Multivariate Analysis Report: Breast Cancer Diagnosis With Supervised and Unsupervised Classification (Wisconsin Dataset)

Multivariate Analysis

Maria Rosário Oliveira,
PhD on Mathematics

Authors:
- **Catarina Rodrigues, ist192434, MMAC**
- **Diogo Monteiro, ist1102093, MECD**
- **Filipa Costa, ist192626, MMAC**
- **Mariana Lopes, ist1102094, MECD**

STRUCTURE AND METHODOLOGY

Breast Cancer Wisconsin (Diagnostic) Dataset

Preprocessing Methodologies

Classification Using Partittion Clusters

Preliminary Analysis

Supervised and Unsupervised Learning

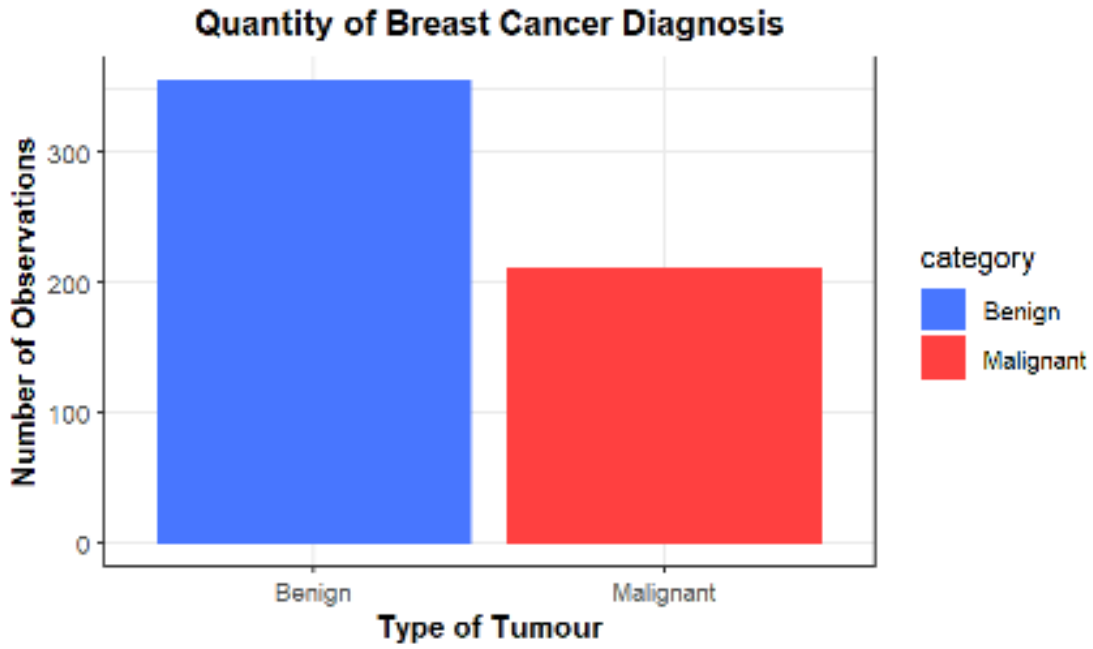
Discussion and Conclusions

BREAST CANCER WISCONSIN (DIAGNOSTIC) DATASET

Attributes of Breast Cancer Wisconsin (Diagnostic) Dataset

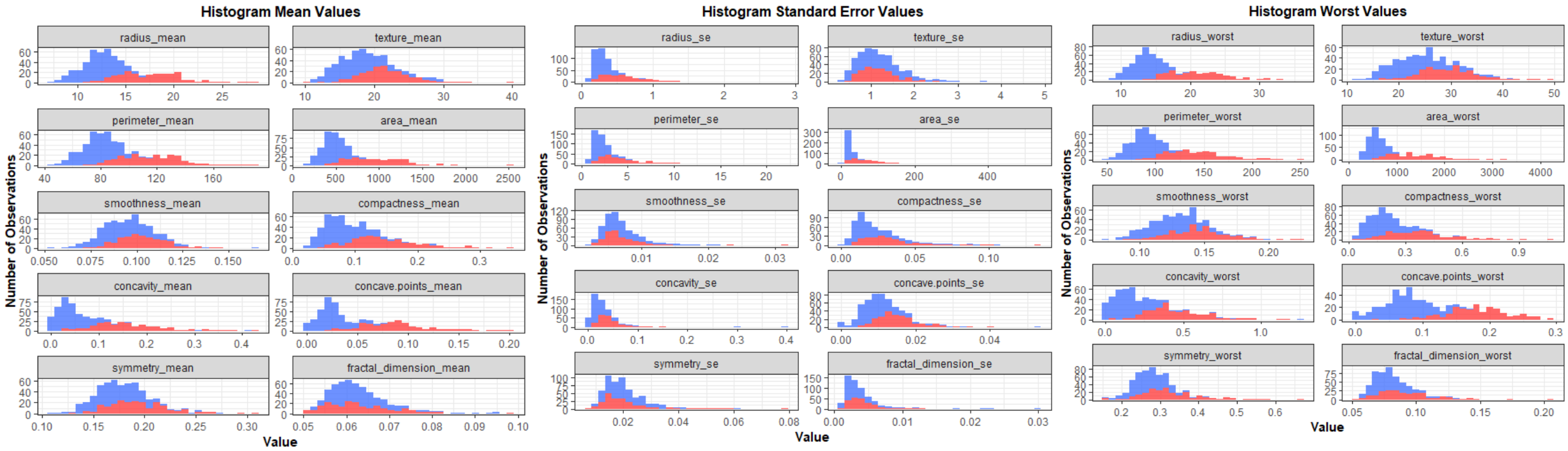
- ID-Number;
- Diagnosis;
- Ten-real values features computed for each nuclear cell:
  - Radius;
  - Texture;
  - Perimeter;
  - Area;
  - Smoothness;
  - Compactness;
  - Concavity;
  - Concave points;
  - Symmetry;
  - Fractal Dimension.

