: Average variable ranking with Real AdaBoost method.

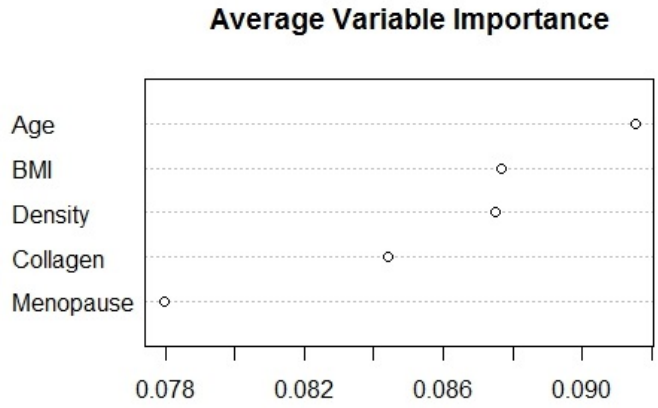

Figure 4.13: Average variable ranking with Gentle AdaBoost method.

- Discrete Adaboost: sensitivity of 0.70, specificity of 0.75 and $1 - TotalErrorRate = 0.72$.
- Real Adaboost: sensitivity of 0.72, specificity of 0.74 and $1 - TotalErrorRate = 0.73$.
- Discrete Adaboost: sensitivity of 0.70, specificity of 0.75 and $1 - TotalErrorRate = 0.72$.

As we can see, also in this case the performances are not very high. The most important things for our work is the ranking of the variables: in every single case the collagen is less "important" than the density; it confirms what we've seen with logistic regression models. The last further confirmation of what said up to this point is what comes from the analysis of the proportions of healthy and sick subjects with values at the extremes of collagen: if you do not notice any difference between the proportions of diseased patients with extremely high value of collagen and diseased patients with extremely low value of collagen, then we have a further proof of the fact that collagen can not be considered a good indicator of the risk of developing cancer.
For this reason we have compared the proportion of diseased patients relative to patients in the first 20% of the ordered values of collagen with the same proportion in the last 20% of patients. The results are reported below:

	Healthy	Diseased
High Collagen	10	11
Low Collagen	11	10

The 95% confidence interval for the logarithm of the OR is:

$$IC[\ln(OR)] = (-1.1748539; 0.8109668) \tag{4.9}$$

This testifies the impossibility to reject the hypothesis of independence of the two categorical variables (high/low collagen and presence/absence of disease), or in other words you can not identify significant differences between proportions with high and low collagen.
The same work has been done with density. The results are as follows:

	Healthy	Diseased
High Density	7	14
Low Density	14	7

The 95% confidence interval for the logarithm of the OR is:

$$IC[\ln(OR)] = (0.2722328; 2.3647497) \tag{4.10}$$

This confirms the results already obtained in previous studies: you can note different proportions of diseased subjects in parts of the population placed at the ends of the density distribution (in contrast to what happened with collagen). This is a further example that the density can be considered actually a risk factor that increases the probability of having the disease because his presence at extremely high values implies an effective increase in the probability of being a sick person. This statement can not be made for collagen because there is no evidence based on the results obtained.
Then it can be concluded by saying that there is probably a difference between collagen and density in terms of power of explanation of the probability of being a sick subject, but the density is a better risk factor in terms of significance in every classifier (both in logistic regression models and in Boosting methods). In particular as you can see from the model (4.6) the use of density as a risk factor makes collagen not significant.

