Danmarks Tekniske Universitet

Data: Feature extraction, and visualization

Project report 1

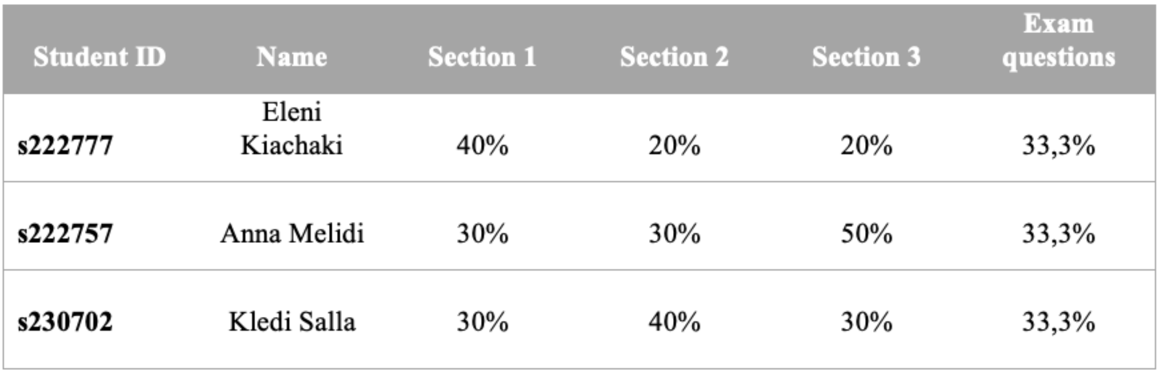
Authors

Eleni Kiachaki - s222777
Anna Melidi - s222757
Kledi Salla - s230702

March 7, 2023

Contents

1 Contribution



Student ID	Name	Section 1	Section 2	Section 3	Exam questions
s222777	Eleni Kiachaki	40%	20%	20%	33,3%
s222757	Anna Melidi	30%	30%	50%	33,3%
s230702	Kledi Salla	30%	40%	30%	33,3%

Figure 1: Contribution of each member to the project

2 Description of Data set

2.1 Problem of interest

Our interest lies in predicting breast cancer based on anthropocentric data and parameters that can be gathered in routine blood analysis. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer and consequently allow for early detection ensuring a greater probability of having a good outcome in treatment.

2.2 Source of data

The UCI machine learning repository was used, from which we retrieved Breast Cancer Coimbra Data Set [1]. This data set came from 64 women newly diagnosed with breast cancer (BC) that were recruited from the Gynaecology Department of the University Hospital Centre of Coimbra (CHUC) between 2009 and 2013. All samples were naive, i.e. collected before surgery and treatment. On the other hand, the 52 controls were female healthy volunteers. All patients had had no prior cancer treatment and all participants were free from any infection or other acute diseases or comorbidities at the time of enrolment in the study.

2.3 Previous analysis summary

2.3.1 Univariate analysis

In a previous study published on BMC Cancer in 2018 [2], the scientists performed an univariate statistical analysis on the data set, to assess the individual diagnosis values of several parameters. For this purpose, each quantitative variable was assessed for normality, using Shapiro-Wilk tests. Upon verifying the normality requirements not being met, median values and interquartile ranges were computed for each variable, while categorical variables were described in terms of absolute frequencies and percentages. Following this, ROC analysis was performed for each of the nine parameters, and the obtained area under the ROC curve, was computed as an indicator of the diagnostic predictive value corresponding to each attribute. For the parameters that scored the biggest diagnostic predictive value, the pair of sensitivity and specificity values that maximize the Youden Index were computed.

2.3.2 Multivariate analysis

Subsequently, a multivariate analysis was applied so as to combine several parameters and generate models to distinguish between healthy subjects and those with breast cancer. More specifically, the importance as breast cancer predictors of each of the variables was determined, using the Gini coefficient to estimate the total decrease in node impurities associated to splitting on the variable in a Random Forest algorithm, averaged over all trees.