The mean, the standard error and the “worst” (mean of the three largest values) of these features were computed for each image, yielding in 30 features.



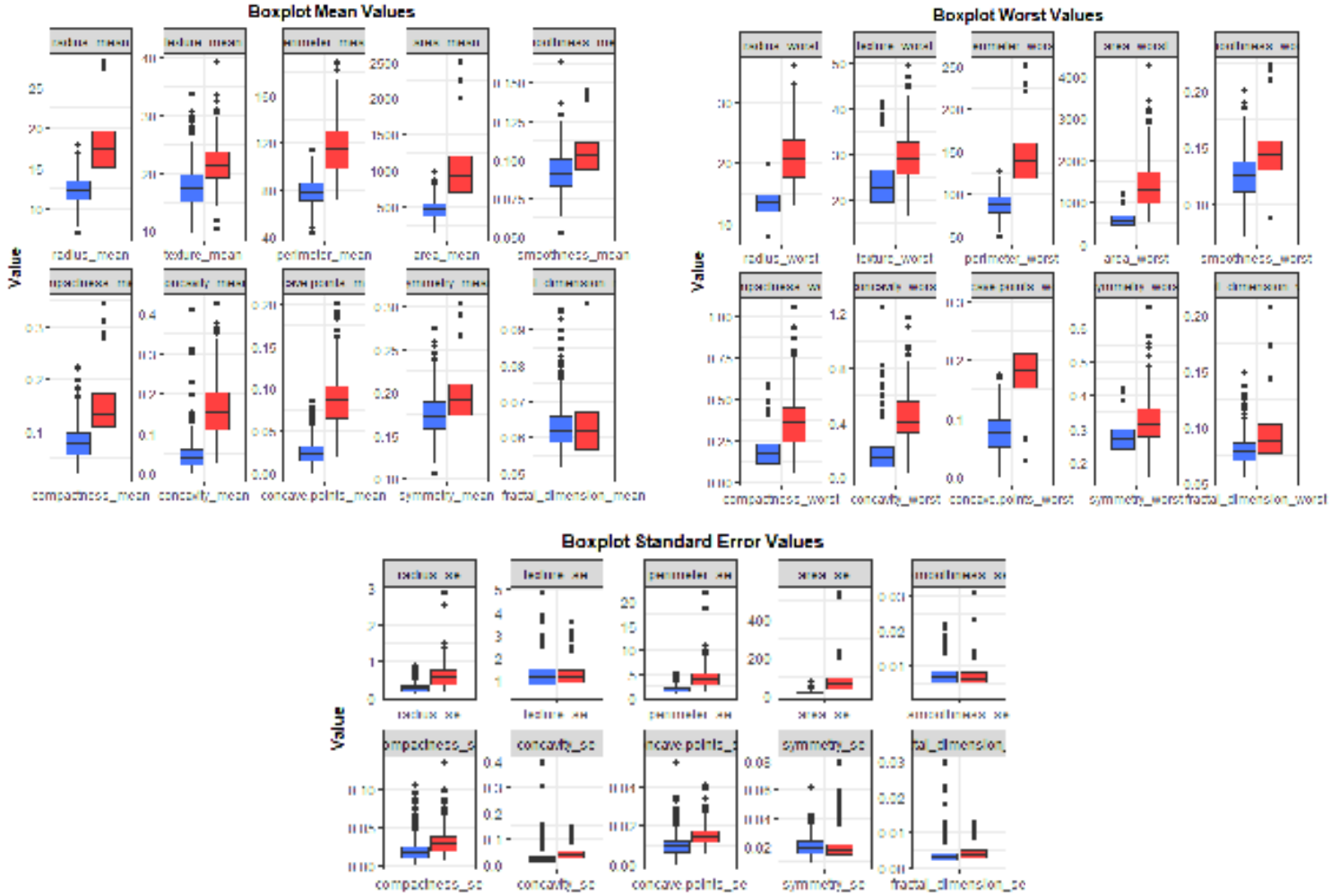
Quantity of the cancer cases in Breast Cancer Wisconsin (Diagnostic) Dataset (Benign=357 , Malignant=212).

PRELIMINARY ANALYSIS



All components mean, standard error and worst values according to the diagnosis (Benign=Blue, Malignant=Red).