Chapter 5

Conclusions

The work was mainly divided in 2 parts: the first one is on finding adequate logistic regression models that could adequately classify observations with malignant or benign tumours. To do this, through several considerations, different models have been fitted.
The most important logistic regression models developed in chapter 3 are reported in section 3.7, they have been fitted using Dataset 2.
The performances aren't very high, because in every case the sensitivity is less than 0.85 and the specificity is less than 0.8.
Based on these models, it is not possible to identify a unique model but of course we can draw some considerations:

- Age is the most important variable in this context: the p-values are very low in every fitted model, so there is a strong evidence in favour of the hypothesis of significant coefficient;
- Collagen is a significant factor in the classification of malignant and benign tumours. High values of delta-Collagen are an indication of increased risk of having a malignant tumour. To achieve this you need to consider that HbO_2 and collagen are closely related (positively), then surely we can come to similar considerations for these two variables;
- Familiarity is a variable with different behaviour from what would be expected intuitively. In fact, there is strong evidence to suggest that familiarity is a factor that would lead to a reduction of the probability of contracting a malignant tumour;
- It is probable, but less obvious in the light of the results obtained, that the Water is a significant variable that increases the likelihood of contracting a malignant tumour.

In the other hand we can classify with absorption data, the most important logistic regression models found are reported in section 3.7.
Even in this case the performances aren't very high, and the values of sensitivity and specificity are almost comparable with those of the logistic regression models calculated with concentrations.
One of the main problems of these models is the fact that it is difficult to reach considerations as we have just done with models built with the concentrations. This is because of the high number of correlated variables and the greater difficulty in giving a physical intuitable meaning to these variables.
You can improve the performance of the logistic regression models (in particular for concentrations) using the Boosting methods, even if these methods are less intuitable than logistic regression models.
A big pro in favour of these methods are the lack of importance of the variable Age, that allows us to built a classifier mainly based on spectrally derived variables.
Since the results using the concentrations and absorptions are similar, it is preferable to use a classifier constructed with concentrations because they have greater applicability and physical meaning.
The second aim of this work was about the evaluation of collagen as a risk factor in favour of the presence or absence of tumours in certain individuals. It's possible to confirm this presumption although it is not possible to say that collagen is a fundamental variable and determining the classification of such subjects.
In any case we can consider the collagen as a factor that increases the likelihood of having a breast cancer, as well as it had been previously confirmed this hypothesis also for the density. However, as suggested by the last part of the work, it is possible to say that the two risk factors are not closely related, and thus can be considered as two different elements that may allow to define a probability of having a breast cancer.
The Boosting methods have been used also in this study: they helped us to understand the importance of the variables. The results in terms of significance of the variables are very similar to those related to logistic regression models: the collagen seems to be less important when compared to the density (weakly significant in the purpose of the classification).

Chapter 6

Codes

6.1 Packages

The whole analysis has been conducted with R. The main packages used are reported below:

- **car**:John Fox and Sanford Weisberg (2011). An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage.
  URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
- **vegan**:Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner (2013). vegan: Community Ecology Package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan
- **glmnet**: Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.
- **mvpart**: rpart by Terry M Therneau, Beth Atkinson. R port of rpart by Brian Ripley <ripley@stats.ox.ac.uk>. Some routines from vegan – Jari Oksanen <jari.oksanen@oulu.fi> Extensions and adaptations of rpart to mvpart by Glenn De'ath. (2014). mvpart: Multivariate partitioning.
  R package version 1.6-2. http://CRAN.R-project.org/package=mvpart
- **MASS**:Venables, W. N.; Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- **lasso2**:Justin Lokhorst, Bill Venables, Berwin Turlach; port to R and tests etc: Mar-

tin Maechler (2014). lasso2: L1 constrained estimation aka 'lasso'. R package version 1.2-19. http://CRAN.R-project.org/package=lasso2

