Benign and Malignant Nuclei in Breast Cancer Diagnosis: a Winconsin case study

Carolina Bellani, Gonçalo Passão, Sofia Jerónimo

Abstract

Breast cancer is among the five most common cancers in the world and it is essential to have accurate diagnosis methodologies to classify malignant and benign tumors in order to decide which is the most appropriate. Here, we performed a K-means and K-medoids clustering to understand the cell characteristics in samples of patients' breast masses. The best result was obtained when using K-medoids and it has an accuracy of 91.4% to distinguish between malignant and benign cells. We designed a decision-tree to understand which variables are more significant and to visualize how to classify the breast mass biopsies. The most important variables for the classification is the worst radius, followed by the worst concave points and the mean of the texture. The accuracy, in this case, was 94.2%.

Keywords

Breast cancer, Breast tumors, Diagnosis, FNA, Clustering, Decision-tree

Introduction

A neoplasm is an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues, and persists in the same excessive manner after cessation of the stimulus which evoked the change (Willis 1953). Cancer can start almost anywhere in the human body, which is made up of 37.200 billion cells (Bianconi et al. 2013). As these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original one. Unlike malignant tumors, benign tumors do not spread into, or invade, nearby tissues. Breast cancer refers to a pathology in which a tumor develops in the breast tissue.

Breast cancer is amongst the most common type of cancer in both sexes since 1975 and causes around 411,000 annual deaths worldwide (Parkin and Fernandez 2006). It is predicted that the incidence for worldwide cancer will continue to increase, with 23,6 million new cancer cases each year by 2030, corresponding to 68% more cases in comparison to 2012 (Bray et al. 2012).

Diagnosis in an early stage is essential to the facilitate the subsequent clinical management of patients and increase the survival rate of breast cancer patients. Mammography is the most common mass screening tool for an early detection of breast cancers because of its sensitivity in recognising breast masses (Subramaniam et al. 2006). After detection of suspicious breast masses, a biopsy test procedure would be carried out, such as Fine Needle Aspirates (FNA' s) (O. L. Mangasarian, Street, and Wolberg 1995). This method has been showed to be safe, cost-effective, accurate and fast (Nasuti, Gupta, and Baloch 2002). A small drop of viscous fluid is aspired from the breast masses by making multiple passes with a 23-gauge needle as negative pressure was applied to an attached syringe (W H Wolberg, Street, and Mangasarian 1993). The aspirated sample is mounted onto a silane-coated glass slide, then fixed and stained to be analysed under the microscope (W H Wolberg, Street, and Mangasarian 1993). Then, a small region of the breast mass cells is photographed in a gray scale image and

Email addresses: m20170098@novaims.unl.pt (Carolina Bellani), m20170450@novaims.unl.pt (Gonçalo Passão), m20170070@novaims.unl.pt (Sofia Jerónimo)

further analysed using an image analysis program 'Xcyt' (W. N. Street, Wolberg, and Mangasarian 1993),(W H Wolberg, Street, and Mangasarian 1993),(William H Wolberg, Street, and Mangasarian 1994). This program uses a curve-fitting to determine the edges of the nuclei from initial dots manually placed near these edges by a mouse (Figure 1).