Finally, using three classification algorithms (logistic regression, support vector machine and random forests), predictive models were built. A Monte Carlo Cross-Validation (MCCV) approach was adopted, in which the three models were constructed on a training set of 69.8% the total amount of data.
Following this workflow, the scientists managed to predict the presence of breast cancer in women on a test data set with sensitivity ranging between 82 and 88% and specificity ranging between 85 and 90% based on the parameters Resistin, Glucose, Age and BMI.

2.4 Classification and regression on our data set

With the classification technique we will try to predict the condition of the individual (breast cancer or healthy) based on the metabolic and anthropocentric attributes. The attributes that will be used will be selected after applying an analysis that will determine which attributes out of the 9 have the maximal positive predictive value to the condition of the individual. As far as the regression task is concerned, even though the dataset is not commonly used for regression purposes, we could predict several attributes such as insulin based on the HOMA and/or glucose levels.

Before starting processing the dataset, we need to normalize as our attributes represent different things and are on vastly different scales. For example, Body Mass Index (BMI) is scaled in kg/m^2 from about 17 to almost 40, while MCP-1 on pg/dL with minimum value around 60 and maximum above 1000. Without normalization, Principal Component Analysis will then accurately detect that the variance in the dataset is primarily in the MCP-1, after all this quantity will vary with many hundreds between individuals while the BMI will only vary with about some 1 kg/m^2 . Therefore by subtracting the mean from each element and dividing it by the standard deviation we ensured that all the attributes have the same scale, and PCA will easier pick out potential correlations.

3 Explanation of the attributes of the data

3.1 Attributes' description

1. Age: discrete, interval (years)

2. BMI: continuous, interval (kg/m^2). It is the measurement of body fat based on a person's weight and height

3. Glucose: continuous, interval (mg/dL). A type of sugar that is a primary source of energy for the cells in the body

4. Insulin: continuous, interval (μU/mL). It plays a key role in regulating the amount of glucose in the bloodstream

5. HOMA (homeostasis model assessment) : continuous, interval. It is a method used to quantify insulin resistance which is typically develped to obese and older people.

6. Leptin: continuous, interval (ng/mL). Leptin is a hormone that regulates appetite and energy expenditure.

7. Adiponectin: continuous, interval (μg/mL). Adiponectin is a hormone that regulates the metabolism of glucose.

8. Resistin: continuous, interval (ng/mL). It is a hormone that it is suggested may promote insulin resistance.

9. MCP-1: continuous, interval (pg/dL). MCP-1 is a type of protein molecule that plays a role in the inflammatory response of the body and is linked to the development of obesity and insulin resistance.

10. Class: nominal (1=Healthy controls, 2=Patients with Breast Cancer)

The attributes 1,2 are anthropometric data, whereas the attributes 3-9 are parameters gathered in blood analysis.

3.2 Data issues

We checked for NaN and null values in the whole dataset but no missing values detected. We also checked for unexpected values in the attributes which possible values are explicitely determined. For example, a value different than 1 or 2 for the Class, would be an invalid value.

3.3 Summary statistics of the attributes

Attribute	mean	variance	median	range	standard deviation
Age	57,30	257,38	56,00	65,00	16,04
BMI	27,58	24,98	27,66	20,21	5,00
Glucose	97,79	503,01	92,00	141,00	22,43
Insulin	10,01	100,49	5,92	56,03	10,02
HOMA	2,69	13,15	1,38	24,58	3,63
Leptin	26,62	364,83	20,27	85,97	19,10
Adiponectin	10,18	46,43	8,35	36,38	6,81
Resistin	14,73	152,20	10,83	78,89	12,34
MCP.1	534,65	118624,06	471,32	1652,60	344,42
Classification	1,55	0,25	2,00	1,00	0,50

Table 1: Data Statistics

From the above table we can extract the following information:

1. The variance and standard deviation in Age,Glucose, Insulin,Homa,Leptin,Resistin, and MCP-1 are significantly high compared to the mean, indicating that these properties are fairly spread out. This might imply that there is substantial variation in the ages and levels of these metabolic markers among people. The variance and standard

deviation of Adiponectin, on the other hand, are fairly high when compared to the mean, indicating that while adiponectin levels in the sample are slightly spread out, the variability is not as extreme as in the other properties. Finally, as compared to the mean, the variation and standard deviation of BMI are quite low. This indicates that the BMI values in the group are not widely distributed, and that the majority of the individuals in the group have BMI values that are reasonably near to the mean.