- **rgl**: Daniel Adler, Duncan Murdoch and others (2014). rgl: 3D visualization device system (OpenGL). R package version 0.95.1157.
  http://CRAN.R-project.org/package=rgl
- **randomForest**:A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.
- **rpart**:Terry Therneau, Beth Atkinson and Brian Ripley (2013). rpart: Recursive Partitioning. R package version 4.1-3. http://CRAN.R-project.org/package=rpart
- **ada**:Mark Culp, Kjell Johnson and George Michailidis (2012). ada: ada: an R package for stochastic boosting.
  R package version 2.0-3. http://CRAN.R-project.org/package=ada
- **adabag**:Esteban Alfaro, Matias Gamez, Noelia Garcia (2013). adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software, 54(2), 1-35. URL http://www.jstatsoft.org/v54/i02/.
- **caTools**:Jarek Tuszynski (2014). caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.. R package version 1.17.1.
  http://CRAN.R-project.org/package=caTools

6.2 Codes

In this section some codes used in the analyzes have been reported. The packages used for a particular item are reported.

- **Multivariate Bartlett Test**:

```
alfa <- 0.05
p <- ncol(paz)  #number of variables
n <- nrow(paz)  #number of patients
k <- 2     #number of groups
pame <- rep(0,k)
nu <- as.vector(table(group))
mat <- mat1 <- array(dim=c(p,p,k))
for (i in 1:k) {
mat[,,i]=((nu[i]-1)/nu[i])*cov(paz[group==i-1,])
mat1[,,i]=(nu[i]-1)*cov(paz[group==i-1,]) }
```

```
Sp <- apply(mat1,1:2,sum)/(n-k)    #pooled variance matrix
c <- 1-(2*p^2+3*p-1)/(6*p-6)*
    (1/(nu[1]-1)+1/(nu[2]-1)-1/(n-k))
test <- c*((n-k)*log(det(Sp))-(nu[1]-1)*
    log(det(cov(paz[group==0,]) ))-(nu[2]-1)*
    log(det(cov(paz[group==1,]) )))
df <- 0.5*p*(p+1)*(k-1)    #degrees of freedom
val <- qchisq(1-alfa,df)
test>val
pvalue <- 1-pchisq(test,val)
```

- **Levene's Test**:

```
cm <- colMeans(deltaM[,1:5])  #malignant mean
cs <- colMeans(deltaB[,1:5])  #benign mean
zm <- rep(0,63)
zs <- rep(0,47)
for (i in 1:63)
        {diff <- deltaM[i,1:5] - cm
        zm[i] <- sqrt(sum(diff^2))
        }
for (i in 1:47)
        {diff <- deltaB[i,1:5] - cs
        zs[i] <- sqrt(sum(diff^2))
        }
#one way anova
z <- c(zm,zs)
fit <- aov(z ~ group)
fit
summary(fit)
boxplot(zm,zs,main="Euclidean distances from centroid",
    names=c("Malignant","Benign"))
```

- **Principal coordinates method**: (**vegan**)

```
dis <- vegdist(paz,method="manhattan")
dis
group <- factor(c(rep(1,63), rep(2,47)),
    labels = c("Malignant","Benign"))

mod <- betadisper(dis, group,type="centroid")
```

```
mod
summary(mod)

## Perform test
anova(mod)
## Permutation test for F
permutest(mod, pairwise = TRUE)

## Tukey's Honest Significant Differences
(mod.HSD <- TukeyHSD(mod))
plot(mod.HSD)

## Plot the groups and distances to centroids on the
## first two PCoA axes
plot(mod)
boxplot(mod)
```