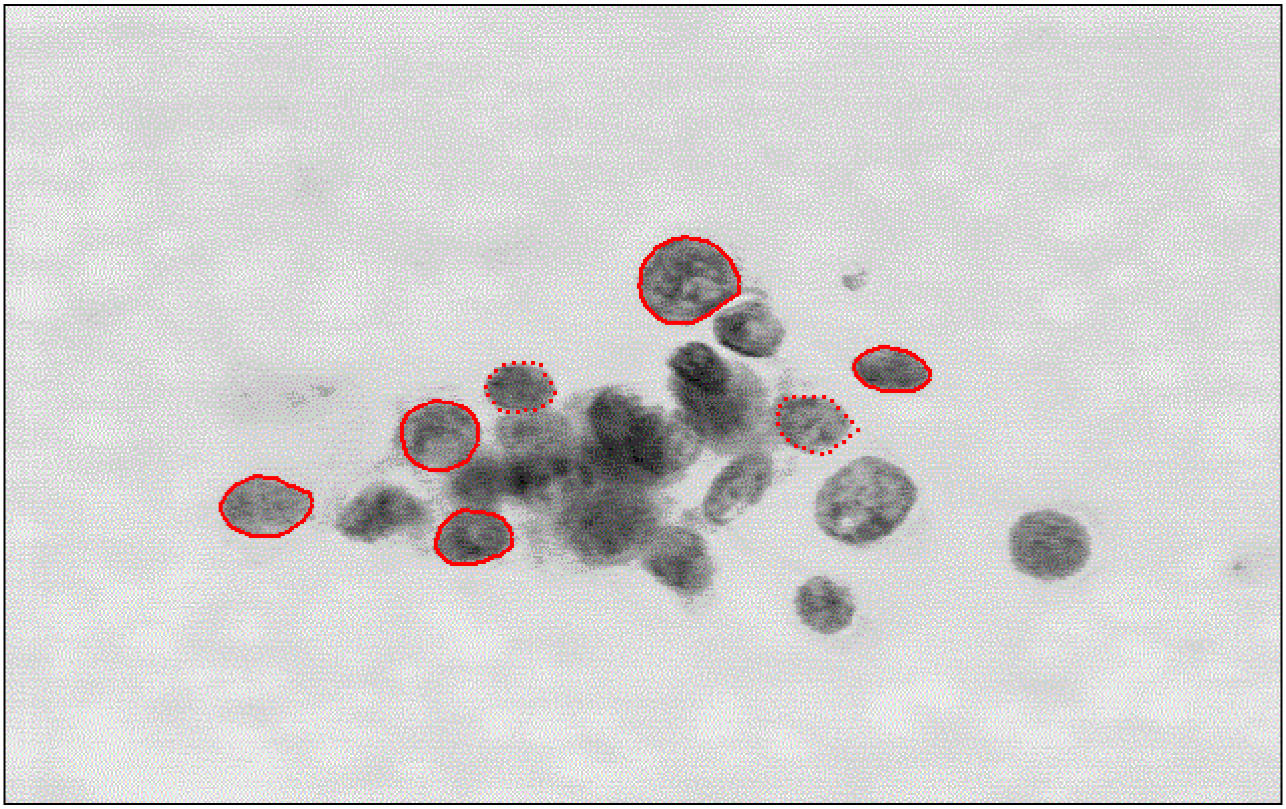

Figure 1. Magnified image of a malignant breast FNA. The edges of the visible cell nuclei were manually placed with a mouse (red dots), 'Xcyt' program will after outline the nuclei (red circle). The interactive diagnosis process takes about 5 minutes per sample.

The data used for this project was collected in 1993 by the University of Wisconsin and it is composed by the biopsy result of 569 patients in Wisconsin Hospital. Ten features were computed for each cell nucleus: 1) radius (mean of distances from center to points on the perimeter), 2) texture (variance of grey-scale values inside the boundary), 3) perimeter, 4) area, 5) smoothness (local variation in radius lengths), 6) compactness (perimeter^2/area - 1.0, this dimensionless number is at a minimum with a circular disk and increases with the irregularity of the boundary, but this measure also increases for elongated cell nuclei, which is not indicative of malignancy), 7) concavity (severity of concave portions of the contour), 8) concave points (number of concave points of the contour), 9) symmetry and 10) fractal dimension of the boundary ('coastline approximation' - 1, described by Mandelbrot (Mandelbrot 1982) a higher value corresponds a less regular contour and thus to a higher probability of malignancy). The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 variables. From this diagnosis, 357 of the cases were classified as benign tumors and 212 were considered malignant tumors. All cancers and some of the benign masses were histologically confirmed (W H Wolberg, Street, and Mangasarian 1993).

To characterise all the features that distinguish the benign from the malignant cells, we did a K-medoid analysis. We expect to have variables that can explain the diagnosis result (benign vs. malignant) and that can be used in the future to accurately classify new patients' samples. Considering our clustering algorithm, we calculated the accuracy, sensitivity and specificity. Finally, we built a decision-tree model to better understand the classification of benign versus malignant cells.