2. The age, glucose, insulin, HOMA, Leptin, Resistin, and MCP.1 value range is quite broad, suggesting that there may be outliers or extreme values in the data set.

3. In BMI and Age, the median and the mean are very close to each other which suggests that their distributions are relatively symmetrical.

4. The mean in Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP.1 is greater than the median which means that the their distribution is skewed to the right.

4 Data visualization

4.1 Outliers

To check for outliers we used boxplots as shown below. The data points seen outside of the boxes and the whiskers, are considered as outliers. However, despite the fact that many outliers in glycose, insulin, HOMA, adiponectin and resistin can be seen, we can not determine if they are irrelevant or not thus we consider them as valid values. It has to be noted, though, that the boxplots of the Figure 2 were constructed on the normalized values. Therefore, as the above mentioned attributes are highly skewed and appear to have many outliers, other normalization or transformation techniques could also be tested.

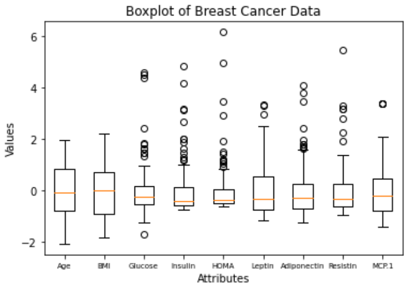


Figure 2: Boxplot of breast cancer data

4.2 Data distribution

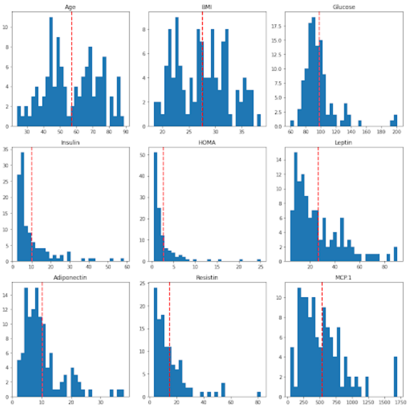


Figure 3: Distribution of all the attributes of our dataset

We examined the shape of the data distribution by constructing the histograms of the Figure 3. Each histogram consists of a series of bars, where the height of each bar represents the frequency of the value of the attribute within a certain range, called a bin. No attribute appears to follow a normal distribution as they do not show any symmetry around the mean and do not form the typical bell curve. We notice multiple peaks in every attribute and skews to one side for most of them which verifies the premises made in the summary statistics of the attributes. Therefore, in the next project when we will do the classification and regression tasks, we will not use statistical methods that assume normality, such as t-tests and linear regression. Instead we will make use of non-parametric methods or transformations, such as logarithmic or square root transformation, that may make the data more normal.

4.3 Covariance

Covariance measures how much one variable y can be expected to change when another variable x changes and visa-versa, but it is affected by the scale of each attribute (Table 2). This can be overcome by standardizing with the empirical standard deviation of the attributes, leading to the correlation (Table 3).

Attribute	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
Age	259,62	0,69	83,52	5,27	7,45	31,72	-24,24	0,55	75,03	-0,35
BMI	0,69	25,20	15,70	7,34	2,09	54,85	-10,40	12,15	389,05	-0,33
Glucose	83,52	15,70	507,38	114,44	57,12	131,83	-18,82	81,31	2063,87	4,32
Insulin	5,27	7,34	114,44	101,36	34,18	58,22	-2,16	18,30	607,21	1,39
HOMA	7,45	2,09	57,12	34,18	13,26	22,86	-1,40	10,43	326,96	0,52
Leptin	31,72	54,85	131,83	58,22	22,86	368,00	-12,52	60,91	92,96	-0,01
Adiponectin	-24,24	-10,40	-18,82	-2,16	-1,40	-12,52	46,83	-21,40	-475,08	-0,07
Resistin	0,55	12,15	81,31	18,30	10,43	60,91	-21,40	153,53	1570,74	1,41
MCP.1	75,03	389,05	2063,87	607,21	326,96	92,96	-475,08	1570,74	119655,57	15,79
Classification	-0,35	-0,33	4,32	1,39	0,52	-0,01	-0,07	1,41	15,79	0,25