- **LASSO, Ridge and elastic net regression**:(**glmnet**)

```
rlas<-glmnet(x, y, family = "binomial",
    alpha = 1, lambda.min = 1e-4)
rrid<-glmnet(x, y, family = "binomial",
    alpha = 0, lambda.min = 1e-4)
renet<-glmnet(x, y, family = "binomial",
    alpha = .5, lambda.min = 1e-4)
summary(rlas)
plot(rlas,xvar="lambda")
plot(rrid,xvar="lambda")
plot(renet,xvar="lambda")
nsteps <- 20
b1 <- coef(rlas)[-1, 1:nsteps]
w <- nonzeroCoef(b1)
b1 <- as.matrix(b1[w, ])

b2 <- coef(rrid)[-1, 1:nsteps]
w <- nonzeroCoef(b2)
b2 <- as.matrix(b2[w, ])

b3 <- coef(renet)[-1, 1:nsteps]
w <- nonzeroCoef(b3)
b3 <- as.matrix(b3[w, ])
ylim <- range(b1,b2,b3)
```

```
matplot( t(b1), type = "o", pch = 19, col = "blue",
    xlab = "Step", ylab = "Coefficients",
        ylim = c(-0.6,0.1), lty = 1)
title("Lasso")
abline(h = 0, lty = 2)

matplot( t(b3), type = "o", pch = 19, col = "blue",
    xlab = "Step", ylab = "Coefficients",
        ylim = ylim, lty = 1)
title("Elastic Net")

matplot( t(b2), type = "o", pch = 19, col = "blue",
    xlab = "Step", ylab = "Coefficients",
        ylim = ylim, lty = 1)
title("Ridge Regression")
```

- **Fisher Discriminant Analysis**:

```
n1 <- nrow(sani); n1
n2 <- nrow(malati); n2

p <- ncol(sani); p
g <- 2
n <- n1+n2; n

#W
S.sani <- var(sani)
S.malati <- var(malati)
W.2gr <- ((n1-1)*S.sani+(n2-1)*S.malati)/(n-2)
W.2gr

#B
mean.sani <- colMeans(sani)
mean.malati <- colMeans(malati)

B.2gr <- ((n1*n2/n)*(mean.sani-mean.malati)%*%
t(mean.sani-mean.malati))
B.2gr

#direction that maximizes the difference based on
mean values
```

```
a.2gr <- solve(W.2gr)%*%(mean.sani-mean.malati)
paz.2gr <- as.matrix(deltarid)
Y.2gr <- paz.2gr%*%a.2gr
dim(Y.2gr)

-1*a.2gr
Y.2gr;

#new coordinates
Y.2gr.sani <- mean.sani%*%a.2gr
Y.2gr.malati <- mean.malati%*%a.2gr

pto.sep <- (Y.2gr.sani+Y.2gr.malati)/2;pto.sep

#performance
Sp.prev <- ifelse(Y.2gr>(c(rep(pto.sep,n))),0,1)
gruppi<-matrix(NA,nrow=100,ncol=1)
gruppi[,1]<-c(rep(1,58),rep(0,42))
alloc <- as.data.frame(cbind(Y.2gr, gruppi,Sp.prev))
table(alloc$V2,alloc$V3)
color.position <- ifelse(gruppi == '1', 'red', 'green')
plot(Y.2gr,col=color.position)
abline(h=pto.sep)
```

- **Random Forest**:(**randomForest**)

```
fit <- randomForest(group~HHb+HbO2+Lipid+Water
    +Collagen+age+familiarity+OC+HRT+children+TAM+BMI,
    type="classification",data=deltatot)
fit
print(fit) # view results
importance(fit) # importance of each predictor
varImpPlot(fit)
pred<-fit$predicted
summary(pred)

#performance
gruppipred<-NULL
for(i in 1:100)
{
if(pred[i]<0.5)gruppipred[i]=0
else gruppipred[i]=1
```

```
}
gruppipred
gruppipred<-as.matrix(gruppipred)
dimnames(gruppipred)[[1]]<-dimnames(deltatot)[[1]]
ms<-cbind(as.double(deltatot[,6]),gruppipred,rep(0,100))
ms
for(i in 1:100)
{
if(ms[i,1]==ms[i,2])ms[i,3]=1
}
ms
mc<-sum(ms[1:58,3]);
hc<-sum(ms[59:100,3]);
sum(ms[,3])/100
```