Methods

The breast cancer dataset is available online on UCI Machine Learning Repository ("UCI Machine Learning Repository" 2017). This dataset does not contain missing values. This dataset is composed of 32 variables, such as the patient ID, the diagnosis result (benign or malignant) and the mean, standard error (SE) and 'worst' or largest (mean of the three largest values) measures for the ten nuclei features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension). All the exploratory analysis and clustering analysis were done in R (Core Team 2015) using RStudio (RStudio Team 2015). The decision-tree was implemented using Python (Team 2017). We checked the distributions of the 30 variables (Figure S1-S3) and we did Shapiro tests to analyse if the distributions were normal. We decided to standardise the units of measurement using Z-score. We checked the presence of outliers through boxplots (Figure S4-S6), but we considered important to keep those values in the analysis because we considered this information relevant to characterize malignant and benign tumors. We carried out a matrix of Person correlation coefficients considering the 30 nuclei features. To avoid redundancy and relevancy, we used the function 'princomp' to calculate the Principal Components Analysis (PCA) and selected seven components to avoid correlated variables that can be detrimental to our clustering analysis (James et al. 2013). We performed two different clustering methods: K-means using the function 'kmeans' from the package 'cluster' (Maechler et al. 2015) and K-medoids using the function 'pamk' from the package 'fpc' (Hennig 2015). For k-means we set 100 as the maximum number of iterations, used the Hartigan-Wong algorithm and set random initialisation seeds. To choose the number of clusters for K-means, we inspected the elbow plot and use the function 'NbClust' from the package 'NbClust' (Charrad et al. 2014). This function provides 26 different indices for determining the number of clusters. We calculated the specificity, sensibility and accuracy of our clustering models. The specificity of a test refers to how well a test identifies patients who do not have a disease. High specificity is more useful when the result is positive, for ruling in patients who have a certain disease. The sensitivity refers to the ability of the test to correctly identify those patients without the disease. High sensitivity is useful for ruling out a disease if a person has a negative result (Lalkhen and McCluskey 2008). To understand the importance of the variables, we decided to perform and visualise the decision-tree model. For this, we used function 'DecisionTreeClassifier' from the library 'Sklearn'. Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. This is a representation used very commonly in medical cases (Hastie, Tibshirani, and Friedman 2009). The tree stratifies the population into strata of high and low outcome, on the basis of patient's characteristics of the breast cancer.

Results and Discussion

The Pearson correlation coefficients indicates that there is high positive correlation (>0.90, Figure 2) between the variables 1) 'mean perimeter', 'mean area', 'worst area', 'worst radius' and 'worst perimeter'; 2) 'se area', 'se radius' and 'se perimeter'; 3) 'mean concavity' and 'mean concave points'; 4) 'worst concave points' and 'mean concave points'. There are not strong negative correlations between the variables (all correlations are > -0.28, Figure 2).

