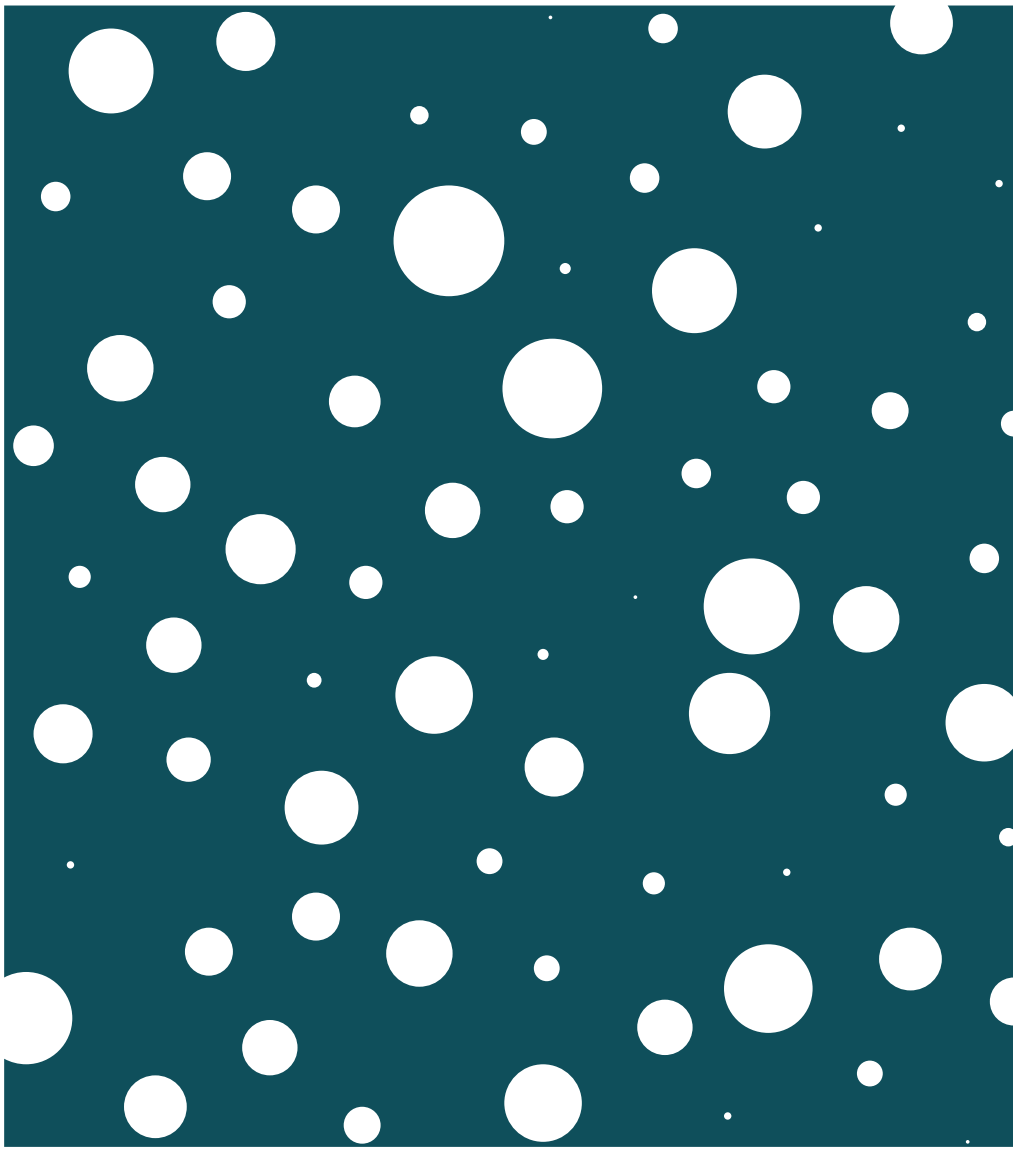
BREAST CANCER REPORT
DATA MINING

Predicting the malignancy of breast tumors



Liam Phan, Michael Bigler and Manuela Giansante

UNIVERSITÉ
DE GENÈVE

MASTER IN BUSINESS ANALYTICS
2022-2023

Table of Contents

1. Overview **2**

2. Data **3**

2.1 Data Analysis 3

2.2 Data Preparation 4

3. Models **5**

3.1 Supervised Learning 5

3.1.1 Logistic Regression 6

3.1.2 Classification Trees 7

3.1.3 K-Nearest Neighbors 9

3.1.4 Neural Network 10

3.1.5 Discriminant Analysis 11

3.1.6 Ensemble Methods 12

3.1.7 Majority of Vote 14

3.1.8 Average of Probabilities 14

3.2. Best Models 15

3.2.1 Best 3 Models on Validation 15

3.2.2 Best 3 Models On Test 16

3.3. Unsupervised Learning 17

3.3.1 K-Means Clustering 17

4. Conclusion **21**

5. References **22**

1. Overview

Each year, cancer affects millions of people worldwide, with more than 18.1 million cases reported in 2020[1]. 8.8 million women were diagnosed with cancer during this period, representing over 25.8% of all possible cancers[2]. The American Cancer Society projects that by 2022 more than 43,250 women will die from breast cancer, and more than 287,850 new cases of invasive breast cancer will be diagnosed[3] (USA only).

Our project focuses on this disease, specifically in the recognition of breast tumors and their potential condition. In medical terms, a tumor can be benign (non-cancerous) or malignant (cancerous). We used a dataset[4] from researchers at the University of Wisconsin in the Clinical Sciences Center and the Department of Computer Science. This dataset was created on the 1st of November 1995 and the method of data measurements is as follows: "Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.[5]"

The aim of this report is to analyze the selected dataset[6] and to apply different statistical and machine learning models from our Data Mining course, given by the University of Geneva and Professor Dr. Roy Welsch (MIT), accompanied by course assistant Anastasia Floru. We want to predict tumor type using these features and compare the best models that can remain globally accurate and perform better at correctly categorizing malignancies to allow patients to be followed up for possible treatment most immediately. We will use a partition to train our models and parameterise them efficiently with a validation partition, and finally we will use a test partition to see the prediction stability of our best models. Thus, the malignant tumor type will be considered as our positive output (M=1) and the benign tumors as our negative output (B=0), so we want to maximize the accuracy and sensitivity of our models.

---

[1] WCRF International. (n.d.). *Worldwide cancer data | World Cancer Research Fund International*. [online] Available at: https://www.wcrf.org/cancer-trends/worldwide-cancer-data/ .

[2] Idem.

[3] American Cancer Society (2022). *How Common Is Breast Cancer?* [online] Cancer.org. Available at: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html .

[4] Uci.edu. (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) .

[5] Idem.