Table 2: Covariance matrix

4.4 Correlation

Correlation tells us how (linearly) related attributes are. A correlation of 0 means that x tells us nothing about y, a positive correlation tells us that when x is large y is also likely to be large and a negative correlation tells us that if x is large y will typically be small. The correlation coefficient is measured on a scale that varies from + 1 through 0 to -1.

Attribute	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
Age	259,62	0,69	83,52	5,27	7,45	31,72	-24,24	0,55	75,03	-0,35
BMI	0,69	25,20	15,70	7,34	2,09	54,85	-10,40	12,15	389,05	-0,33
Glucose	83,52	15,70	507,38	114,44	57,12	131,83	-18,82	81,31	2063,87	4,32
Insulin	5,27	7,34	114,44	101,36	34,18	58,22	-2,16	18,30	607,21	1,39
HOMA	7,45	2,09	57,12	34,18	13,26	22,86	-1,40	10,43	326,96	0,52
Leptin	31,72	54,85	131,83	58,22	22,86	368,00	-12,52	60,91	92,96	-0,01
Adiponectin	-24,24	-10,40	-18,82	-2,16	-1,40	-12,52	46,83	-21,40	-475,08	-0,07
Resistin	0,55	12,15	81,31	18,30	10,43	60,91	-21,40	153,53	1570,74	1,41
MCP.1	75,03	389,05	2063,87	607,21	326,96	92,96	-475,08	1570,74	119655,57	15,79
Classification	-0,35	-0,33	4,32	1,39	0,52	-0,01	-0,07	1,41	15,79	0,25

Table 3: Correlation matrix

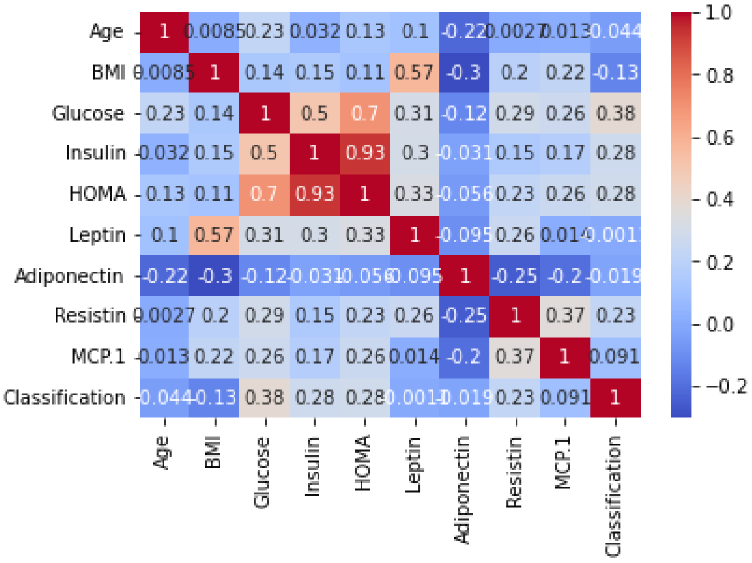


Figure 4: Correlation Heatmap

The heatmap indicates the correlation between all the 10 attributes. We note that insulin and HOMA are highly positively correlated with the correlation coefficient up to 0.93. HOMA is also correlated with glucose with coefficient of 0.7. This linear association between these attributes can also be seen in the matrix scatter plot below (Figure 5), where we scattered plotted all the pairs of attributes classified (healthy or patient) to assess possible dependence between the attributes. Regarding the correlation between the attributes and the target variable, we can say that no attribute has a very strong correlation with the class apart from glucose with coefficient of 0.38. This weak positive correlation between the

class and the glucose can suggest a possible predictive value of this attribute, while further investigation should be made to examine nonlinear relationships or other patterns in our data.
The fact that the correlation coefficients of the 10 attributes range from 1 to around -0.2 suggests that there is a mix of strong positive correlations, weak positive correlations, and moderate negative correlations among the attributes.

