Mazzi Claudio: PhD Student in AI for Society
Department of Computer Science, University of Pisa
MeS Laboratory, Sant'Anna School for Advanced Studies Pisa
e-mail: claudio.mazzi@phd.unipi.it

Classification algorithms on healthcare data:
A comparative analysis on Wisconsin Breast Cancer Data

CONTENTS

- Introduction and data pre-processing
- Exploratory Data Analysis
- Unsupervised Dimension Reduction with Principal Component Analysis
- Supervised Classification Models
  - Linear Discriminant Analysis
  - *k*-NN model
  - Logistic regressionl with LASSO penalization
  - Support Vector Machines
- Results and Future Development

INTRODUCTION AND DATA PRE-PROCESSING - 1

According to the World Health Organization (WHO), breast cancer is the most common type of cancer among women worldwide

What is breast cancer?

It is the resulting of an abonormal growth of cells in the breast tissue, and the formation of a lump of mass, commonly referred to as a Tumor.

How to detect it?

Magnetic resonance imaging (MRI), mammogram, ultrasound and biopsy. Although these techniques have good detection capability, it is critical to act early in prediagnosis.

INTRODUCTION AND DATA PRE-PROCESSING - 2

Machine Learning (ML) and Deep Learning (DL) techniques have enabled a further step forward in the treatment and prevention of breast cancer specifically

What is the aim of this report?

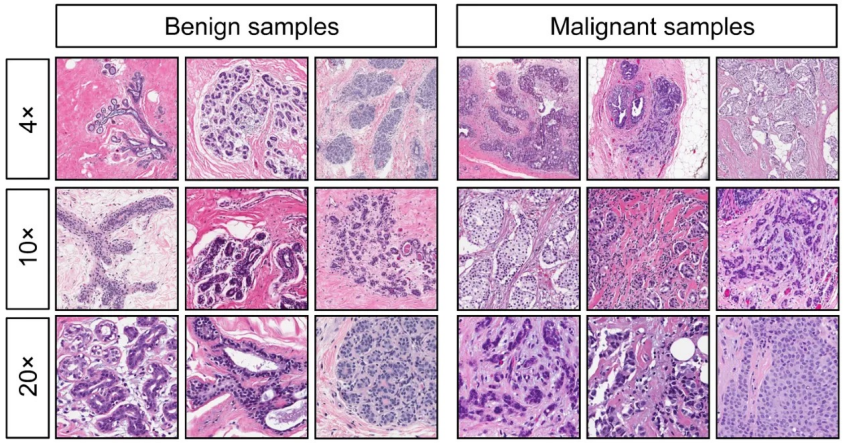
A comparative analysis on different Supervised Classification algorithms on the Wisconsin Breast Cancer Data (WBCD). We study the effeciency in classifying Malignant and Bening tumors of: Linear Discriminant Analysis, k-Nearest Neighbors, logistic regression with LASSO, and Support Vector Machines.

INTRODUCTION AND DATA PRE-PROCESSING - 3

WBDC is composed of 569 instances, with 357 bening (B) and 212 malignant diagnosis (M)
Each instance represents a study of a digitized image of a breast mass.

33 attributes: ID, Diagnosis, X, and 30 numerical features:

- *Radius*
- *Texture (std dev of gray-scale values)*
- *Perimeter*
- *Area*
- *Smoothness (local variation in radius lenghts)*
- *Compactness (pm^2/(area-1)*
- *Concavity (severity of concave portions of the contour)*
- *Concave points (number of concave portions at the contour*
- *Symmetry*
- *Fractal dimension («coastline approx» -1)*



INTRODUCTION AND DATA PRE-PROCESSING - 4

- *Check on missing values*
- *Deleting unuseful columns*
- *Converting «diagnosis» into a categorical value*
- *Standardization of variables*

id	diagnosis	radius_mean	...	symmetry_worst	fractal_dimension_worst	X
842302	M	17.99	...	0.4601	0.11890	NA
842517	M	20.57	...	0.2750	0.08902	NA
84300903	M	19.69	...	0.3613	0.08758	NA
84348301	M	11.42	...	0.6638	0.17300	NA
84358402	M	20.29	...	0.2364	0.07678	NA
843786	M	12.45	...	0.3985	0.12440	NA

diagnosis	radius_mean	texture_mean	...	concave.points_worst	symmetry_worst	fractal_dimension_worst
M	1.0960995	-2.0715123	...	2.2940576	2.7482041	1.9353117
M	1.8282120	-0.3533215	...	1.0861286	-0.2436753	0.2809428
M	1.5784992	0.4557859	...	1.9532817	1.1512420	0.2012142
M	-0.7682333	0.2535091	...	2.1738732	6.0407261	4.9306719
M	1.7487579	-1.1508038	...	0.7286181	-0.8675896	-0.3967505
M	-0.4759559	-0.8346009	...	0.9050914	1.7525273	2.2398308

