Statistical Machine Learning Task

Predicting Prostate Cancer Status : A Machine Learning Approach

Agbatan Fiacre Luc KOUDERIN

African Institute for Mathematical Sciences (AIMS - Rwanda)

December 16, 2024

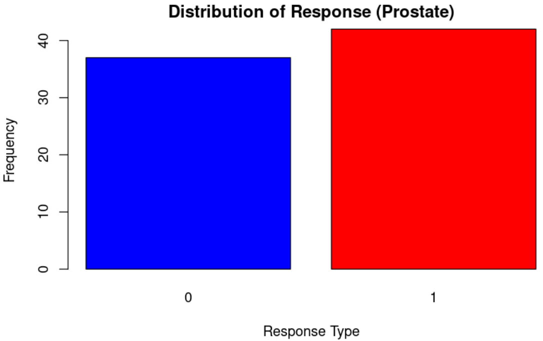
Outline

- Data description and manipulation

- Models fitting and comparison

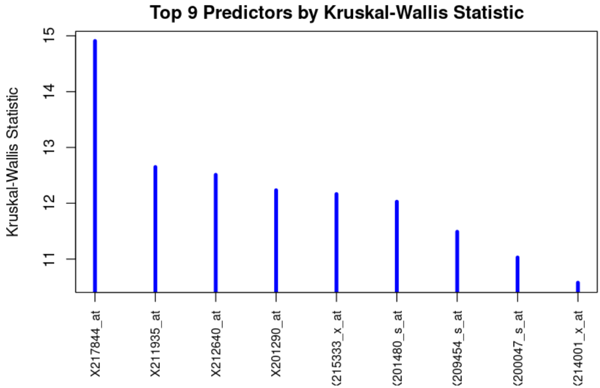
Data description and manipulation

- 501 variables (one response variable, 500 covariates) with 79 observations
- Distribution of the response variable named Y :



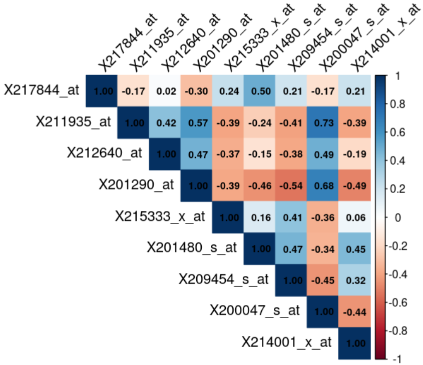
Data description and manipulation

We are in a case of curve of dimensionality, let's see the best predictors.



Data description and manipulation

- Have a look on the correlations between the top best predictors

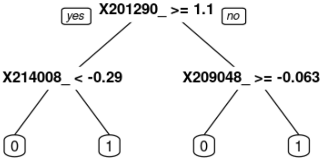


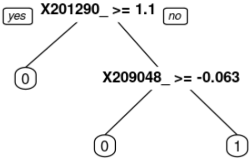
	X217844_at	X211935_at	X212640_at	X201290_at	X215333_x_at	X201480_s_at	X209454_s_at	X200047_s_at	X214001_x_at
X217844_at	1.00	-0.17	0.02	-0.30	0.24	0.50	0.21	-0.17	0.21
X211935_at		1.00	0.42	0.57	-0.39	-0.24	-0.41	0.73	-0.39
X212640_at			1.00	0.47	-0.37	-0.15	-0.38	0.49	-0.19
X201290_at				1.00	-0.39	-0.46	-0.54	0.68	-0.49
X215333_x_at					1.00	0.16	0.41	-0.36	0.06
X201480_s_at						1.00	0.47	-0.34	0.45
X209454_s_at							1.00	-0.45	0.32
X200047_s_at								1.00	-0.44
X214001_x_at									1.00

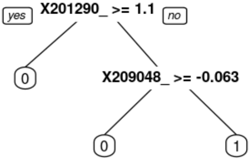
- By doing the eigen decomposition of the correlation matrix we get the ratio $= 17.02$

Models fitting and comparison

- Let's start with a tree with differents values for the complexity parameter.

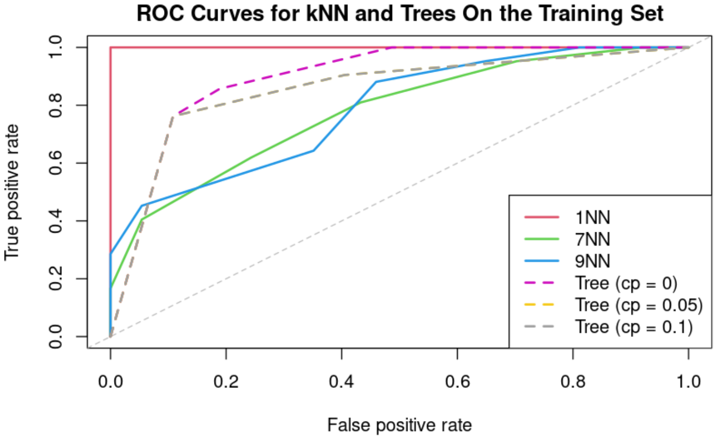






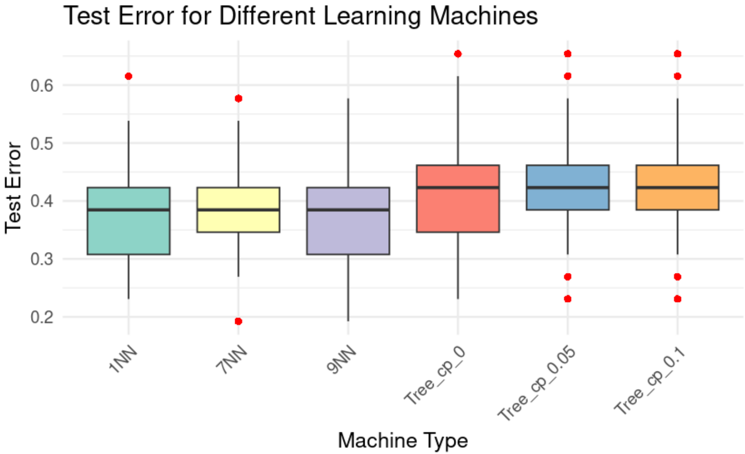
Models fitting and comparison

- Let's add to the previous model a kNN model with differents values of k.



Models fitting and comparison

- Let's do predictions on the test set.



Test Error for Different Learning Machines

Overall Conclusion

- A krustal Wallis Test can be done to have the best predictors variables.
- For this dataset, the kNN machine perform better than the Trees.
- Thers no signficant difference in the predictive merits of 1NN, 7NN and 9NN but in pratice, we will choose 9NN because its the least complex model between this three models.