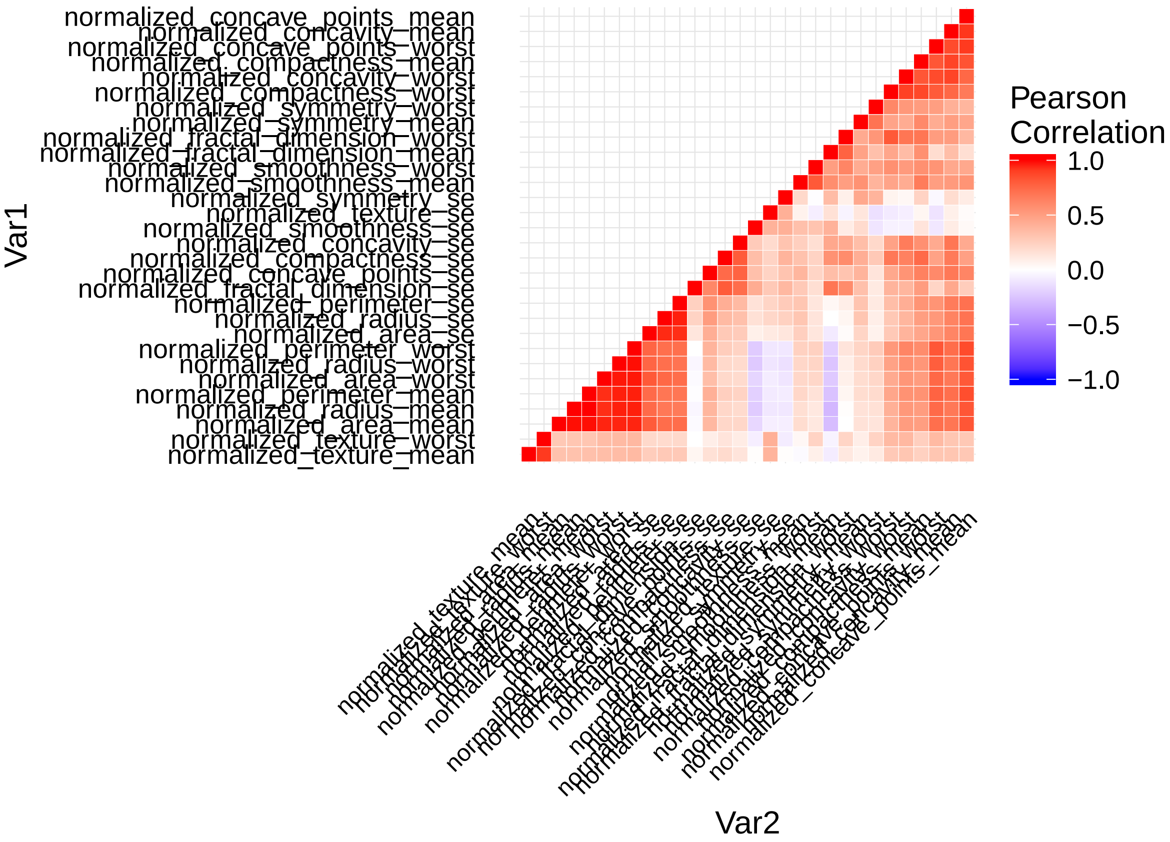

Figure 1. Matrix of the Pearson correlation coefficients between the 30 variables. The red colour indicates the positive correlations and the blue colour the negative correlations.

We considered the first seven principal components from the PCA which explains 91% of the cumulative proportion variance of the data (Table 1).

Table 1. Standard deviation, proportion of variance and cumulative proportion for the 7 components from the PCA analysis.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	3.64	2.39	1.68	1.41	1.28	1.10	0.82
Proportion of variance	0.44	0.19	0.09	0.07	0.05	0.04	0.02
Cumulative proportion	0.44	0.63	0.73	0.79	0.85	0.89	0.91

Considering the elbow plot (Figure 3) and the 26 indices returned by the 'NbClus' function, the best number of clusters is 3 (among all 26 indices: 8 proposed 2 as the best number of clusters, 9 proposed 3 as the best number of clusters. According to the majority rule, the best number of clusters is 3).

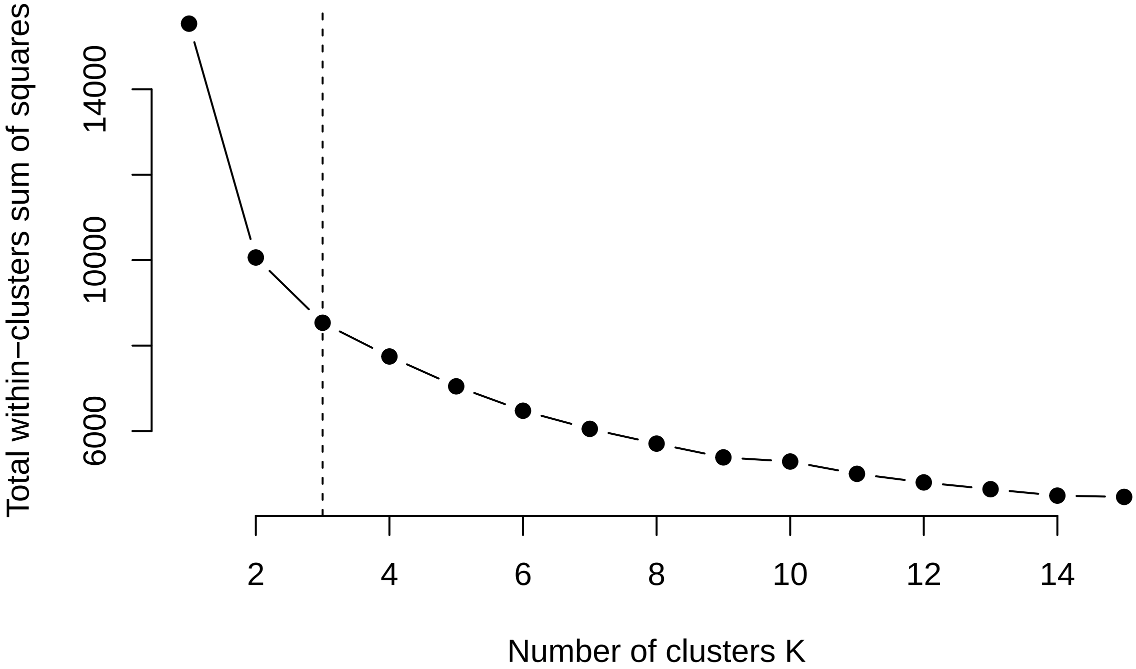

Figure 3. Elbow plot.