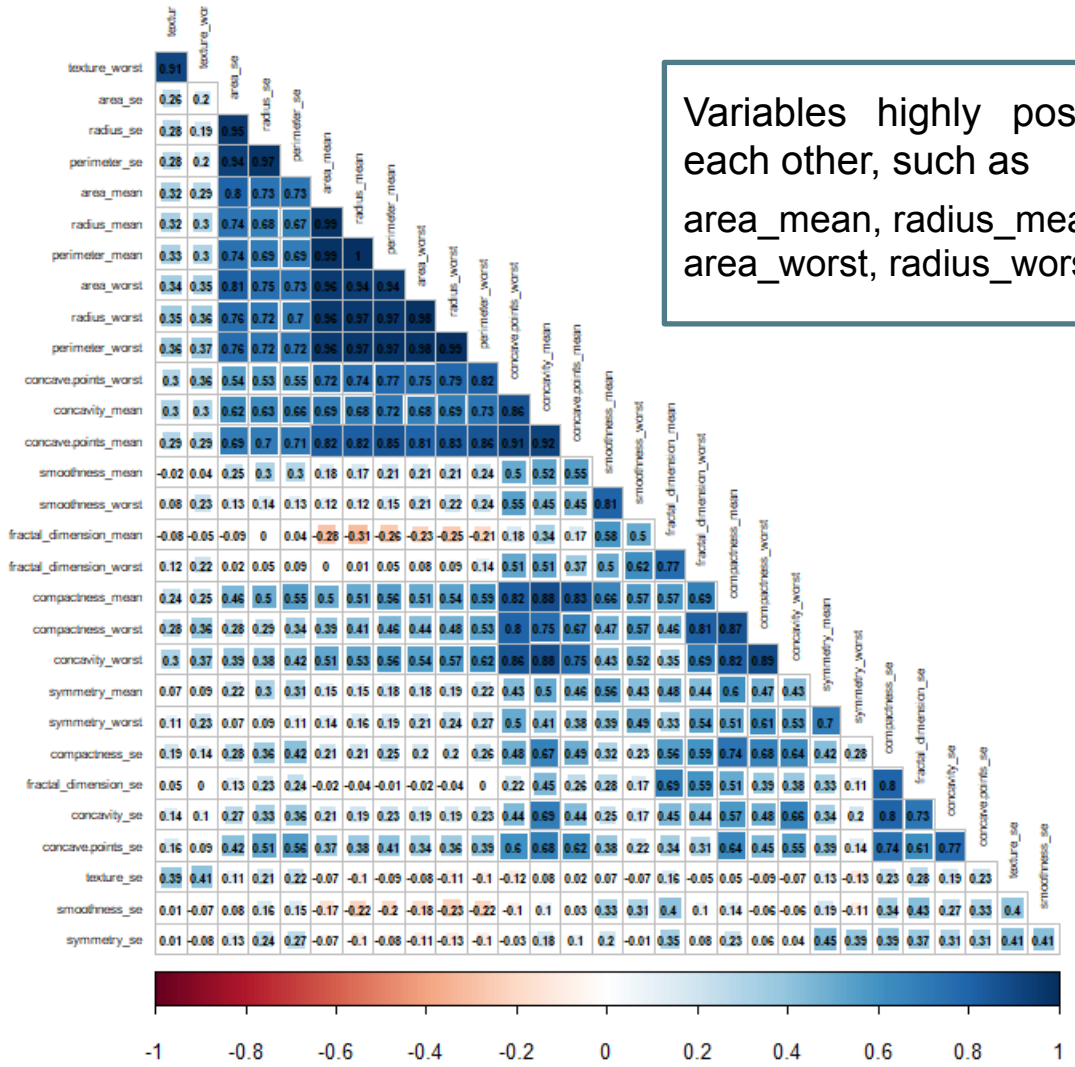
PRELIMINARY ANALYSIS



Variables mean, standard error and worst values according to the diagnosis (Benign=Blue, Malignant=Red).

