Multivariate Analysis Project
Supervised and Unsupervised Learning on Breast Cancer Wisconsin dataset

André Godinho
Data Science MSc.
Instituto Superior Técnico
Universidade de Lisboa, Portugal
`andre.vaz.godinho@tecnico.ulisboa.pt`

Francisco Caldas
Data Science MSc.
Instituto Superior Técnico
Universidade de Lisboa, Portugal
`francisco.caldas@tecnico.ulisboa.pt`

June 27, 2020

Abstract

The work developed in this report intends to analyze in detail the Breast Cancer Wisconsin dataset by applying a thorough preliminary analysis with principal component analysis, feature selection and data visualization.

Once the analysis is completed, the problem is solved by applying different supervised learning methods like Random Forests, Support Vector Machine and K Nearest Neighbor. In order to compare the resulsts, several metrics are used such as confusion matrix, accuracy, f1-score, recall and precision.

The problem in hands is also addressed with Unsupervised methods like Agglomerative clustering and K-means algorithm. Regarding the results of this analysis, metrics such as silhouette score and davies bould v measure, purity and adjusted rand are compared. We will use certain techniques to interpret the results of these clustering methods by applying interpretable methods like PCA, dandogram, silhouette plot and a Decision Tree.

Finally, comparisons between the results of the supervised and unsupervised techniques will be compared, by discussing their advantages for this specific problem and generally.

1 Description of the problem under study

The problem in hands has to do with clinical cases of breast cancer cells. We are going to analyze the *Breast Cancer Wisconsin Dataset* [1] from [2] and analyze the impact of the following attributes in the class of the cell:

1. Clump thickness
2. Cell size uniformity
3. Cell shape uniformity
4. Marginal Adhesion
5. Single Epi Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses
10. **Class: malignant ; benign**

We recommend to view the code of our analysis by the HTML files because all the code and plots can be visualized easily.

2 Objectives

This work is divided into several parts and for each one a different objective is pretended.

First of all, regarding the preliminary analysis, it is mandatory for us to grasp completely the impact of the features in the class label (malignant or benign cell). Thus, we can have a better approache on the Supervised and Unsupervised methodologies.

Second, it is crucial to apply different Supervised methods with the conclusions taken from the preliminary analysis to create unbiased (within possibility) and accurate models to predict if a cell is malignant or benign. Since we are dealing with a Health problem, we have to take special attention to false negatives. Thus, a good performance in Recall(1) is rather important.

Lastly, a good unsupervised approach is very useful for many reasons, and different ones are applicable to different datasets. In this case, we are looking for conclusions that could help improve either the knowledge of the dataset or help give useful information for the classification problem.

3 Estimation and validation methods

3.1 Supervised methodology

For the Supervised approach, we splitted 70% of the dataset for train and 30% for test. Also, we shuffled the samples when doing this split.

The methodology for this part consists in applying supervised learning using four different approaches:

1. Applying principal component analysis (PCA) on the dataset and use a sufficient number of principal components that explained a percentage of variance above 80%.
2. Applying PCA on the dataset and use a number of principal components such that the percentage of variance explained is above 90%.
3. Apply feature selection taking into consideration the correlation among features.
4. Not applying any dimensionality reduction technique as a control test.

For each of these approach we will use three different Supervised methods: Random Forest, Support Vector Machine and K Nearest Neighbor.

For each one of these methods we will use 5-fold cross validation in the train set and for each fold, calculate the accuracy considering the train sample classes as the true label. Furthermore, we will be applying **random search** for Random Forest and K Nearest Neighbor because the number of possible combinations of the hyperparameters is very high, whereas for the SVM classifier we will use **grid search** because we only have 24 different combinations of hyperparameters. For each classifier **we will choose the one that had higher mean accuracy in the 5-folds validation sets**. Using k-fold cross helps us ensure that our model can generalize to an independent data set. Thus, it gives a good validation when applying different strategies to improve our models to guarantee that we accomplish the objective of this task: create a unbiased and accurate model.

All the hyperparameters used are explained in the appendix.

After choosing the classifier with higher mean accuracy in 5-fold cross validation for each approach, we will use it on the test set. We chose the following metrics to compare the results of the four approaches: **accuracy, f1-score, recall(1), recall(0), precision(1) and precision(0)**. As mentioned before, special attention will be taken regarding recall(1) due to the fact that predicting a cell as benign when it was malignant is more penalizing than predicting a cell as malignant when it was benign.