For the K-means clustering method, we classified each cluster as a representative of benign or malignant tumors considering the diagnosis results that have a higher frequency. We classified the first cluster as 'Malignant' because it has 65 cases of malignant versus 36 cases of benign tumors, the second cluster was considered as malignant because it only represents malignant tumors and finally, the third cluster is representative of benign tumors (Table 2). The latter cluster included more patients than the first or second cluster and contains around 63% of the data.

Table 2. Contingency table for the frequency of benign or malignant results included in each cluster using K-means clustering method.

Diagnosis	Cluster 1	Cluster 2	Cluster 3	Total
Benign	36	0	321	357
Malignant	65	110	37	212
Total	101	110	358	569

To calculate the specificity, sensitivity and accuracy of our k-means model, we made a contingency table between the K-means malignant cluster (cluster 1 and 2) and benign cluster (cluster 3) versus the real diagnosis results (Table 3). The specificity was 89.9%, the sensitivity was 82.5% and accuracy was 87.2%.

Table 3. Contingency table between the frequency of the K-means diagnosis result versus real diagnosis result' using K-means algorithm.

	Real malignant	Real benign	Total
Cluster malignant	175	36	211

	Real malignant	Real benign	Total
Cluster benign	37	321	358
Total	212	357	569

Table 4. Centroid average for the seven components using the K-means clustering. The cluster 1 has 101 patients, cluster 2 has 110 patients and cluster 3 has 358 patients.

	Comp. 1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Cluster 1	-2.35	2.96	-0.74	-0.29	0.30	-0.17	-0.03
Cluster 2	-5.34	-1.89	0.43	0.13	-0.18	-0.03	-0.01
Cluster 3	2.30	-0.26	0.07	0.04	-0.03	0.05	0.01

Table 5. Centroid average for the seven components using the K-medoids clustering. The cluster 1 has 175 patients, cluster 2 has 394 patients.

	Comp. 1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Cluster 1	-4.68	-0.97	-0.29	-1.07	-0.13	0.50	0.13
Cluster 2	2.02	-0.25	-0.65	0.06	-0.05	-0.07	0.05

The number of clusters are two. Comparing with the results of ‘NbClus’ there is coherence; in fact, 8 indexes proposed 2, 9 indexes 3. We can notice the second clustering is bigger than the first. Considering the cross tabulation between “Diagnosis” and “clustering” (Table 6), the first segmentation has prevalence of benign cases and the second segmentation of malignant cases.

Table 6. Contingency table for the frequency of benign or malignant results included in each cluster using K-medoid clustering method.

Diagnosis	Cluster 1	Cluster 2	Total
Benign	6	351	357
Malignant	169	43	212
Total	175	394	569

Table 7. Contingency table between the frequency of the K-medoid diagnosis result versus real diagnosis result’ using K-means algorithm.

	Real malignant	Real benign	Total
Cluster malignant	169	43	212
Cluster benign	6	351	357
Total	175	394	569

We calculated the specificity, sensitivity and accurancy for K-medoids. For this model, the specificity was 96.4%, the sensitivity was 96.6% and accuracy was 91.4% (Table 7). The accurancy of K-medoids has a higher value then k-means. The K-means algorithm uses squared euclidean distance which places the highest influence on the largest distances. This causes the procedure to lack robustness against outliers that produce very large distances. These restrictions can be removed at the expense of computation, using K-medoids algorithm (Hastie, Tibshirani, and Friedman 2009).

Thus, we can notice that the results of K-medoids are better: it uses less number of clusters and the three indexes (specificity, sensitivity, accuracy) are higher than the results of K-means.

We examined in depth the profile segmentations of K-medoids and the centroid averages were calculated for each variable (Table 8-10).

Table 8. Centroid average for each mean feature of the nuclei.

	Radius	Texture	Perimeter	Area	Smoothness
Cluster 1	18.35	24.30	121.60	1071.26	0.10
Cluster 2	12.77	18.14	82.10	511.84	0.09

	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
Cluster 1	0.15	0.18	0.09	0.20	0.06
Cluster 2	0.08	0.05	0.03	0.17	0.06

Table 9. Centroid average for each SE feature of the nuclei.