- **AdaBoost methods**: (**ada**) the Discrete AdaBoost application is reported.

```
vars<-rep(0,12)
t1<-proc.time()
for(i in 1:20)
{
rm(gen1)
gen1<-ada(group~Water+HHb+HbO2+Collagen+BMI+age
    +Lipid+familiarity+OC+HRT+TAM+children,
    data=deltatot,type="discrete")
vec1<-varplot(gen1,plot.it=FALSE,type="scores",
    max.var.show=12)
vars<-vars+as.numeric(vec1[order(names(vec1))])/20
cat("i=",i,"␣time=",(proc.time()-t1)/60,"\n")
}
a1<-sort(names(vec1))
a2<-order(vars,decreasing=TRUE)
dotchart(vars[a2][12:1],a1[a2][12:1],main=
    "Average␣Variable␣Imp.")
```

Bibliography

[1] M. J. Anderson, "Distance-Based tests for homogeneity of multivariate dispersion", *Biometrics*, 62, (2006)

[2] J. Martin Bland, Douglas G. Altman, "Measurement in medicine: the analysis of method comparison studies" , *The Statistician*; 32, (1983)

[3] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Trees", 1984 *Wadsworth & Brooks/Cole Advanced Books & Software.*

[4] Breiman, Leo (2001). "Random Forests". *Machine Learning*, 45, Springer

[5] Brown M.B. and Forsythe A.B., "Robust tests for the equality of variances", *Journal of American Statistical Association*, 69, (1974)

[6] Chi-Ling Chen, Noel S.Weiss, Polly Newcomb, William Barlow, Emily White, "Hormone Replacement Therapy in Relation to Breast Cancer", *Journal of American Association*, Vol.287, 734- 741, (2002)

[7] C. D'Andrea, L. Spinelli, A. Bassi, A. Giusto, D. Contini, J. Swartling, A. Torricelli, R. Cubeddu, "Time-resolved spectrally constrained methof for the quantification of cromophore concentrations and scattering parameters in diffusing media", *Opt. Express.*, 14(5), (2006)

[8] Freund Y., Schapire R., "Experiments with a new boosting algorithm", *Machine learning: proceedings of the thirteenth international conference*, (1996)

[9] Friedman J., Hastie T., Tibshirani R., "Regularization paths for generalized linear models via Coordinate descent", *Department of Statistics, Stanford University*, (2009)

[10] J. Friedman, T. Hastie, R. Tibshirani, "Additive logistic regression: a statistical view of boosting", *The annals of Statistics*, 28(2), (2000)

[11] Gower J.C., "Some distance properties of latent root and vector methods used in multivariate analysis", *Biometrika*, 53, (1966)

[12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, "Numerical recipes in C:The art of scientific computing", *Cambridge University Press, New York*, (2002)

[13] Ralph B. D'Agostino SR, Heidy K. Russell, "Multivariate Bartlett Test", *Encyclopedia of Biostatistics*, (2005)

[14] R. W. Sattin, G. L. Rubin, FRACP; L. A. Webster, C. M. Huezo, P. A. Wingo, H. W. Ory, P. M. Layde, "Family History and the Risk of Breast Cancer", *Journal of American Association*, 253(13):1908-1913, (1985)

[15] P. Taroni, A. Bassi, D. Comelli, A. Farina, R. Cubeddu, A. Pifferi, "Diffuse optical spectroscopy of breast extended to 1100nm", *J. Biomed. Opt.*, 14(5), (2009)

[16] P. Taroni, G. Quarto, A. Pifferi, F. Ieva, A. M. Paganoni, F. Abbate, N. Balestrieri, S. Menna, E. Cassano, R. Cubeddu, "Optical identification of subjects at high risk for developing breast cancer",*J.of Biomedical Optics*, 18(6), (2013)

[17] P.Taroni, A.Pifferi, E.Salvagnini, L.Spinelli, A.Torricelli and R.Cubeddu, "Seven-wavelenght time-resolved optical mammography extending beyond 1000nm for breast collagen quantification", *Opt. Expr.* 17, 15932-15946 (2009)

