A MACHINE LEARNING APPROACH TO GENE EXPRESSION ANALYSIS FOR CANCER CLASSIFICATION

Discussed by Vincenzo Marciano',
282004

MATHEMATICS FOR MACHINE LEARNING

WHY MACHINE LEARNING IN GENOMICS

- **Efficient**
- **Rapid**
- **Reliable**
- **Robust to mistakes**

Currently, biomedical research attempts to explain the mechanisms by which develops a particular disease, for this reason, **gene expression studies** have proven to be a great resource.

In a parallel way, **Machine Learning** techniques spread out in a multitude of applications, derived from other severals Science branches: medical diseases' classification is one of them. From Parkinson to Breast Cancer, more and more improvements have be made through the years.

LEUKEMIA: *AML vs ALL*

Our purpose in this work is to study a considerable selection of ML algorithms in order to achieve the best performance possible, at least for those who are the expected final results in cancer classification of two types of Leukemia:

AML → starts in cells that become **lymphocytes**, i.e. white blood cells that are an important part of your immune system. About **80% of childhood leukemia** cases are this type.

AML → begins in early **myeloid cells**, i.e. cells that become white blood cells (other than lymphocytes), red blood cells, or platelet-making cells. It's the most common type of **leukemia in older people**.

PROBLEM OVERVIEW

DATASETS

- Training Set (7129 gene descriptors x 38 patients)
- Test Set (7129 gene descriptors x 34 patients)
- Label Set (categorical, 72 rows)

CLEAN AND TIDY DATA

- Exchanging rows with column: gene activations as features, patients as samples
- Binary Encoding: ALL = 0, AML = 1
- Standardization

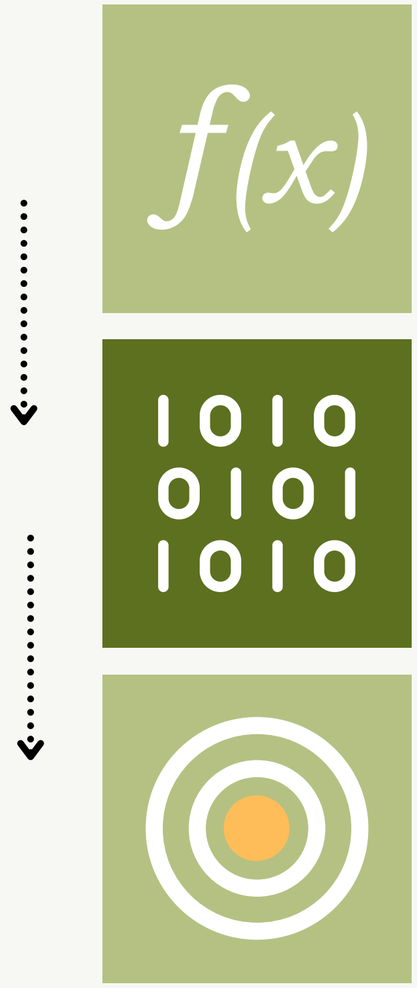
BINARY CLASSIFICATION PROBLEM

WHAT COULD WE EXPECT?

EXPLOITING MACHINE LEARNING TECHNIQUES

- **PCA** mixed with certain classifiers (for example **linear SVM, K-means, Naive Bayes**) is an intuitive idea, since we have a very high dimensionality data to deal with.

- **Random Forest Classifier** behaves very well even if we don't apply any reduction to the set: it works with several subsets of data in a very efficiend **decorrelated** way.



EVALUATION METRICS

NOT THE SAME STUFF!