	Radius	Texture	Perimeter	Area	Smoothness
Cluster 1	0.66	1.37	4.81	83.98	0.001
Cluster 2	0.26	1.04	1.76	18.68	0.006

	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
Cluster 1	0.035	0.045	0.015	0.022	0.004
Cluster 2	0.016	0.017	0.008	0.017	0.002

Table 10. Centroid average for each 'worst' measure feature of the nuclei.

	Radius	Texture	Perimeter	Area	Smoothness
Cluster 1	22.44	33.14	151.07	1586.89	0.14
Cluster 2	14.46	24.45	93.95	662.82	0.13

	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
Cluster 1	0.42	0.50	0.19	0.34	0.09
Cluster 2	0.20	0.19	0.09	0.287	0.08

It is possible to see that the cluster 1, defined as 'Malign', has bigger centroid averages than cluster 2, defined as 'Benign', with the exception of the variable SE Smoothness.

The biggest differences in centroid averages between the two clusters are for the variable worst perimeter, followed by mean concave point, worst radius and mean concavity. Therefore, they are the most important variables to describe our K-medoid model.

We compared the variables' average centroid of each cluster considering the variables' average population (Figure 3.). Considering cluster 1, the variables mean perimeter, worst radius, worst perimeter and worst area have the centroids farthest from the average of the population. Concerning cluster 2, the variables mean concavity, SE radius, SE perimeter and SE concave point are the ones further from the average population.

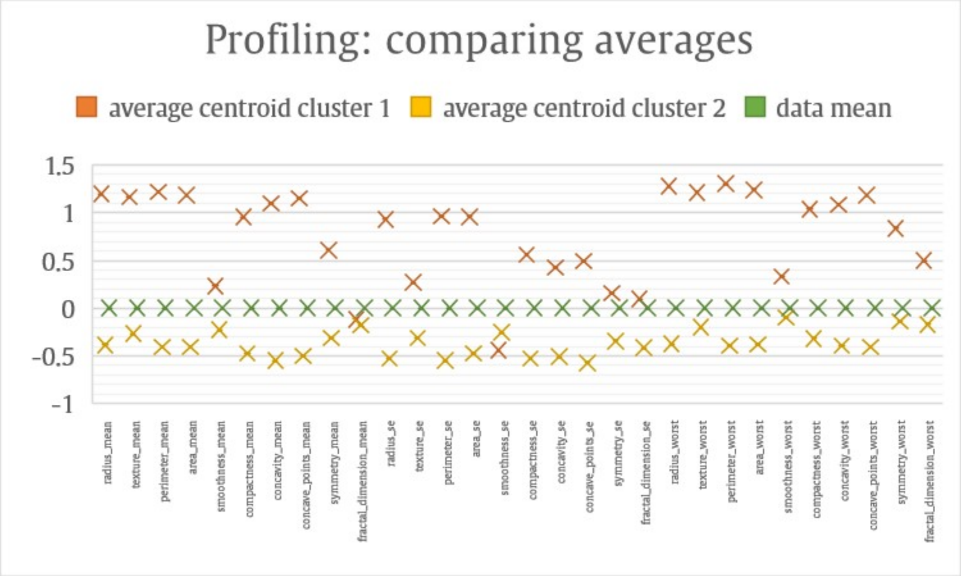

Figure 3. Profiling considering the averages of the two clusters and the average of the population.

In this case, we have more samples of benign than malignant, so we considered only the Gini coefficient because this coefficient splits the largest class into one pure node, and all other nodes into a different node. In opposite, entropy puts their emphasis on balancing the sizes at the two children nodes (Breiman 1996).

Considering the Gini coefficients, the decision tree has an accuracy of 94,2%, a sensitivity of 96,3% and a specificity as 90,9%.

The decision tree shows the most important variables for the decisions: worst radius, worst concave points and mean texture. It is possible to accurantly classify the benign nuclei when both worst radius is smaller or equal to 16.80 and worst concave points is smaller or equal to 0.14 (number of cases= 333, Gini coefficient=0.03). However when the worst concave points are higher than 0.14, it is classified as malignant sample, but this classification has a higher gini coefficient (0.48). In the other way, when the worst radius is higher than 16.80, the most important variable to classify malignant tumors is mean texture higher than 16.11. When the mean texture is smaller than 16.11, it is considered a benign tumor (number of cases=173, Gini coefficient=0.02).