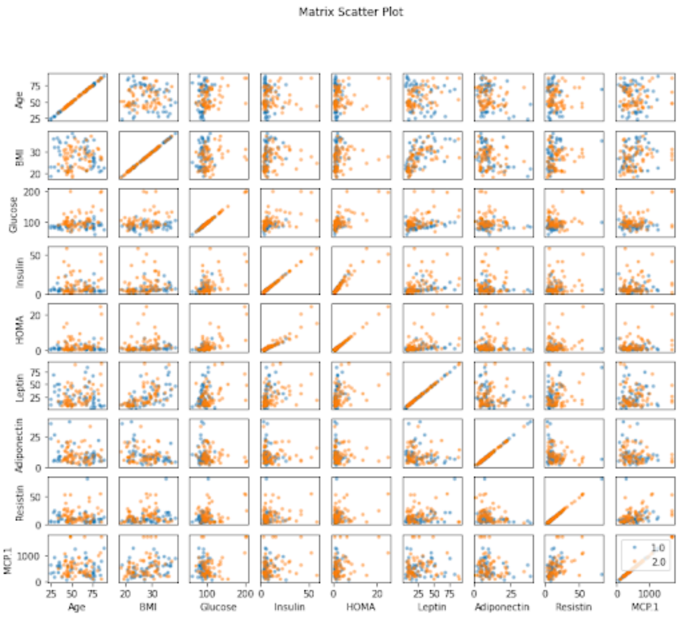


Figure 5: Scatter plot

In addition we made a Parallel coordinates plot in order to visualize the relationship between all the attributes (Figure 6). Data points are represented as lines that connect the values of each attribute, forming a polyline and creating a pattern that may help to identify clusters and trends in our data. As we can see individuals with breast cancer seem to have a deviation in the pattern in particular concerning the BMI, Glucose, Insulin, HOMA and Resistin levels. This may be an indication of a complex relationship between these attributes and the class and will be further investigated in the 2nd report.

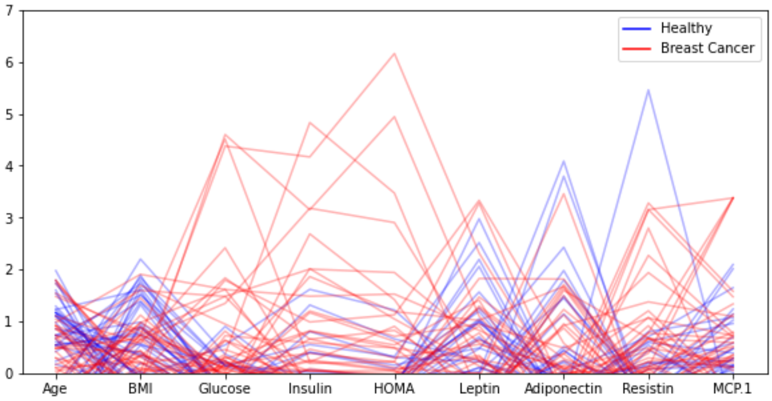


Figure 6: Parallel coordinates

4.5 Feasibility of the ML modeling aim

Combining the information we obtain from all the visualizations above, we can investigate whether the primary machine learning modeling aim appears to be feasible. For this purpose, we took into account the following observations:

- From the summary statistics (Table 1) as well as from the boxplots (Figure 2) in the outliers section, we conclude that no data points are required to be omitted,because there isn't any missing data nor observations that can be certainly considered as irrelevant.

- Having observed the histograms (figure 3), it is apparent that all the examined parameters deviate from a normal distribution, which will render the performance of non-parametric methods or transformations necessary, in the following steps of the analysis.

- The correlation matrix (Table 3) reveals the existence of several attributes that present a strong correlation between them, that are namely: HOMA and insulin, that score a correlation of 0.93, HOMA and glucose with a correlation coefficient 0.7, and leptin and BMI, that seem to be correlated, but to a smaller extend (0.57). However, the rest of the attributes show insignificant correlation.