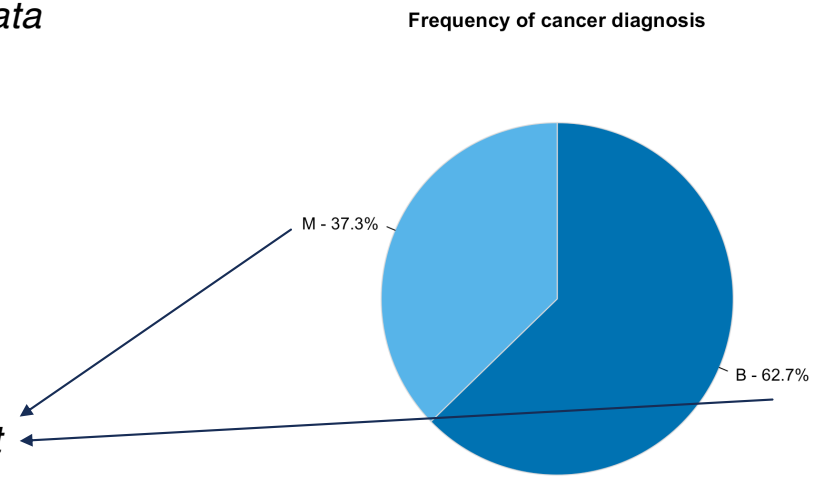
After pre-processing: 569 records vs. 31 features

EXPLORATORY DATA ANALYSIS - 1

Exploratory Data Analysis aims to:

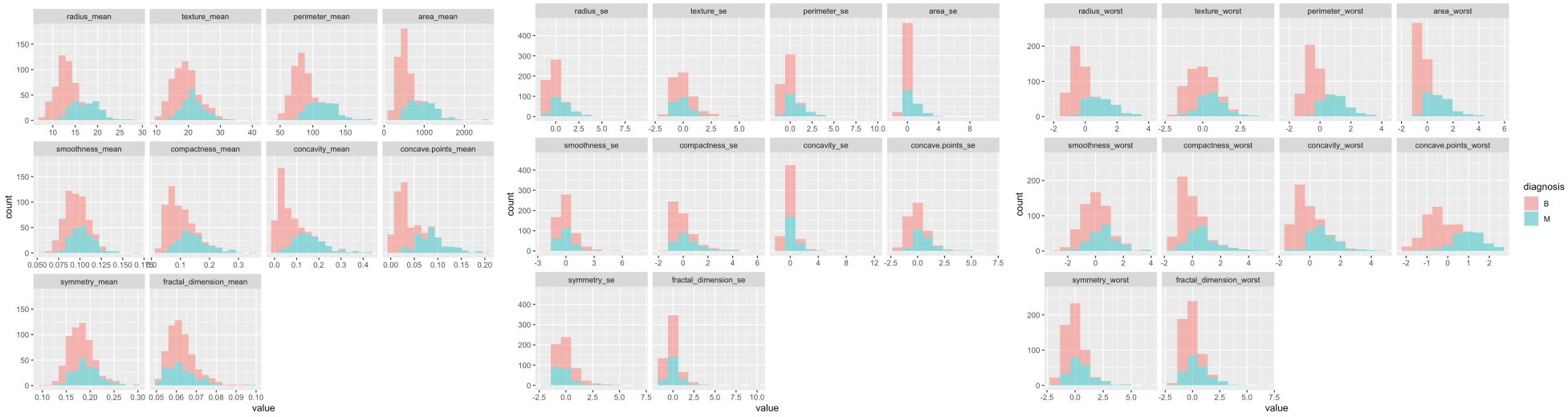
- *Summarize the main characteristics of the data*
- *Identify patterns*
- *Detect anomalies*
- *Check hypotheses and assumptions*