PRELIMINARY ANALYSIS

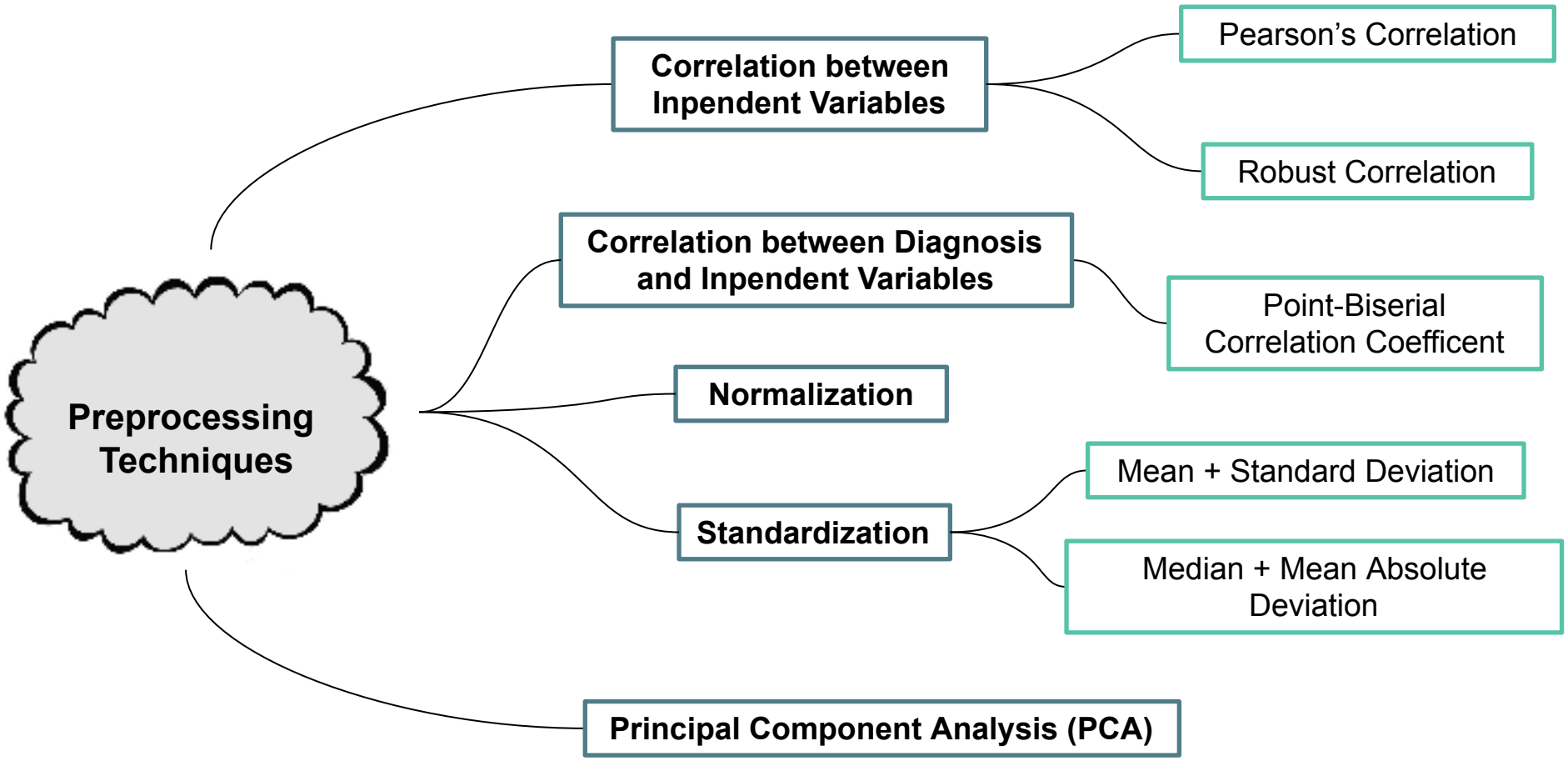
Correlation Matrix



Variables highly positively correlated with each other, such as

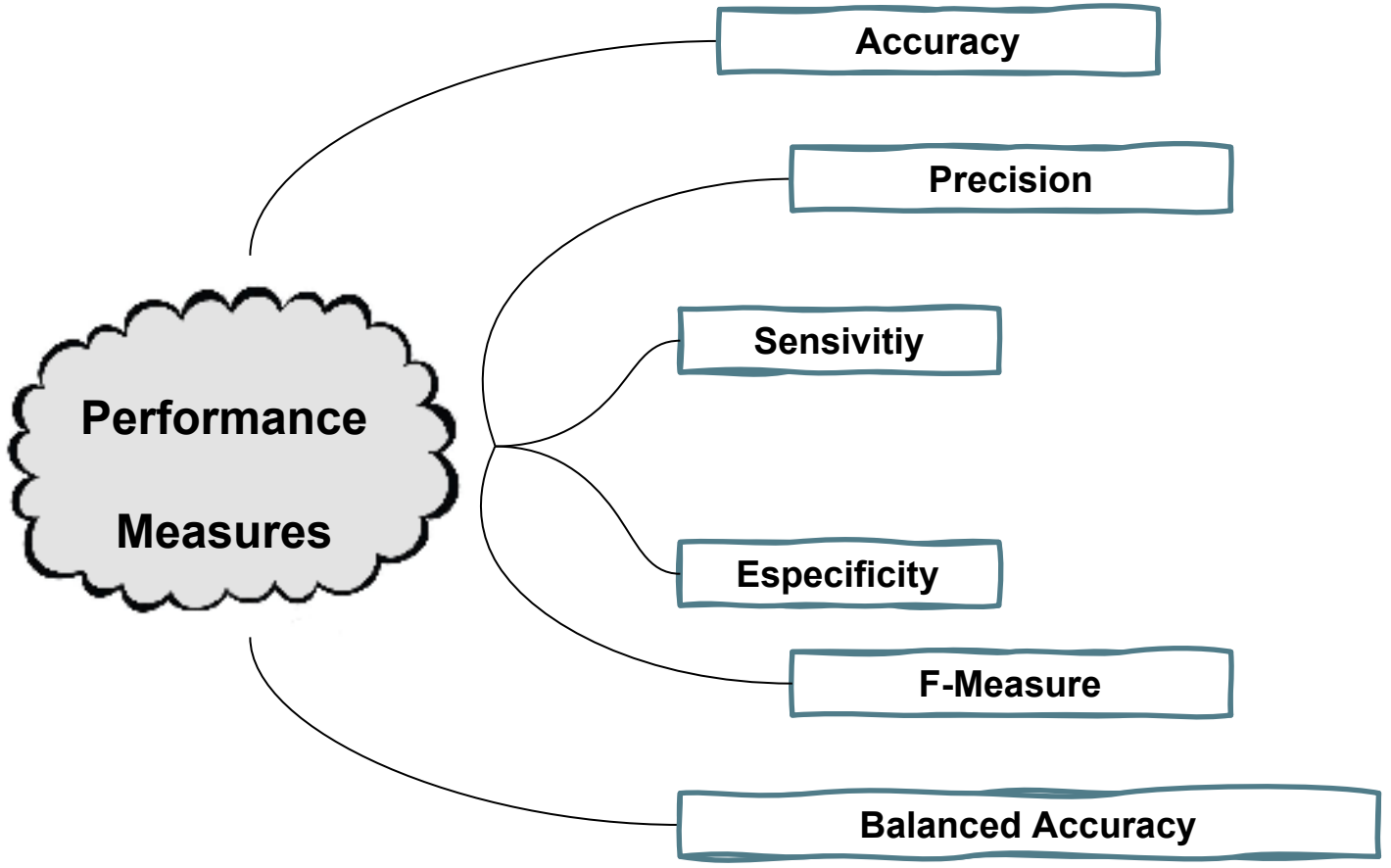
area_mean, radius_mean, perimeter_mean, area_worst, radius_worst.