3.2 Unsupervised methodology

3.2.1 Number of Clusters

Our initial hypothesis to this dataset is to consider that there are two clusters, as there are two classes (malignant and benign). To test this initial hypothesis, we will use the Elbow method by analyzing the plot of the within-cluster sum-of-squares (Inertia).

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

For this method, we trained a kmeans model with N clusters (for N=[1,2..9]) and measured the Inertia for each learnt output. This plot shows us the decrease of inertia with the number of clusters and the "elbow" seems to be located at N=2, confirming our hypothesis.

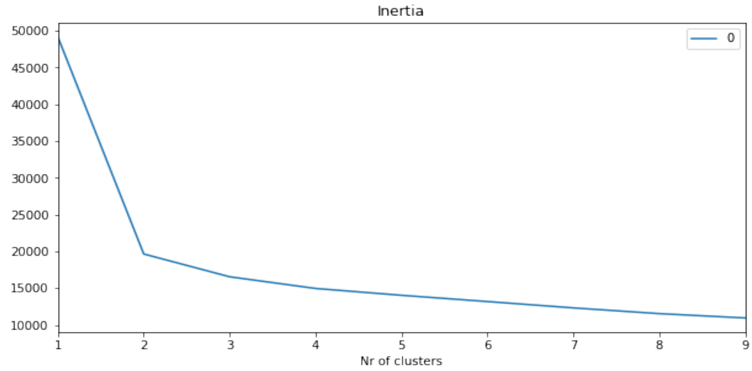


Figure 1: Inertia of a Kmeans clustering versus the possible number of clusters.

3.2.2 Exploratory Data Analysis for clustering

Our hypothesis is currently, the following: the dataset is divided into two clusters and they correspond to the two classes of the original dataset (malignant and benign). We tested the first part of our thesis previously and by plotting the first 2 and 3 pincipal component analysis, our hypothesis seems to hold.

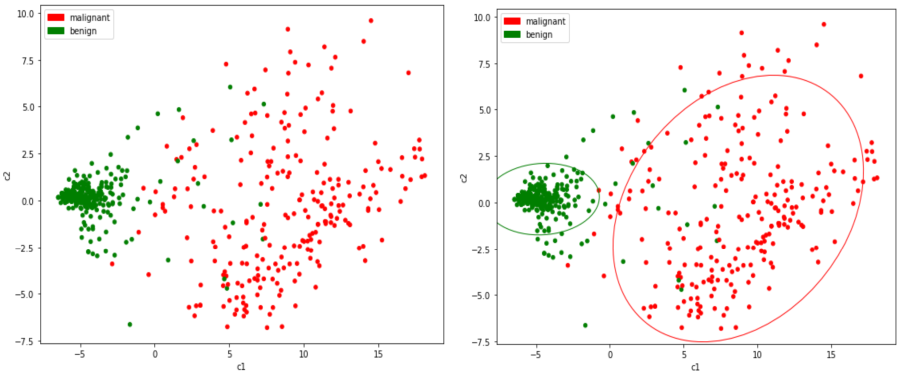


Figure 2: left: First two PCA components colored by the true classes of the dataset. Right: 95.4% interval of confidence of a 2D normal distribution calculated with the variance and mean of the PCA components of each class.

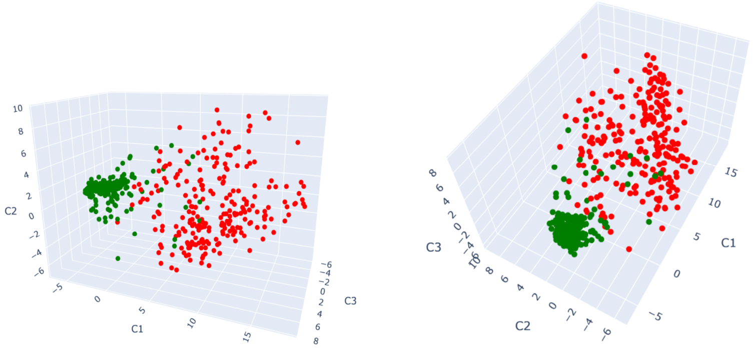


Figure 3: First three PCA components colored by the true classes of the dataset.

To explore more 3D visualizations similar to the ones presented, refer to the notebook attached.

3.2.3 Hierarchical Clusterings and Kmeans

We trained a hierarchical clustering model for different similarity measures and for different distances between clusters. For the similarity, we used: euclidean distance, l1 norm, l2 norm, manhattan distance, and cosine similarity. For the distances between clusters, we used: single linkage, complete linkage, average linkage and the ward method.