Diagnosis distribution:
357 Bening – 212 Malignant



EXPLORATORY DATA ANALYSIS - 2

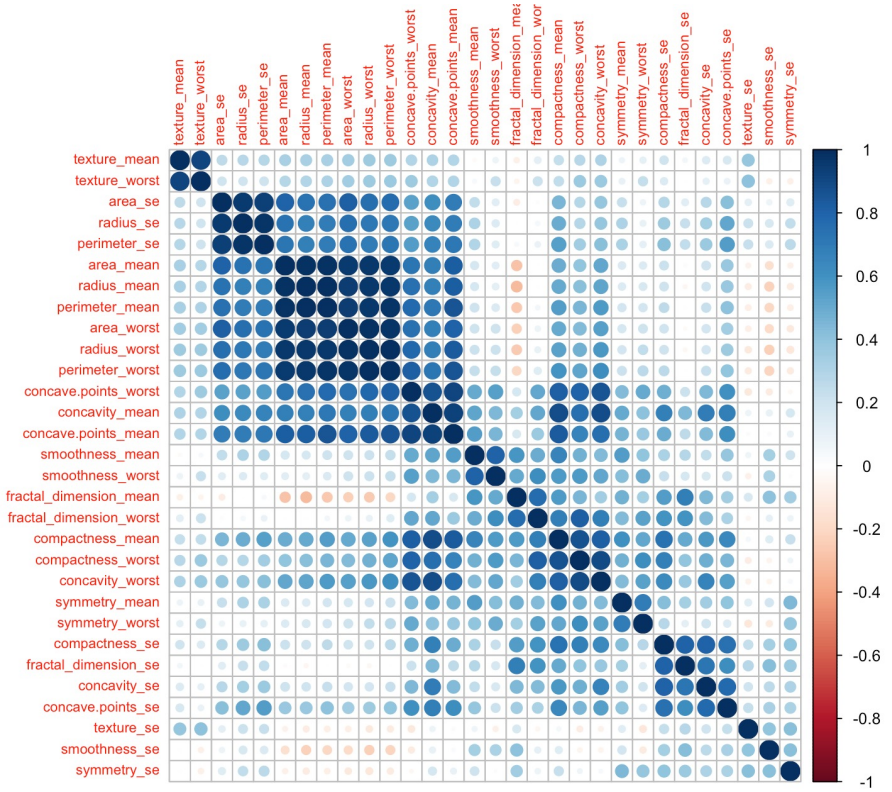
Separation of variables for:
mean – standard error - worst



EXPLORATORY DATA ANALYSIS - 3

Correlation Matrix aims to identify possible dependences within variables. To understand what features are important in the analysis.

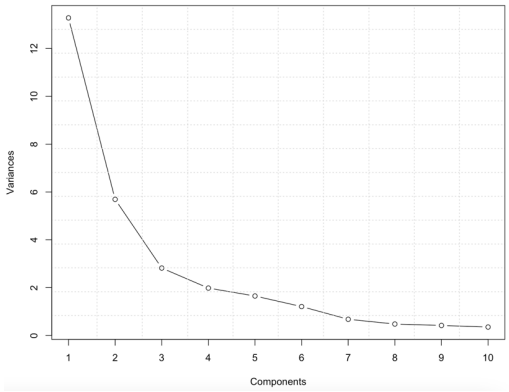
Dimensional reduction is an important fine-tuning prescription for predictive Machine Learning algorithm.



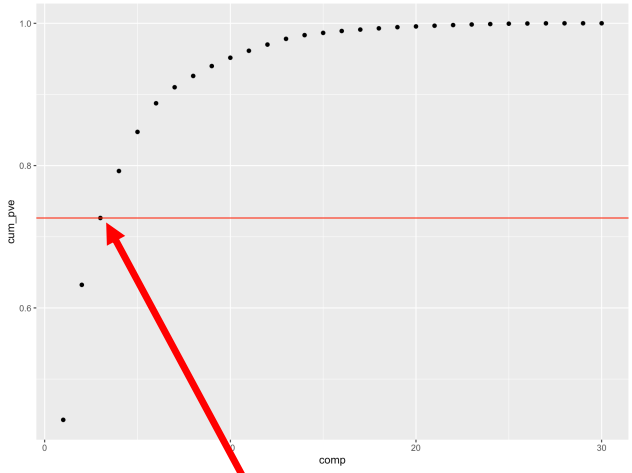
UNSUPERVISED DIMENSION REDUCTION WITH PRINCIPAL COMPONENT ANALYSIS - 1

Principal Component Analysis (PCA) is a Dimension Reduction technique for unsupervisedproblems. It projects the data into a lower-dimensional space so that the lower-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension

Spectral decomposition on Principal Components (eigenvectors of the associated covariance matrix)