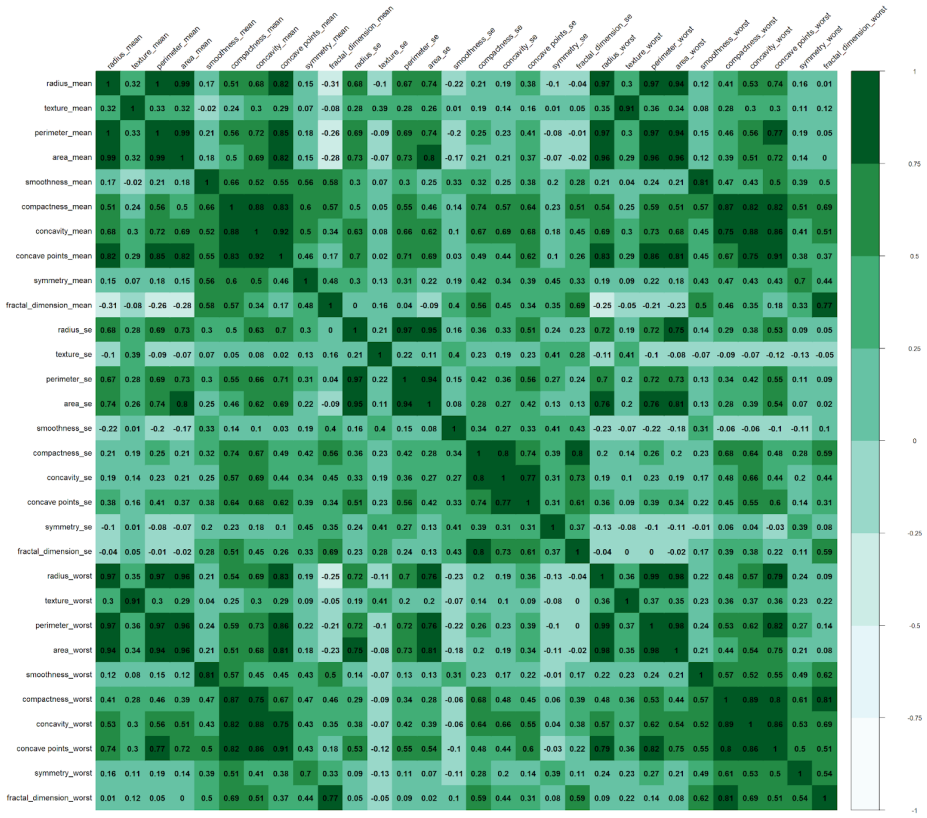
[6] Idem.

2. Data

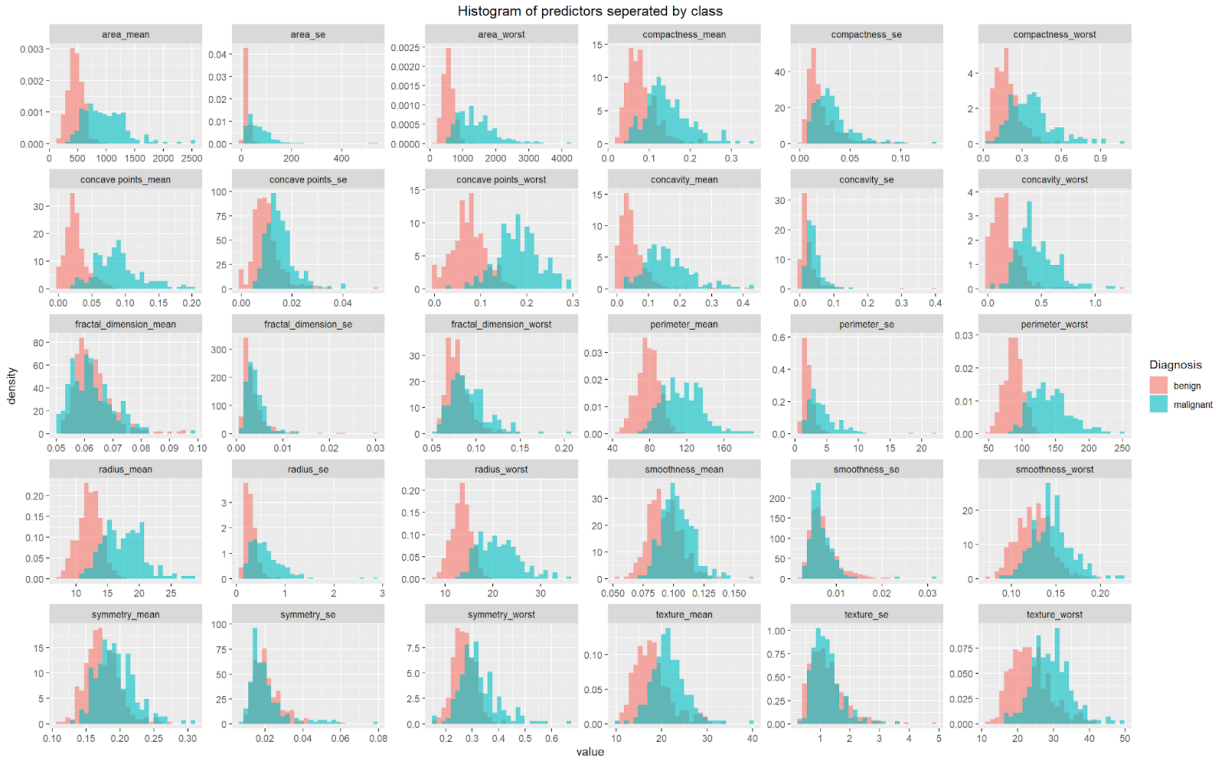
In this section we will analyze the data before we fit the models to it. Additionally, we will prepare the data for further use in this report.

2.1 Data Analysis

In this part we are going to show a correlation plot and the histograms for every variable. As it is categorical data we split the histograms and are showing the distributions per class. The prior distribution between the malignant and benign class is 37.3% to 62.7%. We can observe that we have risks of multicollinearity in our dataset, with variables producing redundant information.



When fitting supervised models to data (classifying) the distribution of it is particularly important. Therefore, we show a histogram for every variable in the dataset down below. Additionally, the data was split according to the diagnosis received. It is observable that there is an overlap between the classes, but we see that almost all variables differ in mean and variance. Normally the benign class has smaller values while the malignant ones are bigger. This makes sense in the way that malignant cancers tend to grow in an uncontrolled way and invade nearby tissues leading them to spread through the entire body.



2.2 Data Preparation

2.2.1 Missing values

Our data set does not contain NAs, therefore no further action is required.

2.2.2 Data Transformation

The data set is going to be normalized or standardized depending on the requirements of the model.
Respectively, we refer to this transformation when naming these two methods[7]:

$$x_{stand} = (x_i - \bar{x})/SD \quad \text{and} \quad x_{norm} = (x_i - x_{min})/(x_{man} - x_{min})$$

This is to maintain the coherence and comparability within our models results. Finally, as previously stated we are interested in the detection of malign cases. The "diagnosis" variable of our dataset is considered as the outcome dependent variable, it is transformed into a dummy 0,1 variable. The malign cases will be coded as 1s, the "Positive Class" is therefore 1.

[7] Zach (2021). *Standardization vs. Normalization: What's the Difference?* [online] Statology. Available at: https://www.statology.org/standardization-vs-normalization/ .

2.2.3 Data Partitioning and exclusion of ID column

The data set “ORIGINAL” is partitioned, with replacement, into a training set, a validation set and a test set. Respectively, they make up the 50%, the 30% and the 20% of the entire dataset.

The training set will be used in the building of the models, therefore used in the variable selection where necessary, and optimization. The validation sample is going to be used to judge the classification performance of every model, therefore drawing confusion matrices and yielding class and probabilities performances.

Finally, the test set after the top 3 models have been selected will be used for stability. The overall process is done to ensure that we are not overfitting our data, which would consequently, cause overestimation of classification performance. Concluding, the column “ID” is not considered in our models because it does not constitute a predictor.

3. Models

In this report we are mainly going to fit supervised models to the data, but we also did one unsupervised approach.

Metrics for Confusion Matrix

Here we introduce the most important metrics to read a confusion matrix that are necessary to understand this report.

Accuracy	The ratio of correctly classified records (of both classes) to the total number of records.
Precision	The ratio of correctly classified malign cancers to the total number of records whose diagnosis is malign.
Sensitivity	Is the true malign rate, meaning it is the percentage of those records whose resulting diagnosis is malign and are predicted/classified as such.
Specificity	Describes the proportions of records that carry a benign cancer and are correctly classified as such. Or the ability of the classifier of “ruling out” records of the class not of interest.