We also trained a Kmeans model in order to also test a partition method. And it converged within 4 to 5 iterations (for 10 runs), and our stopping criteria had a tolerance of 0.0001.

In order to evaluate the clusters, we used two unsupervised indices (silhouette and davies bouldin), and three supervised indices (v measure, purity and adjusted rand).

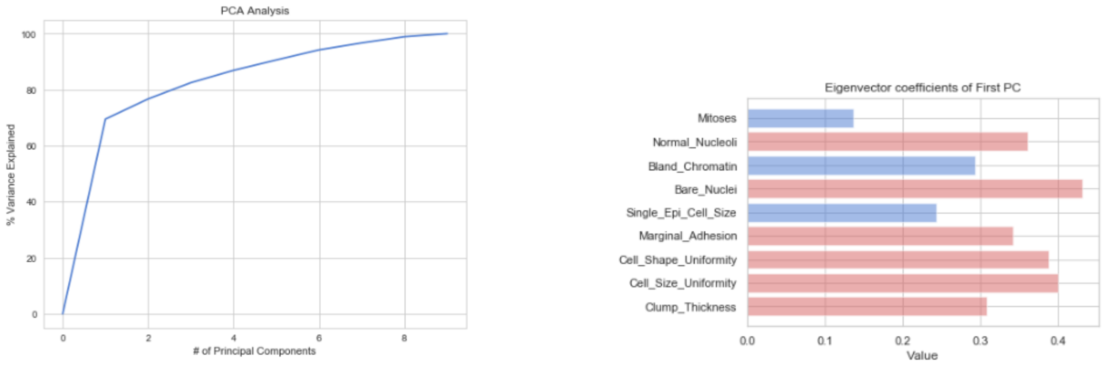
4 Findings in Preliminary Analysis

In this section we will report the findings obtained in the Preliminary Analysis. Take in mind that **all of these were obtained in the train set, we did not use the test set for data analysis to prevent data leakage**.

4.1 Principal Component Analysis

We applied PCA to the train set to estimate the percentage of explained variance for the different principal components. To achieve a percentage of variance explained above 80% we need 3 Principal Components, whereas to achieve above 90% we need 5 Principal components. These results can be observed in figure 4a.

We have also analyzed the eigenvector coefficients of the first principal component. It is a weighted sum of *Normal Nucleoli*, *Bare Nuclei*, *Marginal Adhesion*, *Cell Shape Uniformity*, *Cell Size Uniformity* and *Clump Thickness*. Samples with a high first principal component will have high values of these features. These results can be observed in 4b.



(a) Total percentage variance explained according to the Principal Components

(b) Eigenvector of the first Principal Component

Figure 4: Principal Component Analysis

4.2 Correlation among features

To observe the correlation values among features we ploted the correlation matrix of the features. The feature *Cell Size Uniformity* is highly correlated with *Cell Shape Uniformity* and with *Bland Chromatin*. **These two last features will not be used in Feature Selection approach**. These results can be observed in figure 5.

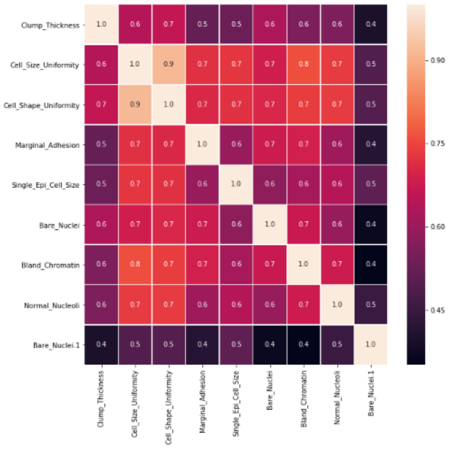


Figure 5: Correlation matrix of features in train set data

4.3 Swarn plot

With this visualization tool that represents the distribution of values of each sample according to its feature and class label. Firstly, we can observe that the minimum value for each feature is 1 and the higher value is 10. Secondly, higher values of each feature tend to correspond to malignant cells whereas lower values tend to correspond to benign cells. Also, for each feature we can clearly observe a separation boundary among classes. For example, for the feature *Cell Size Uniformity*, values above 2 are more likely to belong to a malignant cell, whereas values below 3 tend to correspond to benign cells. Thus, it is a useful feature for classification.