PREPROCESSING METHODOLOGIES

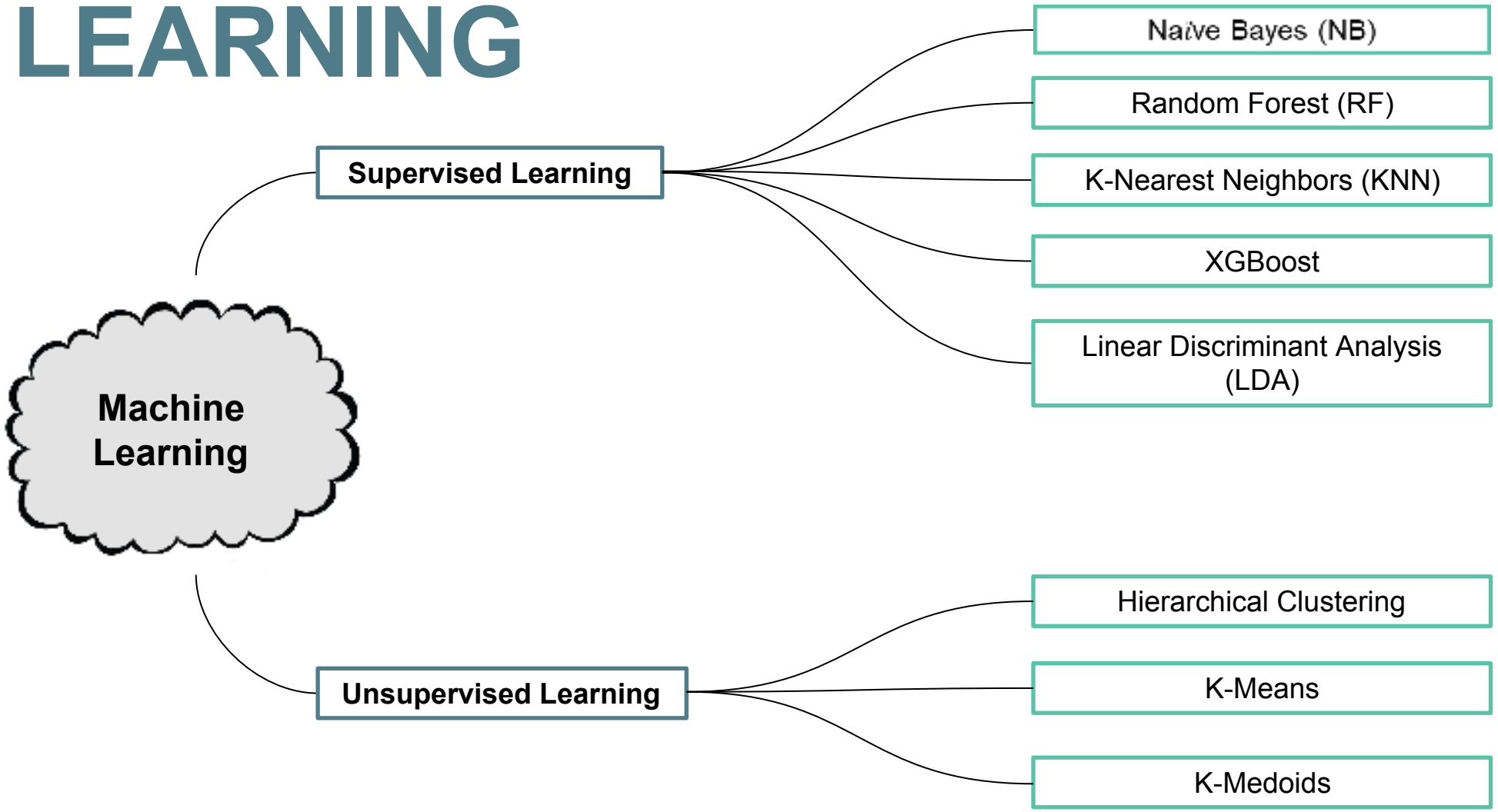


SUPERVISED AND UNSUPERVIDED LEARNING

Validation and Confusion Matrix



SUPERVISED AND UNSUPERVIDED LEARNING

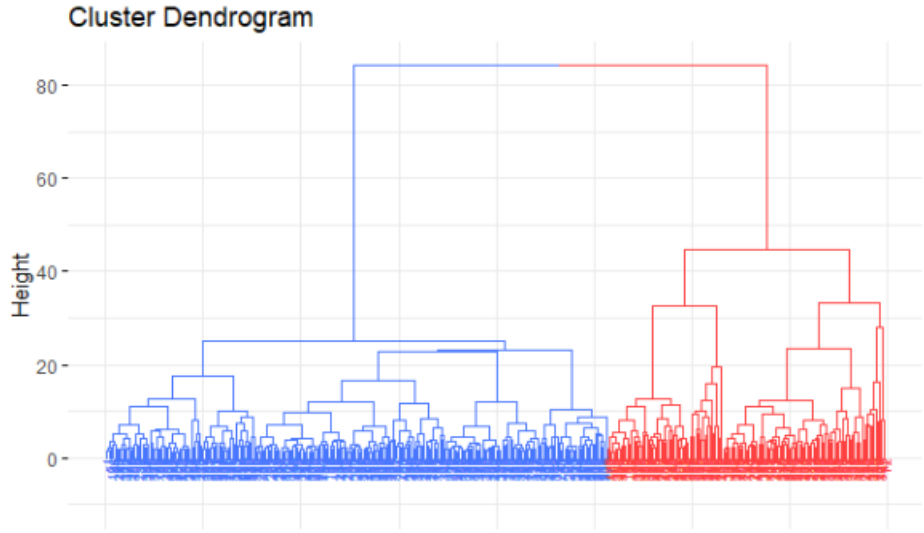


UNSUPERVISED LEARNING

Hierarquical Clustering

TABLE 1: CONFUSION MATRICES USING HIERARCHICAL CLUSTERING(SINGLE LINKAGE, COMPLETE LINKAGE, AVERAGE LINKAGE AND WARD'S METHOD) PERFORMED IN THE TRAIN SET, FOR THE STANDARDIZED DATASET.

Hierarquical Clustering								
Method	Single Linkage		Complete Linkage		Average Linkage		Ward's Method	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
Cluster 1	245	152	246	150	246	150	228	28
Cluster 2	1	0	0	2	0	2	18	124



Dendogram for the Standardized Dataset with Ward's Method

UNSUPERVISED LEARNING

The Best Method

TABLE 2: HIERARQUICAL CLUSTERING, K-MEANS AND K-MEDOIDS FOR THE TRAINING STANDARDIZED DATASET.

Best Clustering Method - train data						
Method	Hierarquical Clustering		K-Means		K-Medoids	
Classes	*B*	*M*	*B*	*M*	*B*	*M*
Cluster 1	228	28	240	28	238	33
Cluster 2	18	124	6	124	8	119
Accuracy	0.8844		0.9146		0.8970	
Sensitivity	0.9268		0.9756		0.9674	