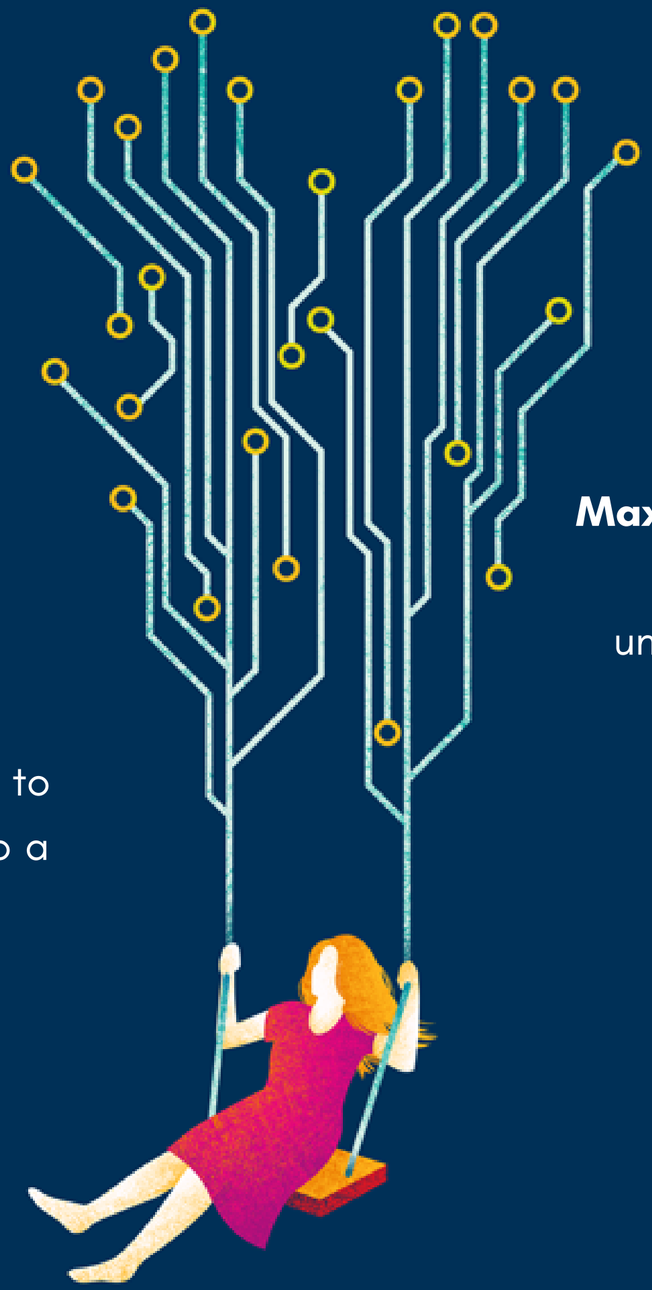
Evaluation Metrics are truly important tools in determining how well our classifier is behaving. The widely-used metric is **Accuracy** (0 to 1) and it will be exploited also in this work.

Nevertheless, taken alone, it says nothing about our model in a "specific" way: we are dealing with cancer classification and we cannot permit to ourselves a large margin of error in False Positives / False Negatives predictions. That's why **MCC** metric comes in handy and explains to the user the overall behaviour of our classifier, from -1 to +1 score.

PCA

PACKED DATA

PCA is a very powerful tool which aims to reduce a high dimensionality set of data into a bunch of compressed features.

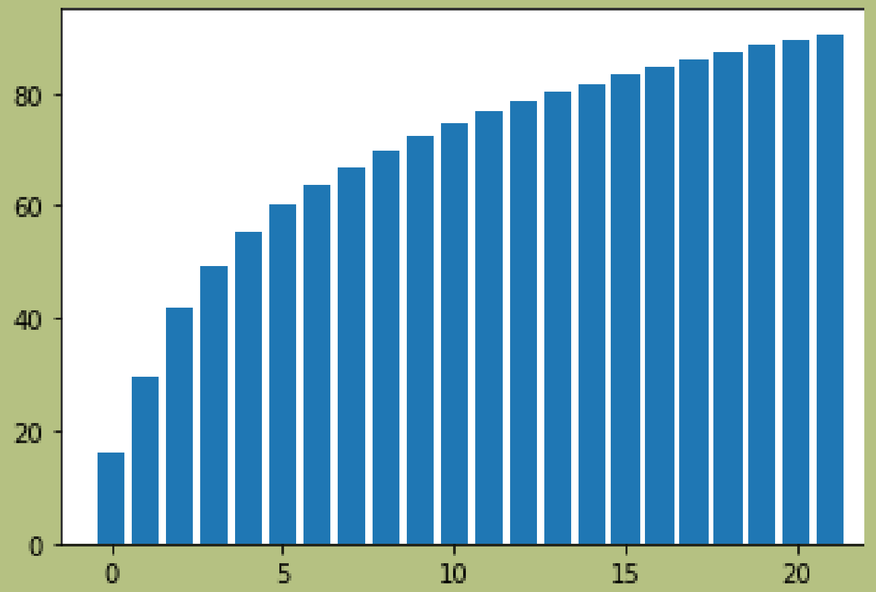


The problem is focused in the **Variance Maximation** carried by our features and, given a reasonably threshold of **90%**, in understanding how much of them contributes to the overall **Explained Variance Ratio** spread.

EIGEN-DECOMPOSITION

DEEPER

Through a deep analysis in the Eigendecomposition of our **Covariance Matrix** and sorting in descending order all the eigenvector by their contribution given to the overall Variance, we achieve a **almost-zero** value by the 38th element.



Applied to our previously defined Explained Variance Ratio threshold we obtain eventually a **22-dimension data**: not bad at all!