5 Results

5.1 Supervised learning methods

The table 1 summarizes the results obtained for each one of the four approaches defined in section 3.1. Also, the classifier method chosen was the one with **higher mean accuracy in the 5-fold crossvalidation** and then applied on the test set.

Table 1: Supervised learning test set results

Approaches	**Accuracy**	**F1-score**	**R(1)**	**R(0)**	**P(1)**	**P(0)**	**Method**
FS	**0.957**	**0.934**	**0.985**	**0.945**	**0.889**	**0.993**	Random Forest
3 PC	0.952	0.928	**0.985**	0.938	0.877	**0.993**	SVM
5 PC	0.943	0.914	**0.985**	0.924	0.853	**0.993**	Random Forest
WDR	0.952	0.926	0.969	0.945	0.887	0.986	SVM

(Note: FS - Feature Selection, WDR - Without dimensionality reduction)

The approach that achieved higher results was the **feature selection one**. Not using the referred features in 4.2. increased the performance of this approach. Furthermore, the approach with 3 principal component had better results than the approach where any dimensionality reduction techniques were applied. However, the approach with 5 principal components had the worst performance.

Regarding Recall(1), all approaches had a very high value, which is a very good result since we are analyzing a health problem. All the confusion matrices can be observed in the *jupyter notebook*.

5.2 Unsupervised learning results

The following table shows the results for the algorithm configurations that produced the best results (according to the average of all the above metrics):

	euclidean and ward	**Kmeans**	**l1 and average**	**manhattan and average**	**euclidean and average**	**l2 and average**
silhouette	0.571	**0.597**	0.590	0.590	0.590	0.590
Davies bouldin	**0.788219**	0.759	0.777	0.777	0.775	0.775
v measure	**0.800**	0.7434	0.698	0.698	0.673	0.673
purity	**0.966**	0.960	0.948	0.948	0.943	0.943
adj. rand index	**0.866**	0.844	0.802	0.802	0.781	0.781

We can clearly see that the hierarchical clustering with euclidean distance and ward method gives the best result.

The following graphs show the results of the 3 best and 3 worst configurations (applied to the PCA data just for plotting in 2D).

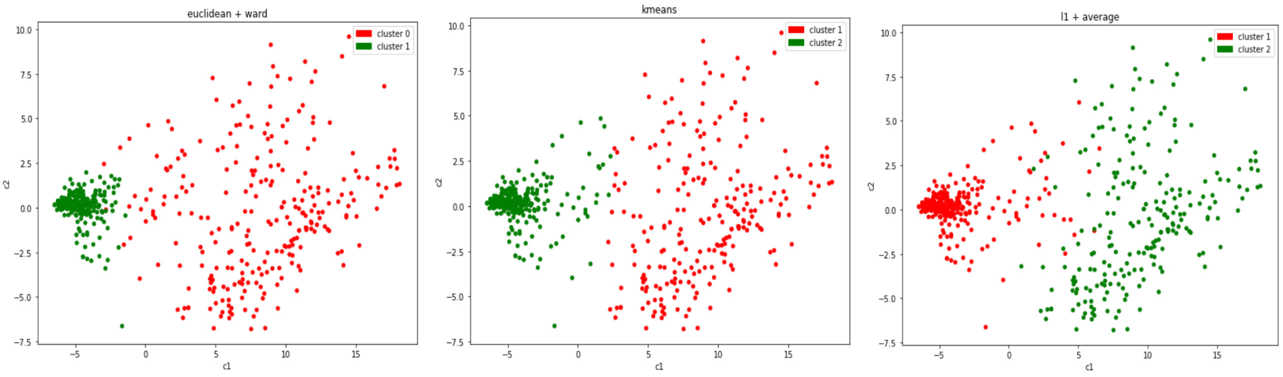


Figure 6: Best 3 results of the clustering models trained. For all the results refer to the notebook.