3.1 Supervised Learning

In supervised learning data with class labels is used to fit models which try to predict the classes based on the other variables. Thereby the best fit should always predict the right class if possible. This is not possible most of the time because there are no clear borders in the data. Thereby the best models are those who predict the class right most of the time.

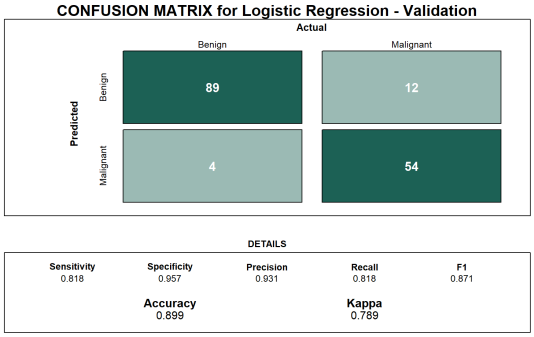
3.1.1 Logistic Regression

The first model we evaluated was logistic regression. We must first check two assumptions, the first being that the dependent variable is categorical in nature and that the independent variables do not contain multicollinearity. Thus, we will be able to use the binomial type of our logistic regression, because we have two classes, benign or malignant. When trying a basic binomial logistic regression, we can observe a convergence error[8], most often related to too many independent variables inserted in the model. The model had more than 31 variables, which partly explains the problem and also the risk of multicollinearity. We therefore conducted a penalization of the model using a logistic lasso regression model to progressively eliminate the non-significant variables until we obtained a convergent model, then we checked the multicollinearity with the VIF[9] (variance inflation factor) method. We were able to observe a very high correlation between several variables as indicated by our correlation plot and with several iterations of the VIF, we progressively eliminated the problematic variables using the VIF > 5 as a threshold.

At the 5th iteration, we are left with 8 independent variables, not suffering from multicollinearity. Using the binomial logistic regression model with only the variables surviving the VIF test, we have the following model:

```
glm(formula = diagnosis ~ area_mean + smoothness_mean + symmetry_mean +
    smoothness_se + `concave points_se` + symmetry_se + `concave points_worst` +
    fractal_dimension_worst, family = binomial(link = "logit"),
    data = Training_Logistic)
```

Keeping this model for our predictions on the validation partition, we obtain the following confusion matrix:



[8] Stack Overflow. (n.d.). *r - Why am I getting 'algorithm did not converge' and 'fitted prob numerically 0 or 1' warnings with glm?* [online] Available at: https://stackoverflow.com/questions/8596160/why-am-i-getting-algorithm-did-not-converge-and-fitted-prob-numerically-0-or .

[9] Zach (2019). *How to Calculate Variance Inflation Factor (VIF) in R*. [online] Statology. Available at: https://www.statology.org/variance-inflation-factor-r/ .

3.1.2 Classification Trees

Full Tree

Classification Trees are a data driven method, with which records are recursively partitioned into mutually exclusive subgroups. It does not require any data transformation, given that no metrics do not play a part in the resulting nodes. The process picks a predictor variable and a value said variable takes, to then split the data in 2 halves, the half whose records show values below and the other with values above. This process is repeated with the goal of reaching homogeneous (in outcome variable class) groups of observations in the end notes.

The “split points” are selected by how much they reduce heterogeneity (in classification) in the resulting node. So, the model goes through variables by importance and then picks the split that will result in the “purest” group achievable.

We run the process and plot the tree over the training set, first with a very low Complexity Parameter to allow for the tree to “grow” to its maximum and

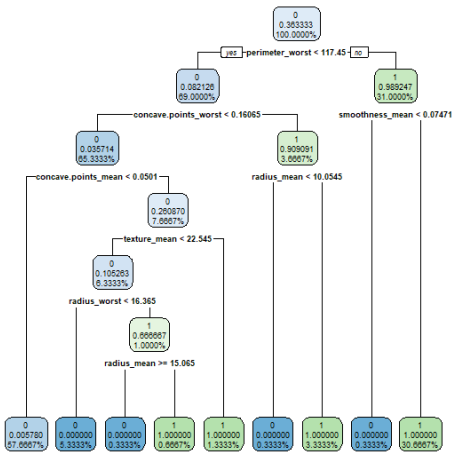


fig.3.ii.ai

overfit the data. The resulting tree is quite small in itself, the resulting end nodes are only 9 in number. We observe the CART property[10] of end nodes being 1 more in number than decision nodes.

It is not reported here (presented in the html file), but we can observe that our first split corresponds to the most important variable classification wise (whether the record is classified as malign or benign), that being “perimeter_worst” . The end notes are made up of 5 bins of “0 records” and 4 for malign cases, which reflects the proportions of benign and malign diagnoses.

[10] Page 124, Wiley.com. (n.d.). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R | Wiley*. [online] Available at: https://www.wiley.com/en-us/Data+Mining+for+Business+Analytics%3A+Concepts%2C+Techniques%2C+and+Applications+in+R-p-9781118879337 [Accessed 13 Dec. 2022].

Best Pruned Tree

As much as our fully grown tree appears to be of moderate size, we need to consider that, out of our 30 plus variables, many describe different characteristics of the same measured feature.

This also carries the possibility of the tree overfitting the data. The pruning process in short, consists of getting rid of the "weakest" branches, those are those branches originating from splits that do not reduce the error rate significantly.

These branches are more likely only modeling over noise in the data, so we proceed to resize our tree into a "Best Pruned Tree".

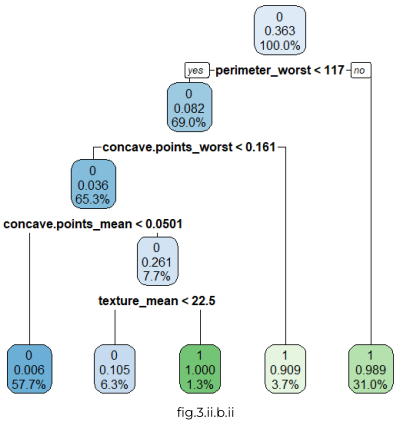


fig.3.ii.b.ii

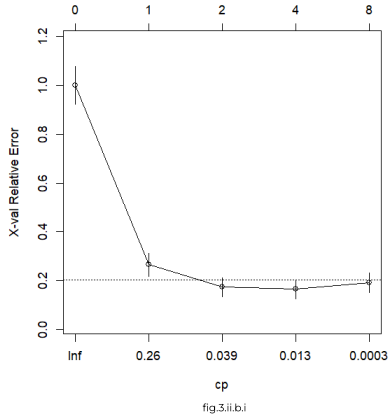
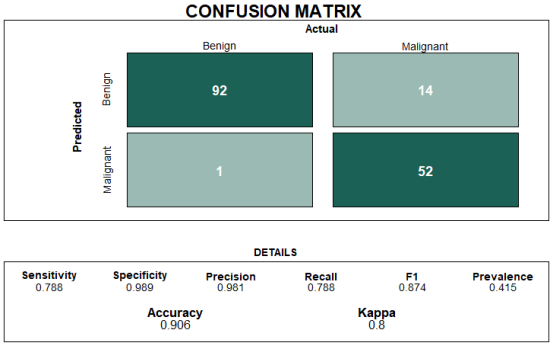


fig.3.ii.b.i

This is done by selecting the tree with the number of decision nodes achieving the lowest cross-validation error. In our case (fig.3.ii.b.i), this minimum x-error is attained with 4 decision nodes, and a complexity parameter of around 0.013.The resulting tree (fig.3.ii.b.ii), is cut to 5 terminal bins, 3 of which collecting malign results. This is good because after all we want to be able to rightly categorize the tumorous cases.