Selection of PCs based on PVE

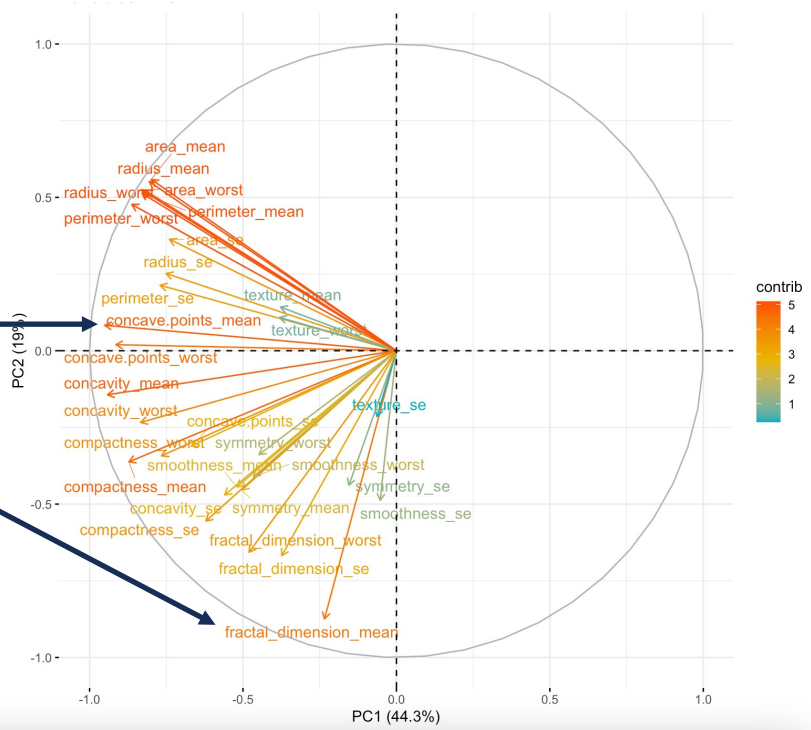


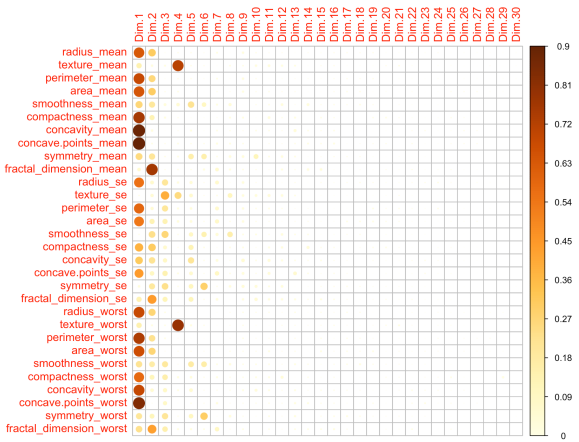
3 first PCs explain
the 73% of the variance

UNSUPERVISED DIMENSION REDUCTION WITH PRINCIPAL COMPONENT ANALYSIS - 2

The PC1 is influenced by *concavity*

The PC2 is influenced by the *fractal dimension*





Rotation matrix after spectral decomposition

SUPERVISED CLASSIFICATION MODELS

Classification models *generalize and improve regression models to work with* *qualitative response variables*

In WBCD → DIAGNOSIS: M/B

1) Divide data set into *training (70%) and test (30%)* data
2) Training process: optimization problem on the training set
3) Validation: by Cross-Validation technique
4) Evaluation of the metrics: *Accuracy, Precision, Recall, F1-score,* and *ROC curve*
5) Prediction process

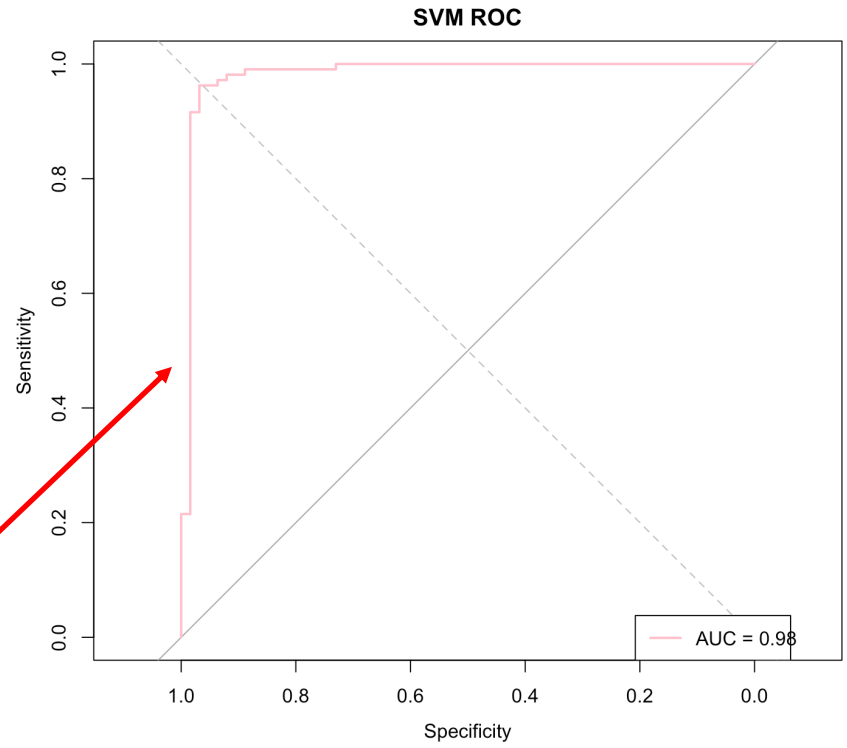
We test the efficiency of LDA, *k*NN, logistic regression with LASSO, and SVMs on classifying Bening and Malignant tumors

LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA):

- Based on *Bayes' Theorem*
- Given predictors' behavior → Gaussian for WBCD

```
> lda.confusion <- confusionMatrix(data = lda.class, reference = data_std$diagnosis, positive = "M")
> lda.confusion
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 355  18
         M   2 194

               Accuracy : 0.9649
                 95% CI : (0.9462, 0.9784)
    No Information Rate : 0.6274
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9236

 Mcnemar's Test P-Value : 0.0007962

            Sensitivity : 0.9151
            Specificity : 0.9944
         Pos Pred Value : 0.9898
         Neg Pred Value : 0.9517
             Prevalence : 0.3726
         Detection Rate : 0.3409
   Detection Prevalence : 0.3445
      Balanced Accuracy : 0.9547

       'Positive' Class : M
```

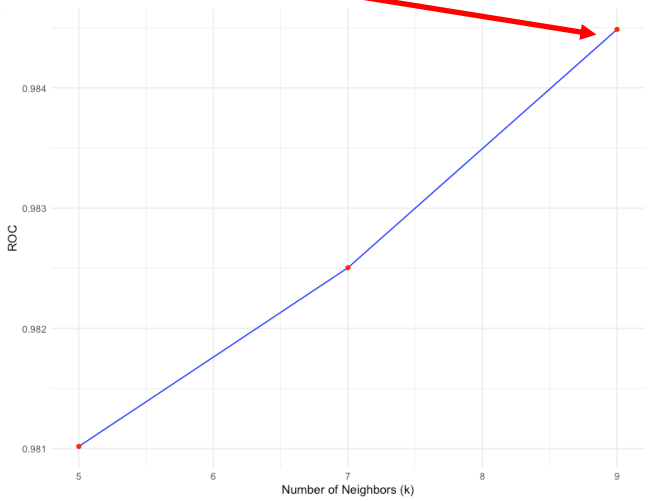


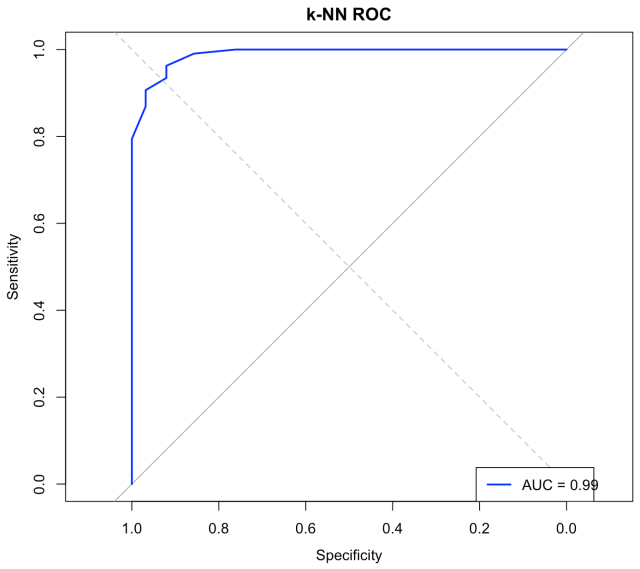
FP vs TP as the threshold for a positive/negative decision changes

K-NEAREST NEIGHBOURS

K-Nearest Neighbors (*k*NN):
non-parametric supervised classification method
Parameter *k* selected t through *cv* → *k=9*

```
> knn_confusion <- confusionMatrix(knnPredict, data_test$diagnosis, positive
> knn_confusion
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 103   5
         M   4  58

               Accuracy : 0.9471
                 95% CI : (0.9019, 0.9755)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8861

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9206
            Specificity : 0.9626
         Pos Pred Value : 0.9355
         Neg Pred Value : 0.9537
             Prevalence : 0.3706
         Detection Rate : 0.3412
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9416

       'Positive' Class : M
```