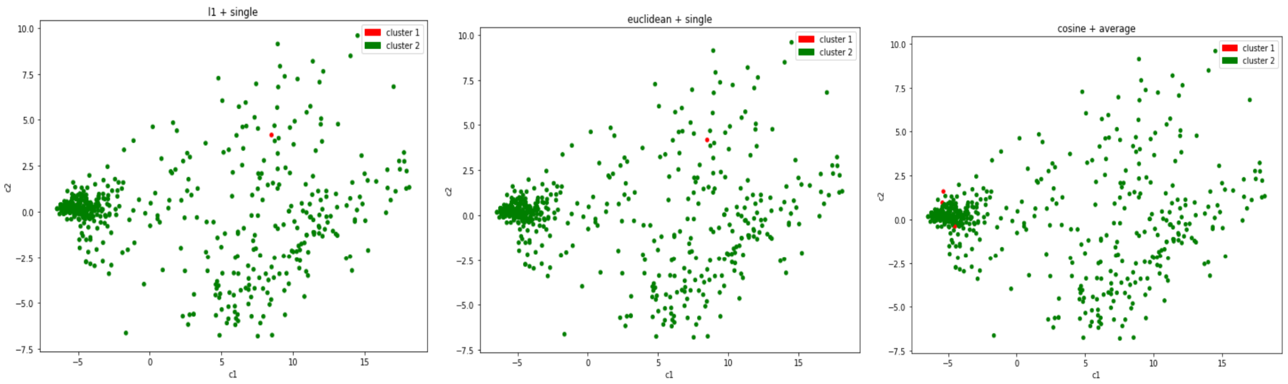


Figure 7: Bottom 3 results of the clustering models trained. For all the results refer to the notebook.

From this point, we will use the best configuration referenced above to evaluate the results. With bigger datasets, the Kmeans model could be a better choice given that it is more computationally efficient, but as our dataset is small, we will use the one with best average metrics.

5.2.1 Analyzing the Clusters

The following dendogram of the Ward + euclidean configuration also strengthens the initial hypothesis that 2 clusters is a number that fits the data well.

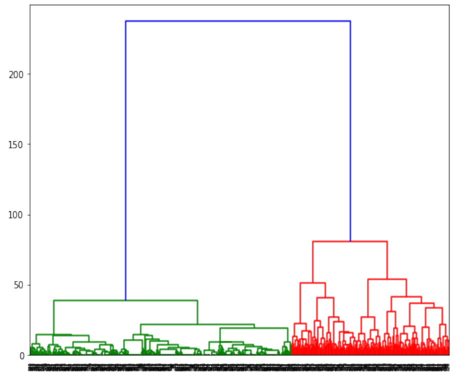


Figure 8: Correlation matrix of features in train set data

We also applied a decision tree to the dataset with the labels being the ones given by the clusters in order to analyze the most important features separating the 2.

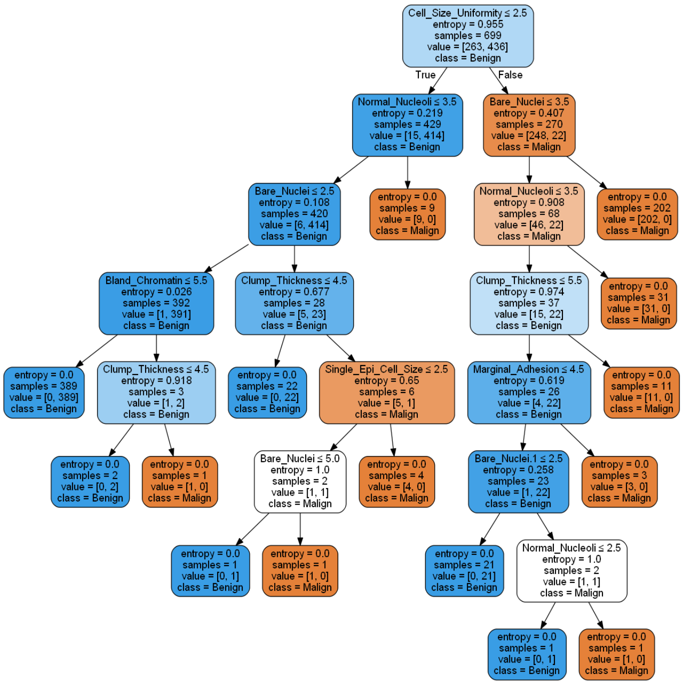


Figure 9: Correlation matrix of features in train set data

The result suggests that **Cell Size Uniformity** is the most important feature that separates the two clusters at **Cell Size Uniformity** = 2.5. **Normal Nucleoli** and **Bare Nuclei** also seem of some importance to refine the clusters further.