Performance of best Pruned Tree

To evaluate the performance of the Best-Pruned Tree we predict diagnosis on our validation set. Our goal is to accurately predict those cancer malign in nature. With the predicted classes, a confusion matrix is drawn.

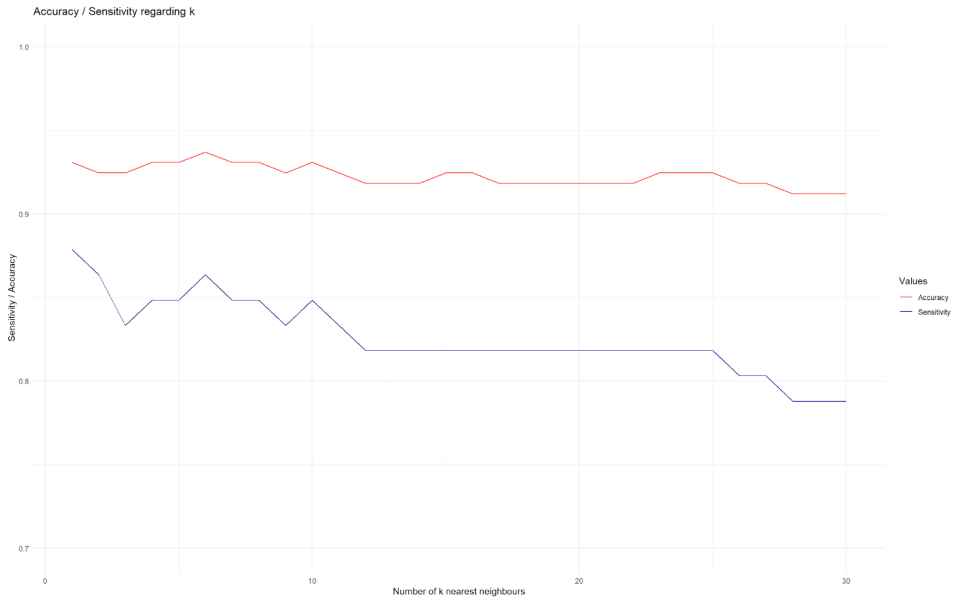


The sensitivity rate is lower than specificity, so the model is not as successful in predicting the cases we are interested in as the benign cases, but overall, the accuracy of the model is 0.90 which is not awful, but certainly we are looking for models that perform better than that.

3.1.3 K-Nearest Neighbors

In K-Nearest Neighbor (KNN) classification data is split into classes depending on the distance to *k* neighbors. It is a non-parametric method which is easy to use, and the data doesn't need to fulfill any distributional assumptions. As algorithms with distance metrics are very sensitive to different scales of variables the data needs to be centered and scaled before being used such that the mean of the data is equal to zero and the variance equal to one.

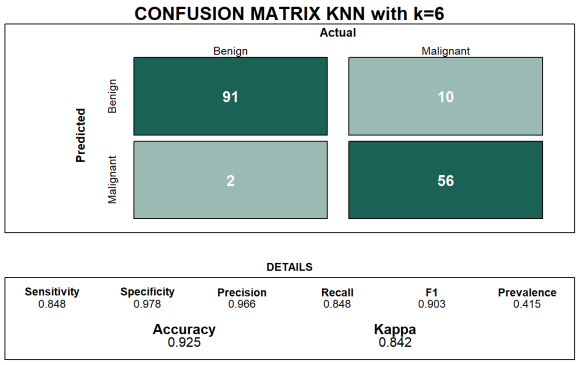
The only problem with KNN is that a *k* needs to be selected. This will determine how many neighbors will be taken into account when trying to classify new data. To find the best *k* for the data we chose an iterative approach where we fitted several models to the data with k ranging from 1 to 30. Then the best can be chosen according to different metrics. In the case of classifying cancer, the metrics that we look at were accuracy and sensitivity. Down below a plot over several k's showing the sensitivity and accuracy of the model.



As it can be seen the accuracy is highest with k equal to 6 but for the sensitivity it is highest when *k* is equal to 1. Although we want the highest sensitivity possible for classifying cancer the best model was determined as the one with *k* equal to 6 as it has the highest accuracy and second highest sensitivity.

Additionally, the model with only one nearest neighbor taken into consideration could be overfitting which should definitely not happen. This leads to the following model being fit onto the training data.

To look at the validity of the model the confusion matrix is shown down below for the KNN with 6 nearest neighbors taken into consideration for classifying data.



With the validation data provided 10 cancers would not get classified as being malignant while actually being malignant. As there are in total 66 malignant cancers in the validation data this leads to a sensitivity of 0.848. Nevertheless, the model predicts the classes very well for the benign class which leads to an overall accuracy of 0.925.

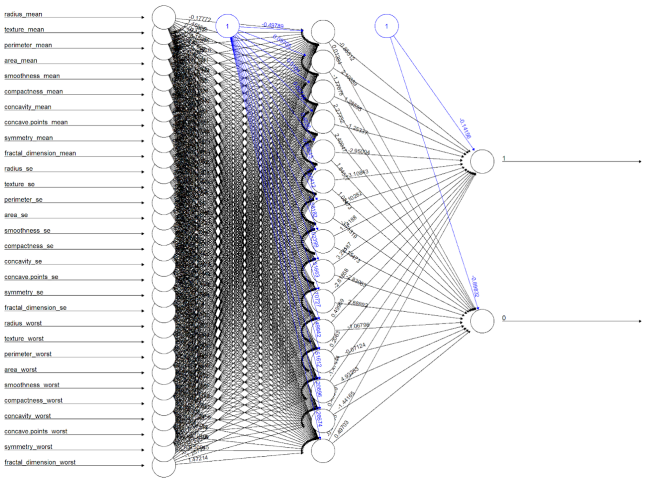
3.1.4 Neural Network

We have also tried several neural network models to predict the state of a tumor in the breast, there is no systematic process when selecting the number of hidden layers or nodes, however different sources seem to indicate a common way forward. We first established a basic model with 3 nodes and 1 hidden layer as a reference for the further process. Then, taking reference to *Introduction to Neural Networks for Java, Second Edition*[11] by Jeff Heaton, we were able to set up two other models trying to improve the predictions on the validation partition. The second model is composed of 15 nodes and 1 hidden layer and the third of 15 nodes and 2 hidden layers. We could observe that the second model obtains the same sensitivity, but the second model seems to be more accurate than the third one, and both better than the basic 3 Nodes Model.

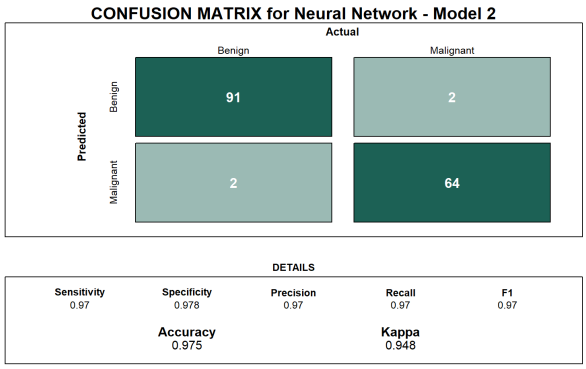
Thus, the following model allowed us to have the best predictions on our validation:

[11] gk_ (2017). *From 'Introduction to Neural Networks for Java, Second Edition'*. [online] Medium. Available at: https://medium.com/@gk_/from-introduction-to-neural-networks-for-java-second-edition-eb9a833d568c [Accessed 17 Dec. 2022].

Neural Network Model with 1 Hidden Layers and 15 Nodes



The sensitivity and accuracy of this model seems really excellent, we could also see that the rate of true negative or true positive seem almost symmetrical, an advantage if the model were also to be used for benign tumors.