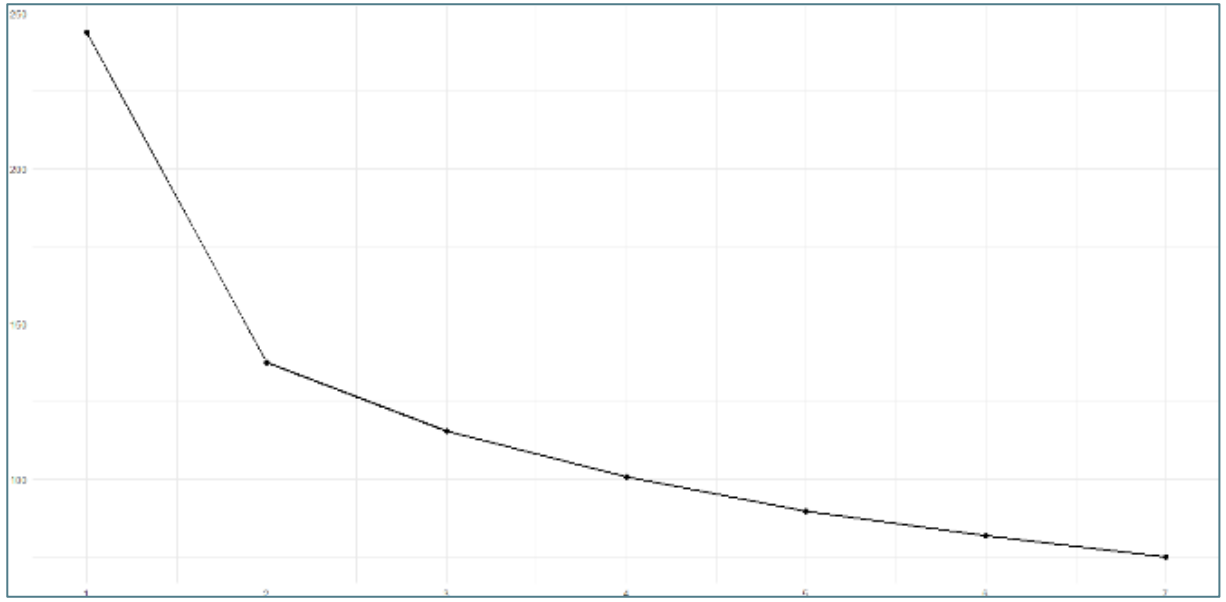
K-Means

UNSUPERVISED LEARNING

The Best Method

TABLE 3: PERFORMANCE MEASURES FOR THE NORMALIZED PCA DATASET, WITH 5 VARIABLES

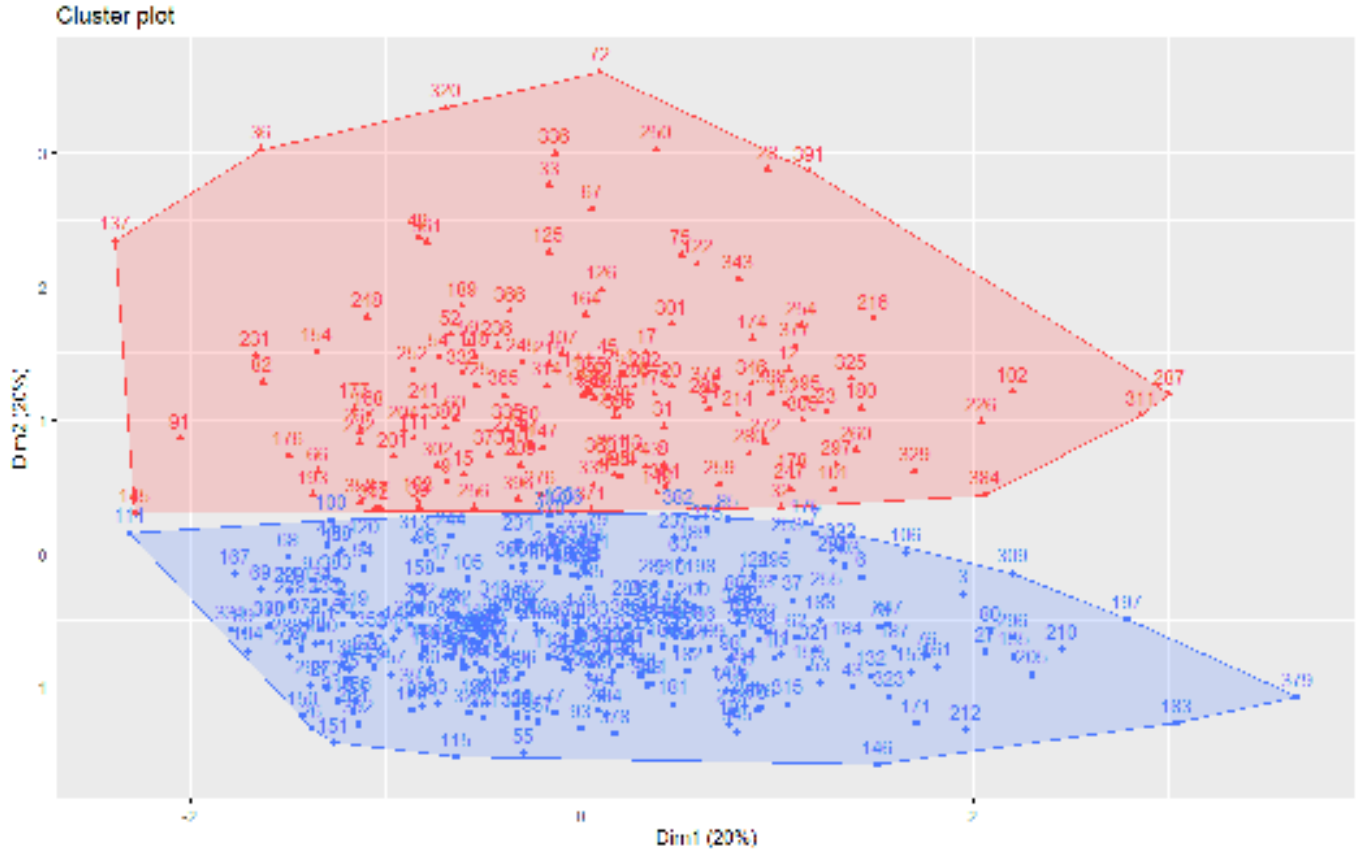
Performance Measures - 5 variables				
Normalization+PCA	Train data		Test data	
Classes	*B*	*M*	*B*	*M*
Cluster 1	237	29	109	8
Cluster 2	9	123	2	52
Accuracy	0.9045		0.9415	
Sensitivity	0.9634		0.9820	
Specificity	0.8092		0.8667	
Balanced Accuracy	0.8863		0.9243	
Precision	0.8910		0.9316	
F1-Score	0.9258		0.9561	