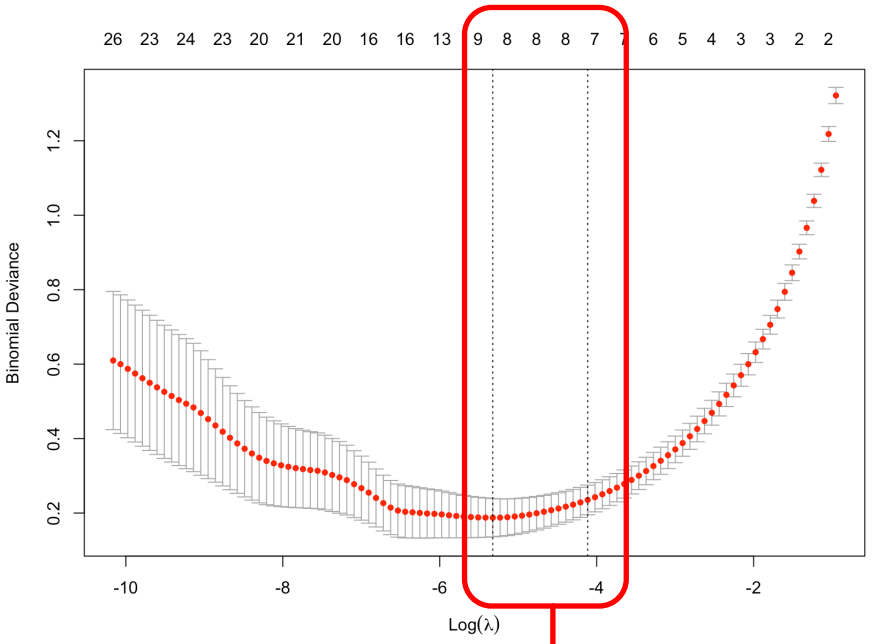


LOGISTIC REGRESSION WITH LASSO

LASSO regression model:

- Introduce a natural regularization on the system
- Generalized Linear Model with a *penalization term driven by* λ → sparsity of the data
- By cross-validation: $\lambda = 0.0049$

```r
# LASSO regression model
fitControl <- trainControl(method = "cv", number = 10, classProbs = TRUE,
    summaryFunction = twoClassSummary)

lassoFit <- train(diagnosis ~ .,
                  data = data_train,
                  method = "glmnet",
                  trControl = fitControl,
                  tuneGrid = expand.grid(alpha = 1, lambda =
                      seq(0.001, 0.1, by = 0.001)))
lassoPredict <- predict(lassoFit, newdata = data_test, type = "prob")
```



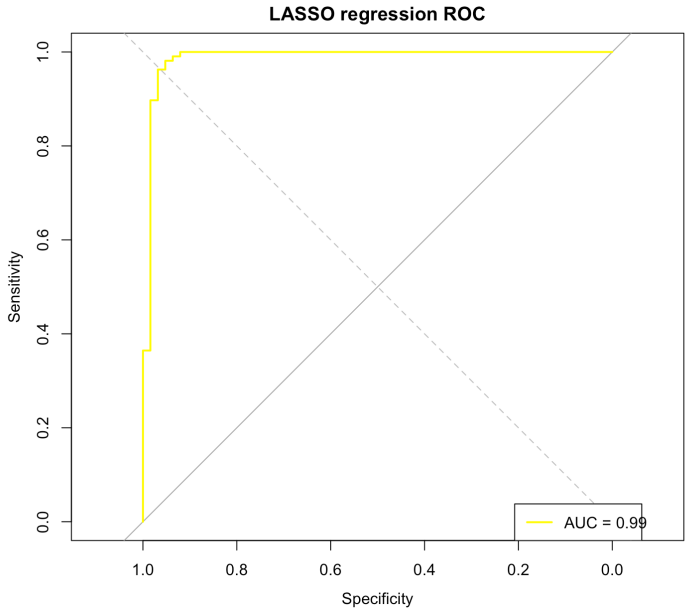

```
> print(selected.variables)
[1] "concave.points_mean"  "radius_se"            "fractal_dimension_se"
[4] "radius_worst"         "texture_worst"        "smoothness_worst"
[7] "concavity_worst"      "concave.points_worst" "symmetry_worst"
```