Overall, we draw the conclusion that the data we are using will require further manipulation in order to achieve the modeling aim.

4.6 Principal component analysis

The goal of PCA is to provide insight into the underlying structure of our complex dataset and to identify the most important variables that drive the variation. The principal components are linear combinations of the original variables that capture the maximum amount of variance in the dataset. By reducing the number of dimensions in the data, PCA may help to simplify our analysis and visualization, while retaining the most important

information.
The first step was to explain the variation as a function of the number of PCA components included as shown below.(Figure 7)

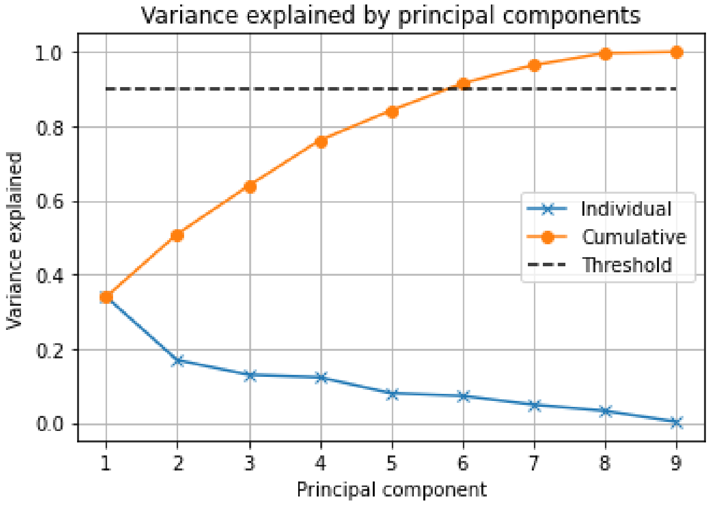


Figure 7: Amount of variation as a function of the number of PCA components included

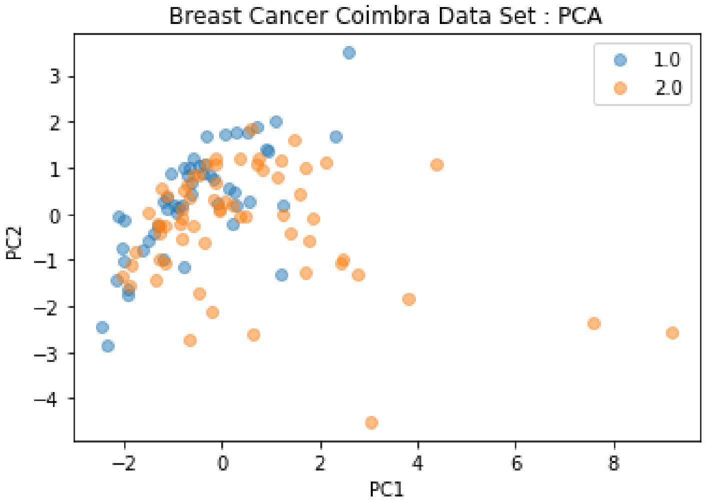


Figure 8: Data projected onto the first two principal components

Visualizing only the first 2 PC (Figure 8), we capture around 50% of the total variance. As it can be seen there is no clear separation or clustering among the data points. This may be due to the fact that there is no strong linear relationship among the attributes or that other PCs are needed to capture the complex patterns in the data. However, capturing 50% of the total variance is still a significant amount, and it suggests that there is some underlying structure in the data. Alternatively other techniques such as non-linear dimensionality reduction methods may be useful and will need further investigation in the next project. Lastly, 90% of the variance of the dataset can be explained by six principle components, indicating that the present dataset is of higher complexity. (Figure 7)