Finally, we obtained the silhouette plot for the best clusters and got the following results:

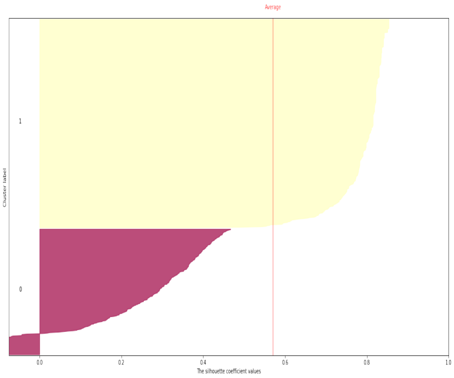


Figure 10: Correlation matrix of features in train set data

Which by analyzing the first plot of Fig 6, is understandable, since the cluster 0 has much more variance, and there are many red points closer to the mean of cluster 1 than to the mean of cluster 0 (even though we are aware that the distances we see in the plot are distorted because of the PCA)

5.2.2 Robustness

For analyzing the robustness of our best model we just compared the results of two different models: Kmeans and Hierarchical Clustering.

Comparing the centroids of Kmeans to the ones of the best Hierarchical Clusterings and the results, we got a mean error of -5.99%, which is low, and shows the robustness of our results.

5.3 Classification results of clustering partitions

The table 2 summarizes the results obtained for the classification using the classes of the partition clusters.

Table 2: Classification results of clustering partitions

Approaches	**Accuracy**	**F1-score**	**R(1)**	**R(0)**	**P(1)**	**P(0)**	**Method**
FS	**0.986**	**0.980**	**0.987**	**0.985**	**0.974**	**0.993**	SVM
3 PC	0.976	0.967	**0.987**	0.970	0.949	**0.992**	Random Forest
5 PC	0.976	0.967	**0.987**	0.970	0.949	**0.992**	Random Forest
WDR	0.995	0.993	1.000	0.993	0.987	1.000	SVM

(Note: FS - Feature Selection, WDR - Without dimensionality reduction)

Overall, the results obtained were higher across all the metrics, as expected. In fact, the heuristic applied in Agglomerative clustering seems to turn data more linearly separable. Thus, the performance of the classifiers increases.

Even though the labels are not the same as the ground truth, we can conclude that the clustering partitions grasped ground truth patterns. In other words, they have a strong association with the original labelling.

5.4 Advantages and disadvantages of the alternatives

5.5 Supervised Learning

5.5.1 Advantages

1. Accurate to predict a malignant and benign cell.
2. Fast predicionts after training the model since the optimal classification rule is established.
3. More interpretable than Unsupervised.

5.5.2 Disadvantages

1. More computationally demanding than Unsupervised in training phase.

5.6 Unsupervised Learning

5.6.1 Advantages

1. Able to discover the impact of features according to a simple heuristic.
2. Less computationally demanding than Supervised in training phase.

5.6.2 Disadvantages

1. To label new data requires a re-run of the entire model.

6 Conclusions

During this project we explored the *Breast Cancer Wisconsin dataset* by applying principal component analysis, analyzing the correlation among feature and data visualization. All of these helped us understand how each feature influenced the class labels in the train set.

Once we had a clear idea of how to the data was distributed, we started working on implementing different approaches in order to come up with an accurate and unbiased supervised learning method. Applying feature selection to the dataset was our best approach, by achieving **95.7% accuracy** and **93.4% f1-score** in the test set.

In the unsupervised phase, we confirmed our two hypothesis. There are two clusters in this dataset and these are hilghly related to the true classes of the original dataset. However, we cannot say that this analysis is particularly useful, given that we have the true classes. It was useful in terms of interpretation and exploration of the dataset, but not so much for the classification goal.

References

[1] Dr. William H. Wolberg. Breast cancer wisconsin (original) data set, 1995.

[2] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

7 Appendix

7.1 Chosen hyperparameters in Supervised methods

7.1.1 Random Forest

1. **n estimators**: number of decision trees in the forest
2. **min samples leaf**: the minimum number of samples required to be at a leaf node.
3. **min samples split**: the minimum number of samples required to split an internal node
4. **max depth**: max depth of the trees

7.1.2 Support Vector Machine

1. **C**: regularization parameter. The strength of the regularization is inversely proportional to C.
2. **kernel**: kernel type to be used in the algorithm. We used Radial Basis Function (RBF) and polynomial with a degree equal to 1 (always).
3. **gamma**: Kernel coefficient.

7.1.3 K Nearest Neighbor

1. **K**: Number of neighbors.
2. **algorithm**: algorithm to compute the nearest neighbors (we used *ball tree* and *kd tree*
3. **leaf size**: Leaf size passed to *ball tree* or *kd tree.* This can affect the speed of the construction and query, as well as the memory required to store the tree.
4. **p**: Distance metric used (manhattan distance and euclidean distance).