3.1.5 Discriminant Analysis

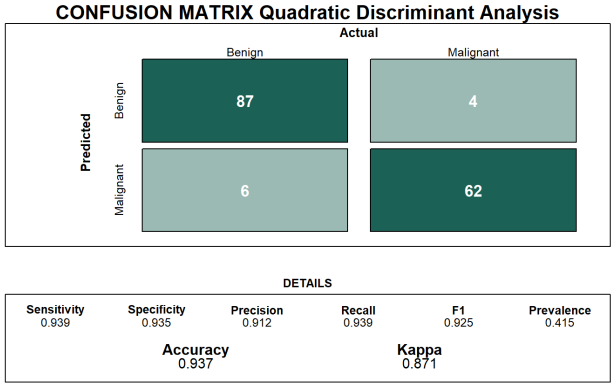
When training a discriminant analysis model, axes are laid through the scatterplots of each predictor pair which separates the classes best, and the means of each subgroup are calculated. Thereby two criteria need to be considered. The split will be where the subgroups have the biggest difference in mean and where the variance is the biggest. To predict new data, one simply goes over all predictors and chooses the mean of the subgroup which is closest to the observation. Then all the means per subgroup are added together and the one subgroup with the highest value in the end is the winning class.

While these are the basic ideas of discriminant analysis there are a lot of different methods such as linear, quadratic, flexible, mixture and regularized discriminant analysis which either loosen or tighten certain assumptions. But as all of them calculate means and distances to the means data again needs to be centered and scaled before fitting a model.

Generally speaking, discriminant analysis will not perform well when there is no separation in the data. If this would be the case logistic regression will

outperform the discriminant analysis. As we have seen in the histograms this is not the case for the data set.
After testing out several discriminant analysis methods we have chosen Quadrat Discriminant Analysis (QDA) as the best performing method. In QDA it is assumed that data comes from a normal distribution but that the covariance per class is different. Down below the model that was fitted is shown. In this model all available predictors were used as it gave back the best results.

To compare the QDA with other methods used in this report we show the confusion matrix here.



It is observable that this model is really good at detecting the malignant class with a sensitivity of 0.939. Additionally, the specificity is also quite high which results in a good overall accuracy.

3.1.6 Ensemble Methods

Bagging

For the ensemble methods regarding Classification Trees, considerations and comparisons for their performances and the Best Pruned Tree will be drawn. As we feel that it is coherent to the final selection of the best models for this subset of classification algorithms to be tackled together.

By Bagging we mean Bootstrap Aggregation. Multiple samples are created through bootstrapping from the training set, and multiple trees are created from these many samples. Following, the results are aggregated, each record will be classified through a majority vote mechanism that accounts for all trees produced.

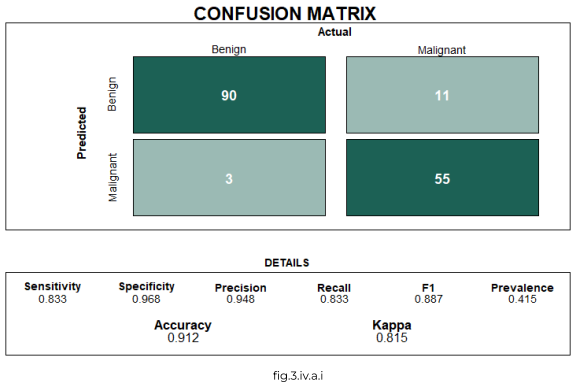


fig.3.iv.a.i

This method compared to regular classification trees reduces the instability of the final categorization, because we have run many trees from as many random samples. Of note, is the peculiarity that the generated trees are correlated, this is due to the fact that predictors importance matters, so all trees

drawn are likely to have the same first split. We expect the bagging model to perform better than our Best Pruned Tree. We draw our confusion matrix. The accuracy metric has increased slightly, compared to the previous model (fig.3.ii.c.i). The sensitivity amounts to 0.833 which is an improved result from the original 0.788 .

Specificity has actually decreased, which again is not necessarily good or bad, but it is something we want to see from the ensemble method, given our interest in detecting the cases belonging to the positive class 1.

Boosting

It is an iterative proceeding that learns from previous iterations. A sample is drawn, and a tree is fit over it, the following sample will be selected such that the misclassified records hold more weight and therefore have higher probability of being drawn. A tree is fit over this new sample and so on, executing sequential learning.

Weighting the misclassified records more, we expect for the emblems methods to first perform better than a regular tree, but also to do a better job than the bagging tree at predicting malign cases of breast cancer.

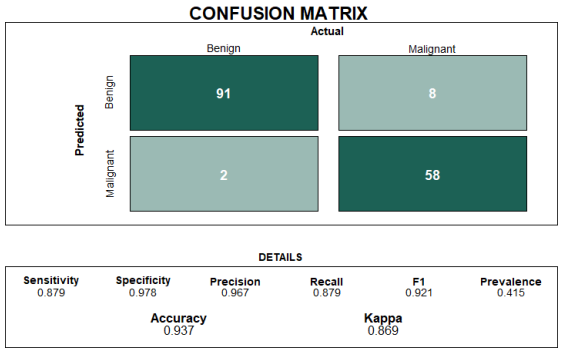


fig.3.iv.b.i

The Boosted Tree achieves the highest classification Accuracy value so far. The 0.879 Sensitivity metric entails that this process does provide better tools for the categorization of tumorous cases, than bagging.

Must be noted that Specificity has incremented as well by a small amount. Still, the metric remains below the Best Pruned Tree's Specificity.

Random Forests

Practically speaking the process is very similar to the one implemented by bagging. So many random samples are drawn, and many trees are fitted to each of these samples, to then vote the final classification.

The difference in this case is that this supervised machine learning algorithm will only consider a few random variables to split nodes, in the case of bagging we had many trees

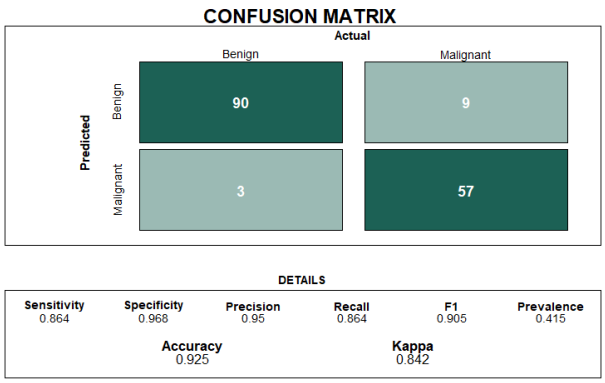


fig.3.iv.c.i

using the same most important variable every run. The misclassification rate amounts to 0.075, which in our situation represents a middle ground between Bagging and Boosting Accuracies in distinction of cases. The sensitivity measure tells us that Random Forests outperforms the Best Pruned Tree, also introducing robustness to the model.

But overall, it does not surpass Boosting which seems to do the better job at classifying those malign cancers as such.