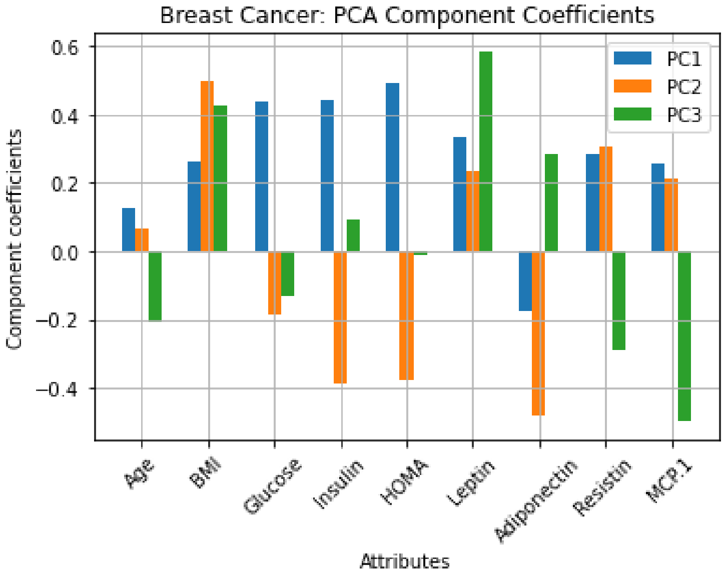


Figure 9: Principal directions of the first three PCA components

Regarding the principal directions of the PCA components we tried to interpret them by plotting the coefficients as vectors in the first 3 principal component space (Figure 9). PC1 represents the direction of maximum variance in the data. We can see that PC1 has high positive values for BMI, leptin, and adiponectin, and negative values for resistin, MCP1, and age, indicating that these variables are strongly correlated with each other in the same direction. This suggests that there is a group of individuals in the dataset who have high BMI, high levels of leptin and adiponectin, low levels of resistin and MCP1, and are relatively young. This group may have a distinct metabolic profile that sets them apart from other individuals in the dataset.

PC2 with high BMI, leptin, resistin and MCP1 levels, while negative values of glucose, insulin, HOMA, and adiponectin suggests that this principal component represents an axis of metabolic dysregulation. High levels of BMI, leptin, resisting, and MCP1 are all associated with insulin resistance and inflammation, which are known risk factors for various chronic diseases, including type 2 diabetes, cardiovascular disease, and certain cancers. On the other hand, negative values of glucose, insulin, HOMA, and adiponectin suggest improved insulin sensitivity and metabolic health [3]. Therefore, PC2 may represent a continuum of metabolic health, with positive values indicating a more dysregulated state and negative values indicating a more favorable metabolic profile.

Lastly, the high values of BMI, leptin, and adiponectin suggests that PC3 reflects the overall adiposity of the individuals. Adiponectin is an adipokine that has been shown to have anti-inflammatory and anti-carcinogenic properties. Therefore, the high loading of adiponectin in PC3 may indicate a protective effect against cancer. On the other hand, the negative loading of age, resistin, and MCP1 may suggest a decrease in inflammation and immune function with aging, as these molecules have been associated with inflammation also.[4,5]

5 Discussion

The examination began with a brief discussion of each attribute so that the observer could understand their biological significance as well as their types. Additional review of the data found no missing or aberrant numbers; nonetheless, there are outliers that cannot be regarded irrelevant, therefore they are considered valid values.
Moving on to the summary statistics, a look at table 1 reveals some interesting information about the distribution of each attribute and how the data was spread. These findings not only confirmed the existence of outliers, but they were also confirmed by the following data distribution graphs developed. Additionally, the correlation was examined, and it was shown that HOMA was favorably corelated with insulin and glucose, although no characteristic has a very significant correlation with the class, with the exception of glucose, which has a coefficient of 0.38. This small positive connection between class and glucose suggests that this attribute may have predictive value, but further research should be done to analyze nonlinear correlations or other patterns in our data. Furthermore, a parallel coordinates plot revealed that people with breast cancer appear to have a divergence in the pattern, particularly in terms of BMI, glucose, insulin, HOMA, and resistin levels. This might indicate a complicated link between these qualities and the class.
Ultimately, PCA indicated that the first two principal components account for around half of the overall variance. which is still a considerable amount and implies that there is some underlying structure in the data, whereas six principle components can explain 90% of the variation in the dataset.