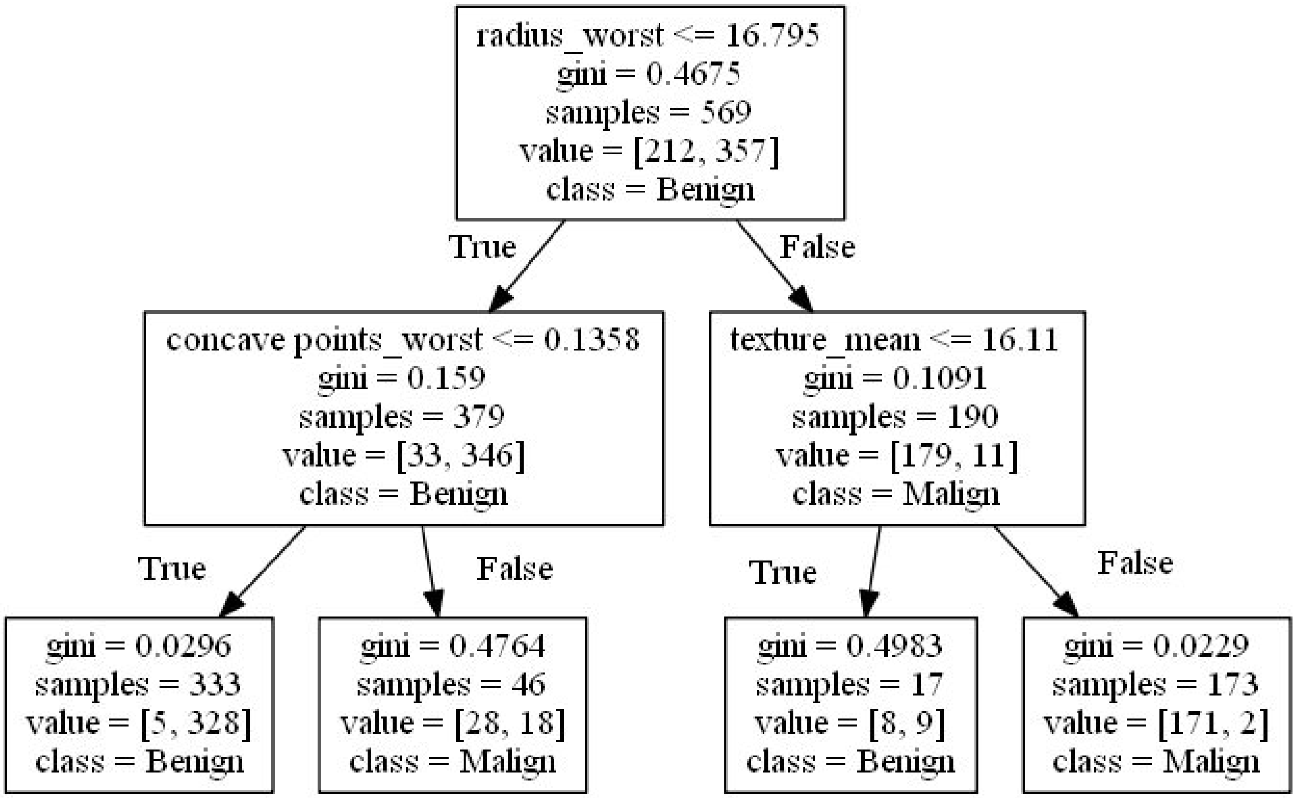

Figure 4. Decision tree considering Gini coefficients.

Conclusions

We collected the best results obtained with K-medoids algorithm and the decision tree model.

The K-medoids has an accuracy of 91.4%, sensitivity of 96.6% and specificity of 94.6% while the decision tree presented an accurancy of 94.2%, a sensitivity of 96.3% and specificity of 90,9%. For both of the models, the worst radius cell characteristic was the most important variable to classify the samples.

We would like to highlight that the clustering and the decision tree have different roles in the analysis. The clustering groups the data points considering all the variables selected, whereas the decision tree creates a path to follow selecting the variables by their importance. Having this in mind, the results of the decision tree could be used to identify potential bearers of the condition with just a small amount of information, while the results of the clustering should be used to identify and rule the individuals affected by cancer.

For future studies, we consider important to verify our results by using a bigger dataset, so the results can have a higher statistical significance. This would also allow to use different data for training, test and validate and therefore have more robust statistical models. It is very important to improve our models because the diagnosis is a important step for the cancer treatment and have a big impact on human survival.