ROC curves for all 3 ensemble models

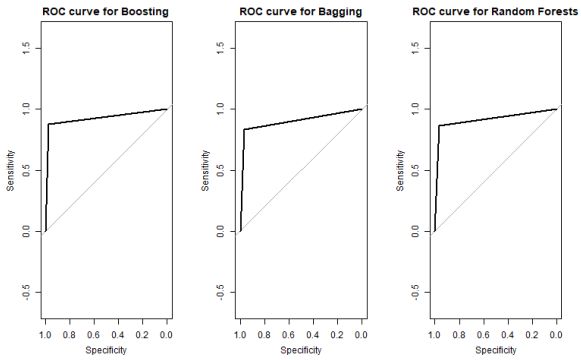
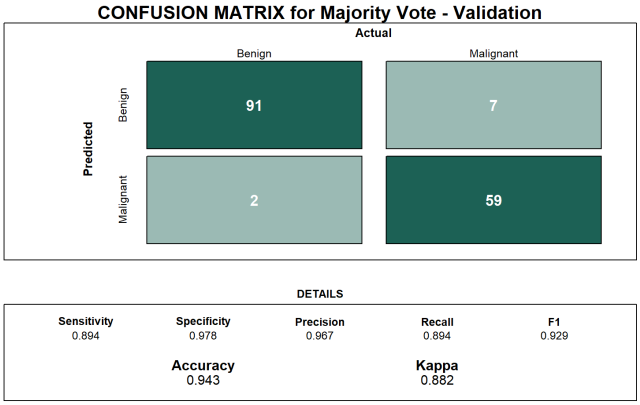


fig.3.iv.c.ii

The plots describe how Sensitivity and Specificity vary as the cutoff moves from 1 to 0. We can then visually compare the different models' ability to rightfully classify malign cases and benign cases. Boosting is the process delivering the more accurate classification of our class of interest.

3.1.7 Majority of Vote

Another way to do an ensemble method which is not based on trees is to fit multiple models to the data and predict the classes based on the most picked class for each observation. In this case we took the voting of the best performing models from before which were the Logistic Regression, KNN, Boosted Trees, Discriminant Analysis and Neural Network. When predicting on validation data this confusion matrix resulted.



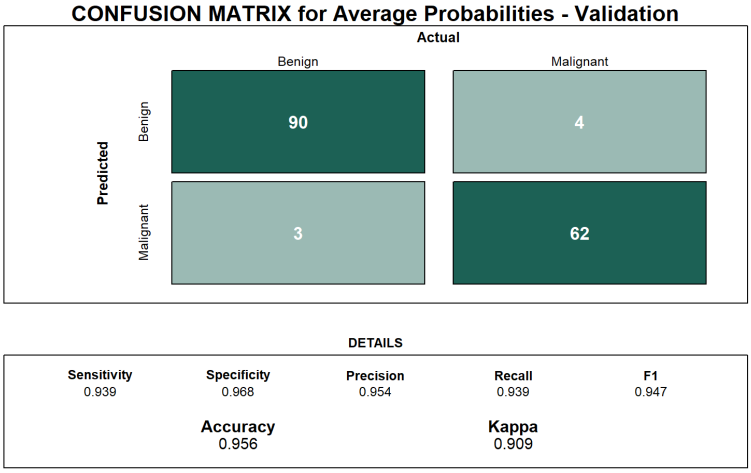
We see from the output that this model is not as good in the case of sensitivity but very good in specificity. But due to the low sensitivity the overall accuracy is drawn down which makes the majority voting not a good model.

3.1.8 Average of Probabilities

Similarly to the majority voting the five best performing models are used to make predictions. But in this instance not the decision is taken but the probability for voting for malignant class. Out of the average of the probabilities one vote is then made. This method will perform differently as the majority vote

as in cases where there are close decisions the prediction will not be as much influenced as in majority voting.

Differently than in majority voting the sensitivity of the average probability model is very good and in combination with a high enough specificity the model has a high overall accuracy.



3.2. Best Models

In this part we will show the sensitivity, specificity and accuracy for all of the best models of each part. We will analyze the performance of all best models on the validation data and analyze the performance for the three best ones on test data.

For clarity, we briefly recount which models these “best” are. In no particular order for the final top three selection, we are considering the 7 best models previously run: Boosted Trees Model, The Neural Network, Discriminant Analysis, Logistic Lasso Regression, K-Nearest Neighbors, Majority Vote and Average Probabilities.

3.2.1 Best 3 Models on Validation

As we fit the data to training we need to check whether the model is good in general or if it only performs good on the training data. Therefore, we use the mode fitted on training data to predict the classes on validation. Down below we display the table of the sensitivity, specificity and accuracy of all the best models of each part.

7 Models on Validation - Order by Sensitivity

	Sensitivity	Specificity	Accuracy
Best Neural Network	0.9697	0.9785	0.9748
Best DA	0.9394	0.9355	0.9371
Average Probabilities	0.9394	0.9677	0.956
Majority Vote	0.8939	0.9785	0.9434
Boosted Tree Model	0.8788	0.9785	0.9371
Best KNN	0.8485	0.9785	0.9245
Best Logistic Lasso Regression	0.8182	0.957	0.8994

As we can observe from the table the best models on validation data are the Neural Network, the Discriminant analysis and the average probability model. Thereby they will get verified on testing data as well.

3.2.2 Best 3 Models On Test

To conclude our analysis, what we proceed to do is select our top three models' performance wise, and appraise their classification work on the test set (which amounts to 20% of the original data-set).

This step is taken to simulate the usage of one of our models on unseen data, therefore testing the stability of it. Implying, no further optimization of our models will follow from this. We are looking to draw considerations on whether the proposed classifiers overfit the data or not, overall, we do expect worse performance, but we also hope that they do not result too far off.

As previously reported, our top three includes (in this order) Neural Network, Discriminant Analysis and finally the ensemble yielded from averaging the classification probabilities from the selection of 5 we compare in the section above. For the former 2 models, we simply test them on the smallest hold out set. For the latter case, the process requires multiple stages, in fact to test the Averaging Ensemble we need to first build predictions on all models for the test set and average them once again.

7 Models on Test - Order by Sensitivity

	Sensitivity	Specificity	Accuracy
Best Neural Network	0.973	0.9863	0.9818
Average Probabilities	0.973	0.9726	0.9727
Best DA test	0.9459	0.9726	0.9636
Majority Vote	0.9459	0.9726	0.9636
Boosted Tree Model on Test	0.9189	0.9863	0.9636
Best Logistic Lasso Regression	0.8919	0.9589	0.9364
Best KNN test	0.8919	1	0.9636

The result table is ordered by sensitivity, it is evident that on the Test set the Average Ensemble has gained in this metric relative to Discriminant Analysis. The outcome of this testing is quite surprising in fact, because all models display higher Accuracy measures than they did on the Validation set.

This of course, can be considered a coincidence that can be attributed to sampling, after all the technique is not faultless. But it is also encouraging because it means that our models do not dramatically drop in sensitivity and other metrics once they have to perform on new data, they remain adequately stable.

3.3. Unsupervised Learning

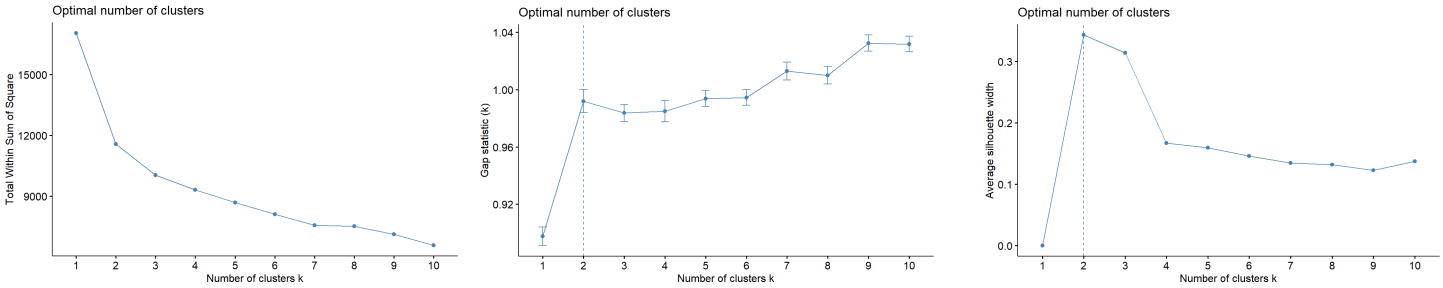
The purpose of this part using unsupervised models is to explore our dataset and to be able to glimpse added information, using some medical knowledge to open up a wider search and understand the limitations of our data.