LOGISTIC REGRESSION WITH LASSO

LASSO optimization reduces bias and introduce a regularization in the data set. Now, we apply *logistic regression* on the unbiased and reduced set, spanned by the previous 9 selected features.

```
> logistic_confusion <- confusionMatrix(logisticPredict, data_test_selected$diagnosis, positive = "M")
> print(logistic_confusion)
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 105   3
         M   2  60

               Accuracy : 0.9706
                 95% CI : (0.9327, 0.9904)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9367

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9524
            Specificity : 0.9813
         Pos Pred Value : 0.9677
         Neg Pred Value : 0.9722
             Prevalence : 0.3706
         Detection Rate : 0.3529
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9668

       'Positive' Class : M
```

On the new data set with 9 features.
In this case we notice no differences in applying the regression in the entire data set, and in the reduced one.



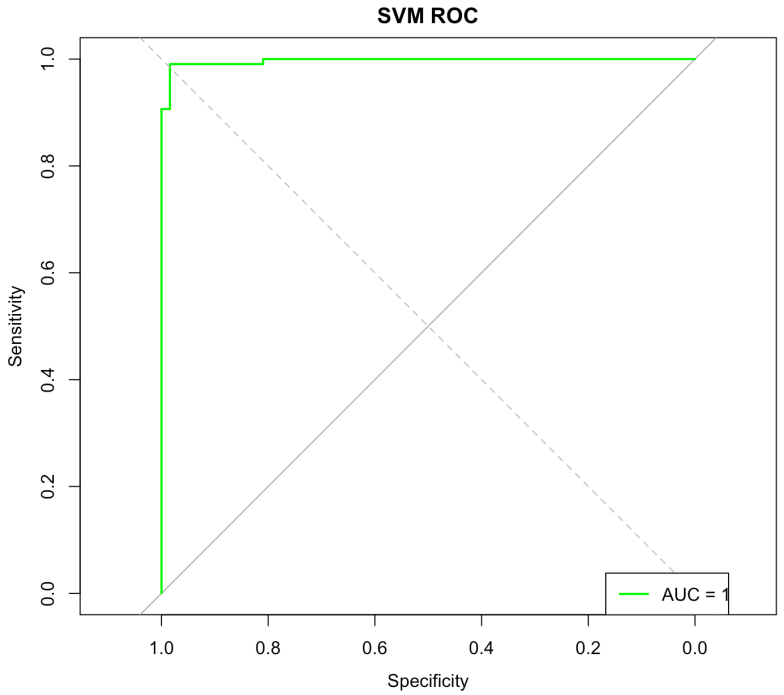
SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs):

- Robust on noisy data
- *Linear* Kernel trick
- More expensive computational costs

Best performance

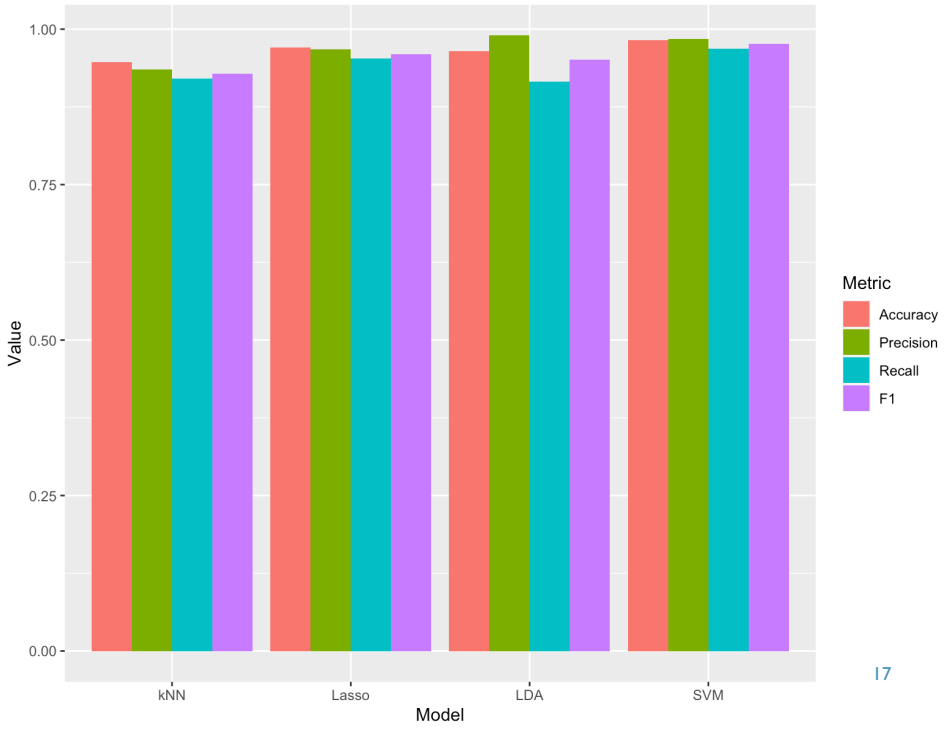
```
> svm_confusion <- confusionMatrix(svmPredict, data_test$diagnosis, positive = "M")
> svm_confusion
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 106   2
         M   1  61

               Accuracy : 0.9824
                 95% CI : (0.9493, 0.9963)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.962

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9683
            Specificity : 0.9907
         Pos Pred Value : 0.9839
         Neg Pred Value : 0.9815
             Prevalence : 0.3706
         Detection Rate : 0.3588
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9795

       'Positive' Class : M
```