Acknowledgements

We would like to thank Jorge Antunes for all the help answering our doubts throughout this paper.

References

Bianconi, Eva, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, et al. 2013. "An estimation of the number of cells in the human body." *Annals of Human Biology* 40 (6): 463–71. doi:10.3109/03014460.2013.807878.

Bray, Freddie, Ahmedin Jemal, Nathan Grey, Jacques Ferlay, and David Forman. 2012. "Global cancer transitions according to the Human Development Index (2008–2030): a population-based study." *The Lancet Oncology* 13 (8): 790–801. doi:10.1016/S1470-2045(12)70211-5.

Breiman, Leo. 1996. "Technical note: Some properties of splitting criteria." *Machine Learning* 24 (1): 41–47. doi:10.1007/BF00117831.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software*. http://www.jstatsoft.org/v61/i06/.

Core Team. 2015. "R: A language and environment for statistical computing." Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Vol. 99. Springer Series in Statistics 466. New York, NY: Springer New York. doi:10.1007/978-0-387-84858-7.

Hennig, Christian. 2015. "fpc: Flexible Procedures for Clustering." http://cran.r-project.org/package=fpc.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics 9-12. New York, NY: Springer New York. doi:10.1007/978-1-4614-7138-7.

Lalkhen, Abdul Ghaaliq, and Anthony McCluskey. 2008. "Clinical tests: sensitivity and specificity." *Continuing Education in Anaesthesia Critical Care & Pain* 8 (6). British Journal of Anaesthesia: 221–23. doi:10.1093/bjaceaccp/mkn041.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2015. "cluster: Cluster Analysis Basics and Extensions."

Mandelbrot, Benoit B. 1982. "The fractal geometry of nature." *Earth Surface Processes and Landforms* 8 (4): 460. doi:10.1002/esp.3290080415.

Mangasarian, Olvi L., W. Nick Street, and William H. Wolberg. 1995. "Breast Cancer Diagnosis and Prognosis Via Linear Programming." *Operations Research* 43 (4): 570–77. doi:10.1287/opre.43.4.570.

Nasuti, Joseph F., Prabodh K. Gupta, and Zubair W. Baloch. 2002. "Diagnostic value and cost-effectiveness of on-site evaluation of fine-needle aspiration specimens: Review of 5,688 cases." *Diagnostic Cytopathology* 27 (1): 1–4. doi:10.1002/dc.10065.

Parkin, D. Maxwell, and Leticia M. G. Fernandez. 2006. "Use of Statistics to Assess the Global Burden of Breast Cancer." *The Breast Journal* 12 (s1): S70–80. doi:10.1111/j.1075-122X.2006.00205.x.

RStudio Team. 2015. "RStudio: Integrated Development Environment for R, v. 0.99.896." Boston, MA: RStudio, Inc. http://www.rstudio.com/.

Street, W. N., W. H. Wolberg, and O. L. Mangasarian. 1993. "Nuclear feature extraction for breast tumor diagnosis." In *ISandT/SPIE International Symposium on Electronic Imaging: Science and Technology*, edited by Raj S. Acharya and Dmitry B. Goldgof, 1905:861–70. doi:10.1117/12.148698.

Subramaniam, Esugasini, Tan Kuan Liung, Mohd Yusoff Mashor, Nor Ashidi, and Mat Isa. 2006. "Breast Cancer Diagnosis Systems : A Review." *International Journal of The Computer, the Internet and Management* 14 (2): 24–35.

Team, Python Core. 2017. *Python: A Dynamic, Open Source Programming Language*. Python Software Foundation. https://www.python.org/.

"UCI Machine Learning Repository." 2017. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\protect\T1\textbraceleft/%\protect\T1\textbraceright28Diagnostic\protect\T1\textbraceleft/%\protect\T1\textbraceright29.

Willis, R.A. 1953. *The Spread of Tumors in the Human Body.* 3rd edition. Vols. 35-B. 2. Butterworth-Heinemann.

Wolberg, W H, W N Street, and O L Mangasarian. 1993. "Breast cytology diagnosis with digital image analysis." *Analytical and Quantitative Cytology and Histology* 15 (6): 396–404. http://www.ncbi.nlm.nih.gov/pubmed/8297430.

Wolberg, William H, W.Nick Street, and O.L. Mangasarian. 1994. "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates." *Cancer Letters* 77 (2-3): 163–71. doi:10.1016/0304-3835(94)90099-X.