6 Exam questions

Question 1: option D

- x1 (Time of day) is interval, because it appears in the form of numerical values, where the distance between the two points is standardized and equal (3o min).
- y (Congestion level) is ordinal, as its possible values have a meaningful order and ranking, but the magnitude between successive values is not known.
- x6 (Traffic lights) and x7 (Running over) are ratio because in both cases zero value has a physical meaning.

Question 2: option A
To answer this question we used the definition of p-distance:

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_p = \begin{cases} \left(\sum_{i=1}^{M} |x_i - y_i|^p\right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty \\ \max\{|x_1-y_1|, |x_2-y_2|, \ldots, |x_M-y_M|\} & \text{if } p = \infty. \end{cases}$$

where M: the number of attributes in the dataset
So for p=∞ :

$$d_p(x,y) = max|26-9|, |0-0|, ..., |0-0| = 7$$

Question 3: option A
The variance is given by the formula:

$$V_{arianceExpl} = \frac{\sigma_k^2}{\Sigma_{j=1}^{M}\sigma_j^2}$$

where M: the number of attributes in the dataset
So for the first 4 PC:

$$\frac{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2+\sigma_5^2} = \frac{13.9^2+12.47^2+11.48^2+10.03^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2}$$

= 0.86679314748 > 0.8

Question 4: option D
The projection onto PC2 is:

- x1: Time of Day = -0.5 (low)
- x2: Broken truck = 0.23 (high)
- x3: Accident victim = 0.23 (high)
- x4: Immobilized bus = 0.09 (high)
- x5: Defects = 0.8 (high)

Question 5: option A
The Jaccard similarity was calculated using the type:

$$J(S_1, S_2) = \frac{f_{11}}{f_{01}+f_{11}+f_{00}}(1)$$

where f11 is the number of common words in S1 S2 (2: the, words), while f11 is the number of non common words between S1, S2 (11: bag, of, representation, becomes, less,

parsimoneous, if, we, do, not, stem).
So from (1), we have:

$$J(S_1, S_2) = \frac{2}{11+2} = \frac{2}{13} = 0.15384615384$$

Question 6: option B

The probability an observation to have x2 = 0 given light congestion, was calculated as it is shown below:

$$p(x_2 = 0|y = 2) = p(x_2 = 0, x7 = 0|y = 2) + p(x_2 = 0, x_7 = 1|y = 2) = 0.081 + 0.03 = 0.84$$

List of Figures

List of Tables

Nomenclature

BC	Breast Cancer
BMI	Body Mass Index
MCP-1	Monocyte Chemoattractant Protein-1
PC	Principal Component
PCA	Principal Component Analysis

7 References

1. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

2. Patricio M, Pereira J, Crisostomo J, Matafome P, Gomes M, SeiÃ§a R, Caramelo F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. 2018 Jan 4;18(1):29. doi: 10.1186/s12885-017-3877-1. PMID: 29301500;

3. Jiralerspong S, Goodwin PJ. Obesity and Breast Cancer Prognosis: Evidence, Challenges, and Opportunities. J Clin Oncol. 2016 Dec 10;34(35):4203-4216. doi: 10.1200/JCO.2016.68.4 Epub 2016 Nov 7. PMID: 27903149.

4. Eggink HM, van Nierop FS, Schooneman MG, Boelen A, Kalsbeek A, Koehorst M, Ten Have GAM, de Brauw LM, Groen AK, Romijn JA, Deutz NEP, Soeters MR. Transhepatic bile acid kinetics in pigs and humans. Clin Nutr. 2018 Aug;37(4):1406-1414. doi: 10.1016/j.clnu.2017.06.015. Epub 2017 Jun 19. PMID: 28669667.

5. Elman JS, Li M, Wang F, Gimble JM, Parekkadan B. A comparison of adipose and bone marrow-derived mesenchymal stromal cell secreted factors in the treatment of systemic inflammation. J Inflamm (Lond). 2014 Jan 7;11(1):1. doi: 10.1186/1476-9255-11-1. PMID: 24397734; PMCID: PMC3895743.