RESULTS AND FUTURE DEVELOPMENT - 1

Final comparison of the four models on predicting Bening and Malignant breast cancer

	Model	Accuracy	Precision	Recall	F1
1	SVM	0.98	0.98	0.97	0.98
2	kNN	0.95	0.94	0.92	0.93
3	LDA	0.96	0.99	0.92	0.95
4	Lasso	0.97	0.97	0.95	0.96



RESULTS AND FUTURE DEVELOPMENT - 1

- SVMs stand out for their outstanding accuracy and precision of 98% → the most reliable model for distinguishing between benign and malignant tumors. Moreover, the F1-score, which balances accuracy and recall, suggests that SVMs are the most balanced model among the four options, with a score of 98%.

- *k*NN and LDA show competitive results, with close to 95% accuracy, but with slight differences in other metrics. However, the LDA model generally performs better than kNN due to the possibility of a linear splitting of the response, as well as the Gaussian behavior of the predictors.

- LASSO logistic regression shows that only a few predictors are informative for tumor classifications. Regression with the LASSO penalization also provides an automatic regularization of the data set → better then LDA and kNN.

It is fundamental to note that the choice of model also depends on the context and the specific goal of the analysis.

RESULTS AND FUTURE DEVELOPMENT - 1

In the future, we are interested in moving the analysis outside the diagnosis prediction context and testing supervised and unsupervised learning in data sets describing the clinical pathways for generic chronic patients (IMA, diabetes). Using statistical learning algorithms, we want to mine the clinical pathways for chronic patients from data concerning hospitalization, pharmaceutical prescriptions, and ambulatory services.

REFERENCES

1) World Health Organization (WHO), 2024. https://www.who.int/news-room/factsheets/detail/breast-cancer.

2) MartInez-Campa C, Menendez-Menéndez J, Alonso-Gonzalez C, González A, Alvarez Garciaa V, Cos S. «What is known about melatonin, chemotherapy and altered gene expression in breast cancer.» Oncol Lett. 2017.

3) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor. «An introduction to Statistical Learning: with Applications in R.» New York: Springer, 2013.

4) Liu Congcong, Gao Jinsong, Liu Juntao, Wang Xietong, He Jing, Sun Jingxia, Liu Xi-aowei, and Liao Shixiu. «Predictors of Failed Intrauterine Balloon Tamponade in the Management of Severe Postpartum Hemorrhage.» Frontiers in Medicine, Vol.8. 2021.

5) *Bennett Kristin P. and Campbell Colin. «Support Vector Machines: Hype or Hallelujah?» SIGKDD Explorations, Vol. 2. 2000*

6) *Cortes C. and Vapnik V. «Support-vector networks.» Machine Learning, Vol. 20, 273{297}. 1995.*

7) *Ezio Bottarelli and Stefano Parodi. «Un approccio per la valutazione della validit a dei test diagnostici: le curve R.O.C. (Receiver Operating Characteristic).» Ann. Fac. Medic. Vet. di Parma Vol. XXIII. 2003.*

8) *Francesca Ferr e, Chiara Seghieri, Andrea Burattin, and Andrea Vandin «Process Mining and Clinical Pathways: an application to Breast cancer data in Tuscany.» Proceedings of 4th International Workshop on Process-Oriented Data Science for Healthcare. 2021*

Q&A SECTION

THANK YOU FOR THE ATTENTION

Mazzi Claudio: PhD Student in AI for Society

Department of Computer Science, University of Pisa

MeS Laboratory, Sant'Anna School for Advanced Studies Pisa

e-mail: claudio.mazzi@phd.unipi.it