Cluster Plot for the Normalized+PCA dataset

UNSUPERVISED LEARNING

The Best Method



Cluster Plot

SUPERVISED LEARNING

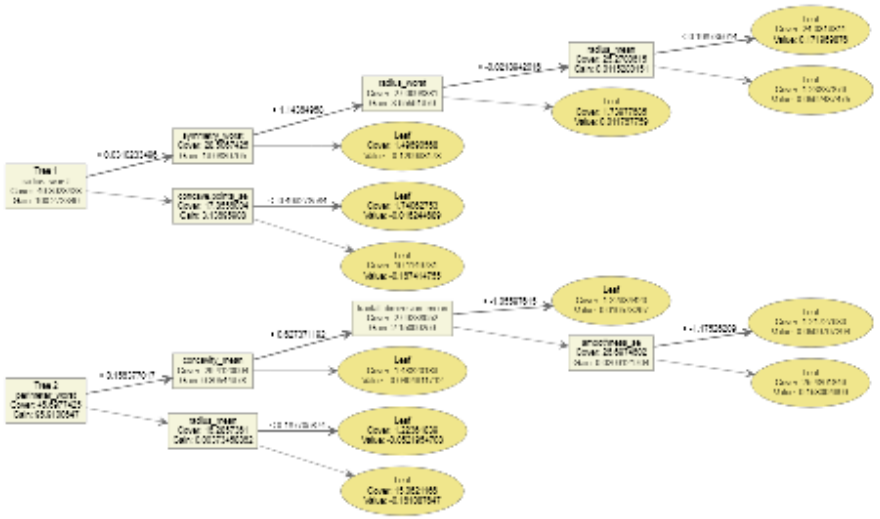
The Best Method

TABLE 4: CONFUSION MATRICES - STANDARDIZED TEST SET

Method	kNN		NB		XGBoost		RF		LDA	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
B	110	4	107	5	110	1	110	3	110	4
M	1	56	4	55	1	59	1	57	1	56

TABLE 5: PERFORMANCE MEASURES - STANDARDIZED TEST SET

Method	kNN	NB	XGBoost	RF	LDA
Accuracy	0.9707	0.9415	0.9883	0.9766	0.9707
Sensitivity	0.9909	0.9549	0.9909	1.000	0.9909
Specificity	0.9333	0.9166	0.9833	0.9333	0.9333
Balanced Accuracy	0.9621	0.9358	0.9871	0.9666	0.9621
Precision	0.9649	0.9549	0.9909	0.9652	0.9649
F1-Measure	0.9777	0.9549	0.9909	0.9823	0.9777



XGBoost results trees for the dataset arising from the analysis of the correlation between independent features and diagnosis variable, comprising three variables.

SUPERVISED LEARNING

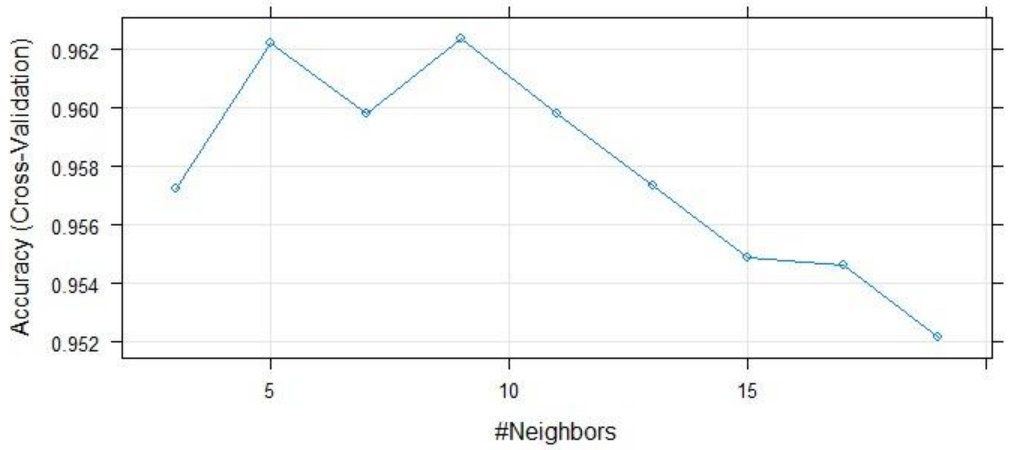
The Best Method

TABLE 4: CONFUSION MATRICES - STANDARDIZED TEST SET

Method	kNN		NB		XGBoost		RF		LDA	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
B	110	4	107	5	110	1	110	3	110	4
M	1	56	4	55	1	59	1	57	1	56

TABLE 5: PERFORMANCE MEASURES - STANDARDIZED TEST SET

Method	kNN	NB	XGBoost	RF	LDA
Accuracy	0.9707	0.9415	0.9883	0.9766	0.9707
Sensitivity	0.9909	0.9549	0.9909	1.000	0.9909
Specificity	0.9333	0.9166	0.9833	0.9333	0.9333
Balanced Accuracy	0.9621	0.9358	0.9871	0.9666	0.9621
Precision	0.9649	0.9549	0.9909	0.9652	0.9649
F1-Measure	0.9777	0.9549	0.9909	0.9823	0.9777



Relation between k values and accuracy of the kNN model.