[18] P. Taroni, A. Pifferi, E. Salvagnini, L. Spinelli, A. Torricelli, R. Cubeddu, "Seven-wavelenght time-resolved optical mammography extending beyond 1000nm for breast collagen quantification", *Opt. Express*, 17(18), (2009)

[19] Tibshirani R., Bien J., Friedman J., Hastie T., Simon N.,Taylor J., (2012) "Strong Rules for Discarding Predictors in Lasso-type Problems", *JRSSB* vol 74

[20] Tibshirani R., "Regression shrinkage and selection via the LASSO", *Journal of the Royal Statistical Society*, 58, (1996)

[21] Vantini S., "I CART e il Problema della Classificazione Statistica: Teoria e Applicazioni", master's degree thesis, available at www1.mate.polimi.it/biblioteca.it

[22] Woolf B., "On estimating the relation between blood group and disease", *Annals of Human Genetics* 19, 251-253, (1955).

[23] Collaborative Group on Hormonal Factors in Breast Cancer. "Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53297 women with breast cancer and 100239 women without breast cancer from 54 epidemiological studies". *Lancet*, 347: 1713-1727, (1996)

[24] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ringraziamenti

I primi naturali ringraziamenti per questo lavoro sono indirizzati a chi mi ha direttamente seguito durante questi mesi di ricerche ed applicazioni, ossia la Prof.sa Paganoni, che più volte mi ha saputo dare importanti indicazioni e ha trovato il tempo di rispondere alle decine di domande che le ponevo.

Allo stesso tempo vorrei ringraziare anche la Prof.sa Taroni e il Prof. Pifferi per il prezioso contributo nozionistico e il supporto "fisico".

Penso sia fondamentale sottolineare che questo lavoro non sarebbe mai stato lo stesso se non avessi conosciuto tutte quelle persone che ho avuto modo di conoscere e apprezzare in questi anni: Chiara penso sia stata fondamentale soprattutto negli ultimi anni, il suo carattere stupendo mi insegnato cosa vuol dire l'ottimismo nella vita; Luca è il compagno di risate e avventure che è diventato nel tempo un mio riferimento; Alessandra mi ha saputo aiutare sempre, con i modi e le parole di chi sa sempre cosa dire al momento giusto; Riccardo è stato il mio riferimento in Collegio e fuori, ho sempre seguito alla lettera ogni suo consiglio e lo reputo una persona estremamente disponibile e di cuore; Federico, Nina, Marco e Francesca per la spensieratezza, ho condiviso con loro gran parte del mio percorso e non avrei potuto fare a meno di loro; Alberto e Sirio per i momenti di studio passati insieme a sopportarci a vicenda; con Paola e Valentina ho avuto meno opportunità per conoscerci, ma le sento parte di questo meraviglioso gruppo e quindi meritano sicuramente di essere nominate.

Naturalmente anche la realtà del mio paesino mi ha permesso di raggiungere questo obbiettivo, senza i miei amici sarebbe stato tutto più difficile, per questo motivo vorrei ringraziarli uno ad uno.

Per ultima, ma solo perchè la reputo più importante di tutto, voglio ringraziare la mia famiglia. Penso che 'supporto' e 'fiducia' siano le due parole che meglio descrivono come

mi sono sentito in questi 5 anni: sapere di avere qualcuno che mi aiuta nel miglior modo possibile, dandomi tutte le possibilità di questo mondo senza mai mettermi vincoli e permettendomi ogni opportunità di scelta basandosi esclusivamente su una estrema fiducia nei miei confronti è la prova più grande di come loro stessi credano in me e vogliano per me solo il meglio. Parlo dei miei genitori, e per questo non vorrei mai deluderli.
Stesse parole valgono per Filippo: un ragazzo dalle potenzialità indescrivibili che è in grado di trascinare chiunque gli stia intorno, me compreso. Io ho la fortuna di averlo come fratello.