3.3.1 K-Means Clustering

The idea behind the k-means clustering method[12] is to separate our data (or points) into a number of separable K groups, each of which is assumed to remain homogeneous and compact (within their respective cluster). By calculating the Euclidean distance between each observation, we can separate them into an optimal number of clusters, this number being calculated from 3 different methods. Thus, we could for example find that for the whole population of breast tumors, 2 clusters seems to be the optimal number.

The methods applied are the elbow method, the silhouette method and the gap statistics[13]:

Three Methods for Optimal Number of Cluster - For Model 1

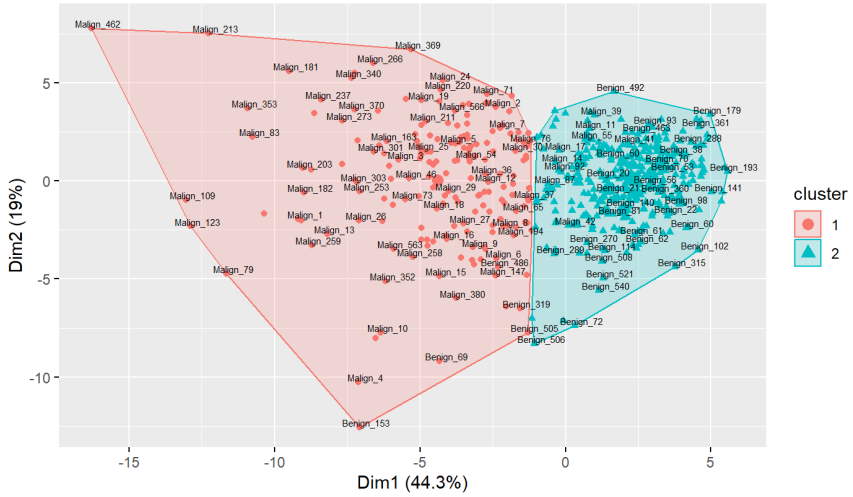


[12] Formation Data Science | DataScientest.com. (2020). *K-means : Focus sur cet algorithme de Clustering & Machine Learning*. [online] Available at: https://datascientest.com/algorithme-des-k-means [Accessed 17 Dec. 2022].

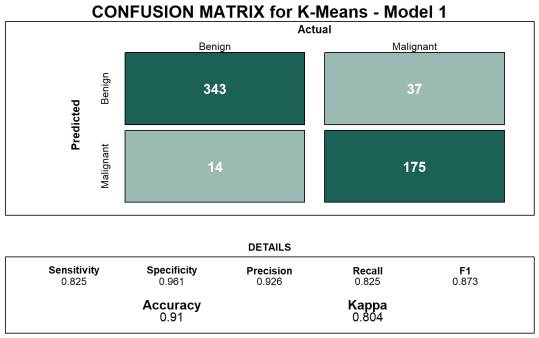
[13] Datanovia. (n.d.). *Determining The Optimal Number Of Clusters: 3 Must Know Methods*. [online] Available at: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/ [Accessed 17 Dec. 2022].

Thus, we have a first cluster model, which once visualized, shows to have separated automatically and without indication, the malignant and benign tumors.

Cluster Model 1 - Whole Population of Tumors



We can also extract the members of each cluster to make a performance test in the form of a confusion matrix:



We can see that this type of model has been quite successful in classifying benign tumors from malignant tumors, with an accuracy of 0.9, however compared to supervised models, the sensitivity is rather low and does not really allow this model to be a good predictive model for our needs.

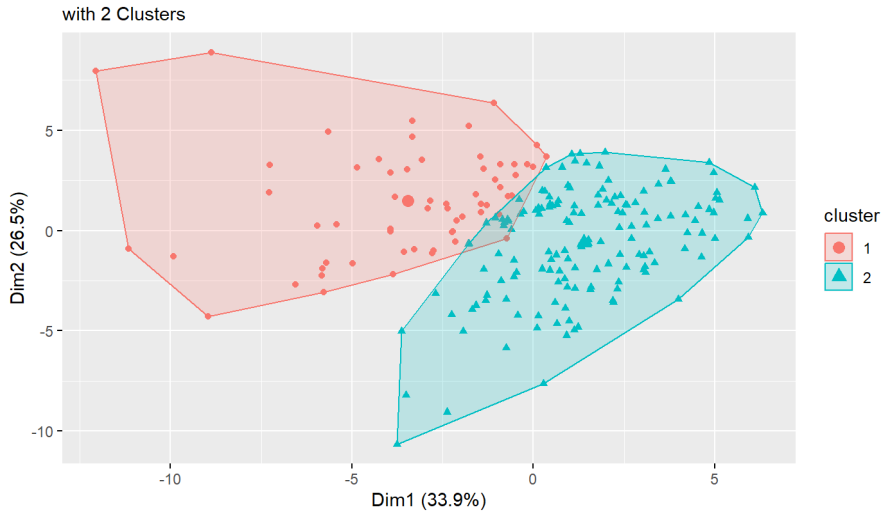
Furthermore, we could make a cluster model but only for malignant tumors. First with an optimal number of 3 and then 2 depending on which methods we choose. Then by looking at the centroids of each cluster, we could notice that there are indeed quite different averages for each variable.

When coupling information on the stages of malignant tumors using the staging system by the *American Cancer Society*[14] as well as the *ASCO*

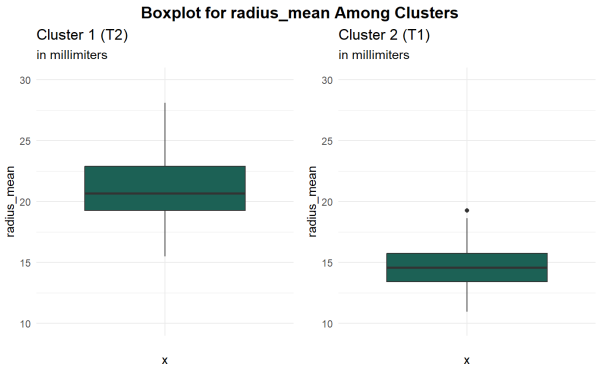
[14] American Cancer Society (2018). *Breast Cancer Stages*. [online] Cancer.org. Available at: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html .

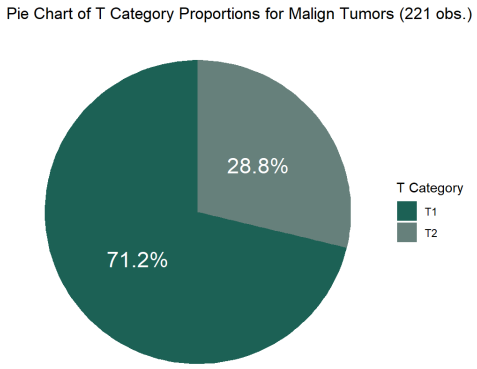
organization[15], we can observe that the global size of the tumor can determine the possible states and therefore medical measures to be taken for treatment or surgical operation.

Cluster Model 2 - Subset of Population: Only Malignant Tumors



It is by looking at the tumor radius variable that we can see that our two clusters can give important information about the possible states of each tumor and thus understand the limitations of our data to better categorize our malignant tumors.