REPETITION USING CLUSTERS

The Best Method

TABLE 6: CONFUSION MATRIX USING CLUSTERS RESULTS

Method	XGBoost				kNN			
Dataset	Norm		PCA (norm)		Norm		PCA (norm)	
Classes	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
B	110	10	110	7	110	7	110	7
M	1	50	1	53	1	53	1	53

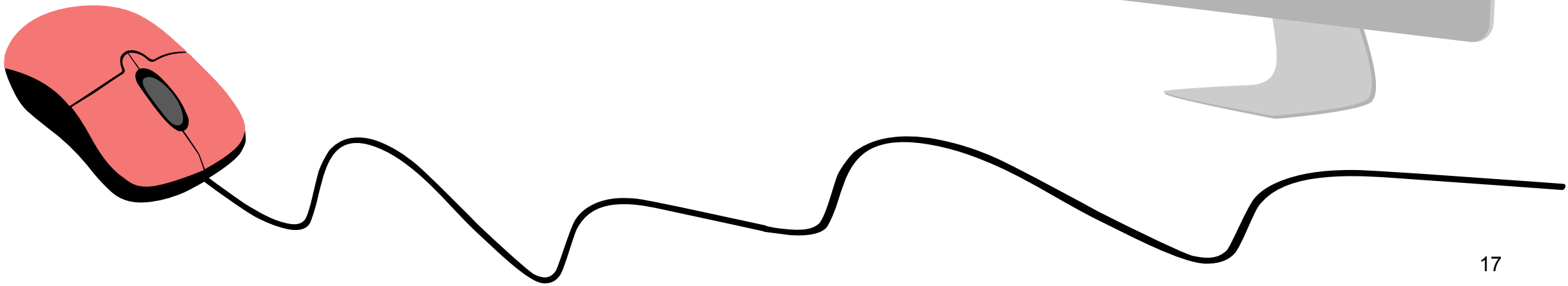
TABLE 7: PERFORMANCE MEASURES USING CLUSTERS RESULTS

Method	XGBoost		kNN	
Dataset	*Norm*	*PCA (norm)*	*Norm*	*PCA (norm)*
Accuracy	0.9356	0.9532	0.9532	0.9532
Sensivity	0.9909	0.9909	0.9909	0.9909

DISCUSSION AND CONCLUSIONS

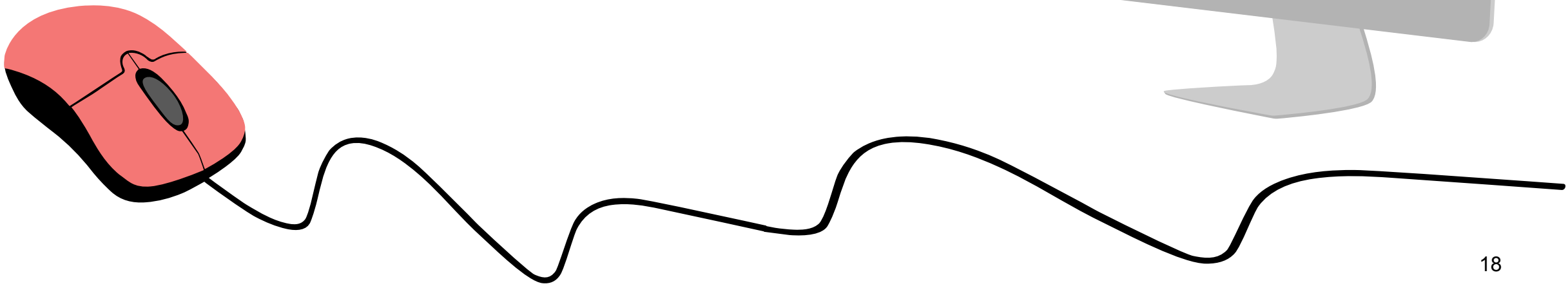
- XGBoost can obtain predictive models around 98% and kNN around 97%;
- The dataset with 3 variables: radius_worst, perimeter_worst and concave. points_worst presented an accuracy of 95% when predicted by XGBoost;

It's possible to stipulate that these three variables are essential and sufficient to classify if a tumour if either benign or malignant.



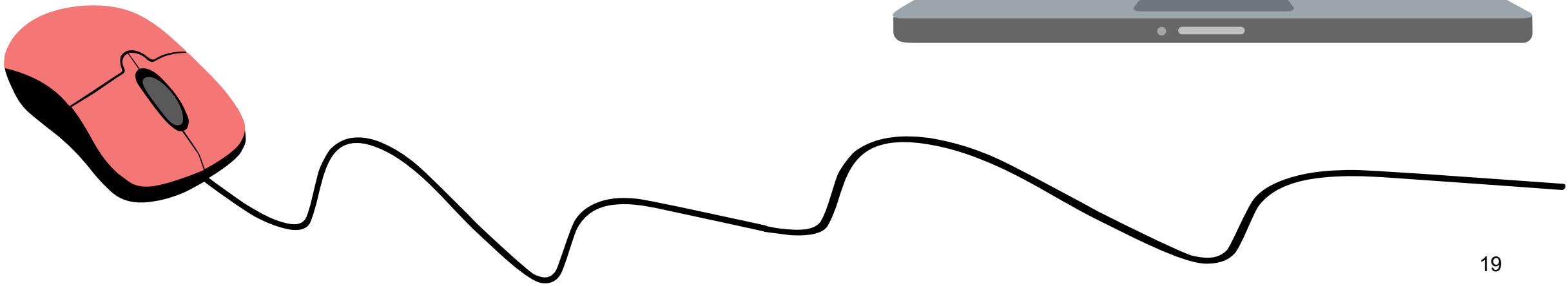
DISCUSSION AND CONCLUSIONS

- Very good results were achieved with the ward's method, and a almost perfect separation of the observations was achieved with a sensitivity of 93%;
- However the best unsupervised method was the K-means because it's vital to have a high sensitivity;
- This way, the K-Means approach not only presented the highest sensitivity (98%), but also the best accuracy (91%);
- When observing both supervised and unsupervised methods, the best results were obtained in datasets that suffered transformations, namely variable reduction.



REFERENCES

- W. N. S. Dr. William H. Wolberg and O. L. Mangasarian. (September 2016) Breast cancer wisconsin (diagnostic) dataset. (accessed: 27.12.2021). [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.



TÉCNICO LISBOA

U LISBOA | UNIVERSIDADE DE LISBOA

Thank you for your attention!

Catarina Rodrigues
catarinasaleirorodrigues@tecnico.ulisboa.pt

Diogo Monteiro
diogo.pinto.monteiro@tecnico.ulisboa.pt

Filipa Costa
filipabcosta@tecnico.ulisboa.pt

Mariana Lopes
mariana.simoes.lopes@tecnico.ulisboa.pt