With the help of the 7 Key pieces information provided by the *American Cancer society*[16], we can see that we can only provide 1 key, that of the size of the tumor denoted **T**. The transition from **T1** to **T2** is mainly characterized by the size of the tumor, from a radius smaller than 2 cm in size to a measure equal or larger than. However, size can also be defined by area or other metrics thus **T1** and **T2** are hypothetical here and only based on how the clustering occurred, thus some **T2** members still have smaller radius than 2cm but may have other

[15] Cancer.net. (2019). *Breast Cancer - Stages*. [online] Available at: https://www.cancer.net/cancer-types/breast-cancer/stages .
[16] American Cancer Society (2018). *Breast Cancer Stages*. [online] Cancer.org. Available at: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html .

metrics showing bigger sizes. We can notice the difference with the boxplots of our clusters on the radius variable. There is still a lot of other information required, which limits the possibilities of identification.

Here are the only possible assumptions[17] we can make with the information we have:

Cluster 2 with **T1** could potentially lead to	***Stage IA**: The tumor is small, invasive, and has not spread to the lymph nodes (**T1**, N0, M0).* ***Stage IB**: Cancer has spread to the lymph nodes and the cancer in the lymph node is larger than 0.2 mm but less than 2 mm in size. There is either no evidence of a tumor in the breast or the tumor in the breast is 20 mm or smaller (T0 or **T1**, N1mi, M0).* ***Stage IIIC**: A tumor of any size that has spread to 10 or more axillary lymph nodes, the internal mammary lymph nodes, and/or the lymph nodes under the collarbone. It has not spread to other parts of the body (any T, N3, M0).* ***Stage IV** (metastatic): The tumor can be any size and has spread to other organs, such as the bones, lungs, brain, liver, distant lymph nodes, or chest wall (any T, any N, M1). Metastatic cancer found when the cancer is first diagnosed occurs about 6% of the time. This may be called de novo metastatic breast cancer. Most commonly, metastatic breast cancer is found after a previous diagnosis of early stage breast cancer.*
Cluster 1 with **T2** could potentially lead to	***Stage IIA**: Any 1 of these conditions: The tumor is larger than 20 mm but not larger than 50 mm and has not spread to the axillary lymph nodes (**T2**, N0, M0).* ***Stage IIB**: The tumor is larger than 20 mm but not larger than 50 mm and has spread to 1 to 3 axillary lymph nodes (**T2**, N1, M0).* ***Stage IIIA**: The tumor of any size has spread to 4 to 9 axillary lymph nodes or to internal mammary lymph nodes. It has not spread to other parts of the body (T0, T1, **T2**, or T3; N2; M0).* ***Stage IIIC*** ***Stage IV** (metastatic).*

[17]Cancer.net. (2019). *Breast Cancer - Stages*. [online] Available at: https://www.cancer.net/cancer-types/breast-cancer/stages .

4. Conclusion

We were able to run all the models we had planned and rank the best models in terms of predictive ability, focusing on the sensitivity of the model and its overall accuracy in predicting malignancies. A different ranking can be made depending on the prediction goal and the importance that the medical profession puts on incorrectly classified benign tumors, however in our case, classifying malignant tumors more correctly allows us to invest more money and time in confirming the diagnosis and thus to avoid making this kind of effort if the probability of malignant tumors is low. As in any model, the larger the data size, the greater the risk of redundant information or multicollinearity problems may occur.

Moreover, the paper produced from this dataset uses only 3 variables at a time in this method: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34][18]. Most of our models have been able to use the majority of the measurements made on breast tumor scans, however, in the context of generalization and external application (in the case of limited time and medical devices), it would probably be more interesting to use a dataset containing the measurements that are important today to determine the nature of a tumor and to be able to predict more precisely the state of malignant tumors and to be able to offer the medical profession pre-diagnostic advice on the probability of surgical or chemical operations according to the staging system.

We can also conclude that the main purpose of this report was to provide predictive tools rather than explanatory models, we could also consider opening up this kind of research to be able to collect data and possible ranges of adequate values in the future to conduct better performance tests and thus better predict potential tumors, not only in the breast but also by making the best use of the scans and analyses of the tumors, as the structure of the nodes in the tumors and the directions of spread to other organs are also very important in identifying the cancer and its stage of development.

[18] Uci.edu. (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) .

5. References

- Uci.edu. (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) .
- Wiley.com. (n.d.). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R | Wiley*. [online] Available at: https://www.wiley.com/en-us/Data+Mining+for+Business+Analytics%3A+Concepts%2C+Techniques%2C+and+Applications+in+R-p-9781118879337 [Accessed 13 Dec. 2022].
- www.javatpoint.com. (n.d.). *Logistic Regression in Machine Learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/logistic-regression-in-machine-learning.
- Stack Overflow. (n.d.). *r - Why am I getting 'algorithm did not converge' and 'fitted prob numerically 0 or 1' warnings with glm?* [online] Available at: https://stackoverflow.com/questions/8596160/why-am-i-getting-algorithm-did-not-converge-and-fitted-prob-numerically-0-or [Accessed 13 Dec. 2022].
- www.sthda.com. (n.d.). *Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net - Articles - STHDA*. [online] Available at: http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net [Accessed 13 Dec. 2022].
- Zach (2020). *Lasso Regression in R (Step-by-Step)*. [online] Statology. Available at: https://www.statology.org/lasso-regression-in-r/ .
- www.digitalocean.com. (n.d.). *Plotting ROC curve in R Programming | DigitalOcean*. [online] Available at: https://www.digitalocean.com/community/tutorials/plot-roc-curve-r-programming [Accessed 13 Dec. 2022].
- DataTechNotes (n.d.). *How to create a ROC curve in R*. [online] Available at: https://www.datatechnotes.com/2019/03/how-to-create-roc-curve-in-r.html [Accessed 13 Dec. 2022].
- Zach (2019). *How to Calculate Variance Inflation Factor (VIF) in R*. [online] Statology. Available at: https://www.statology.org/variance-inflation-factor-r [Accessed 13 Dec. 2022].
- www.datacamp.com. (n.d.). *ANN (Artificial Neural Network) Models in R: Code & Examples on How to Build Your NN*. [online] Available at: https://www.datacamp.com/tutorial/neural-network-models-r .
- in (2010). *How to choose the number of hidden layers and nodes in a feedforward neural network?* [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw .
- Heaton, J. (2008). *Introduction to Neural Networks with Java*. [online] *Google Books*. Heaton Research, Inc. Available at: https://books.google.it/books?id=Swlcw7M4uD8C&lpg=PA158&dq=Introduction%20to%20Neural%20Networks%20for%20Java%2C%20Second%20Edition%20The%20Number%20of%20Hidden%20Layers&hl=it&pg=PA158#v=onepage&q=Introduction%20to%20 Neural%20Networks%20for%20Java [Accessed 13 Dec. 2022].
- finnstats (2021). *Cluster Analysis in R | R-bloggers*. [online] Available at: https://www.r-bloggers.com/2021/04/cluster-analysis-in-r .
- Cross Validated. (n.d.). *machine learning - Do we need to set training set and testing set for clustering?* [online] Available at: https://stats.stackexchange.com/questions/268934/do-we-need-to-set-training-set-and-testing-set-for-clustering [Accessed 13 Dec. 2022].
- Imad Dabbura (2018). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. [online] Medium. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a .
- www.cancer.org. (2021). *Types of Breast Cancer | Different Breast Cancer Types*. [online] Available at: https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html .
- American Cancer Society (2018). *Breast Cancer Stages*. [online] Cancer.org. Available at: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html .
- Cancer.net. (2019). *Breast Cancer - Stages*. [online] Available at: https://www.cancer.net/cancer-types/breast-cancer/stages .
- Chelliah, I. (2021). *Bagging Decision Trees — Clearly Explained*. [online] Medium. Available at: https://towardsdatascience.com/bagging-decision-trees-clearly-explained-57d4d19ed2d3 .