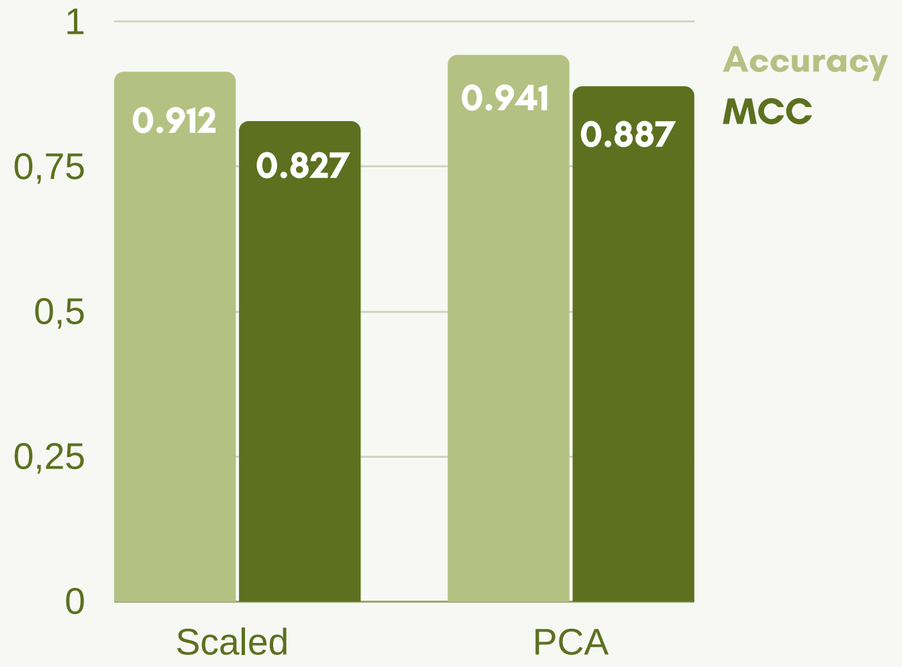
MACHINE LEARNING MODELS **SVM**

MASTERING THE HYPERPLANE

Support Vector Machines *(SVM) are widely used to tackle the sample complexity and computational complexity challenges raised up by the high dimensionality of feature space.*

Kernel operations:

- exploited in order to avoid long and computationally extensive calculations
- dealing with **inner products** among instances instead of raw functions mapping



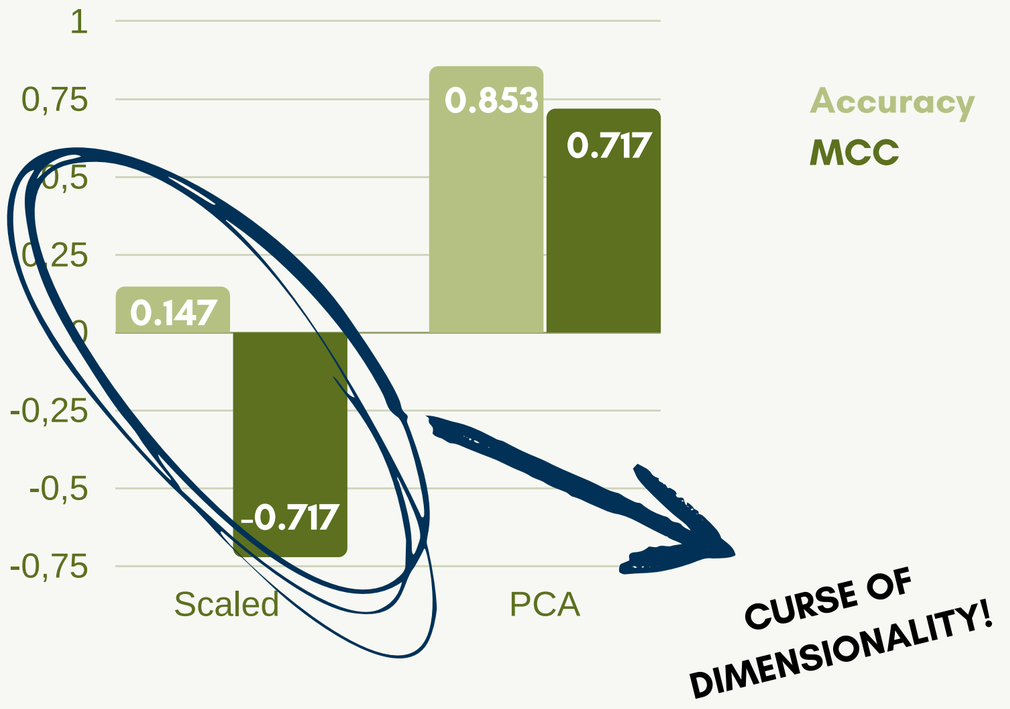
MACHINE LEARNING MODELS K-MEANS

GROUP 'EM ALL!

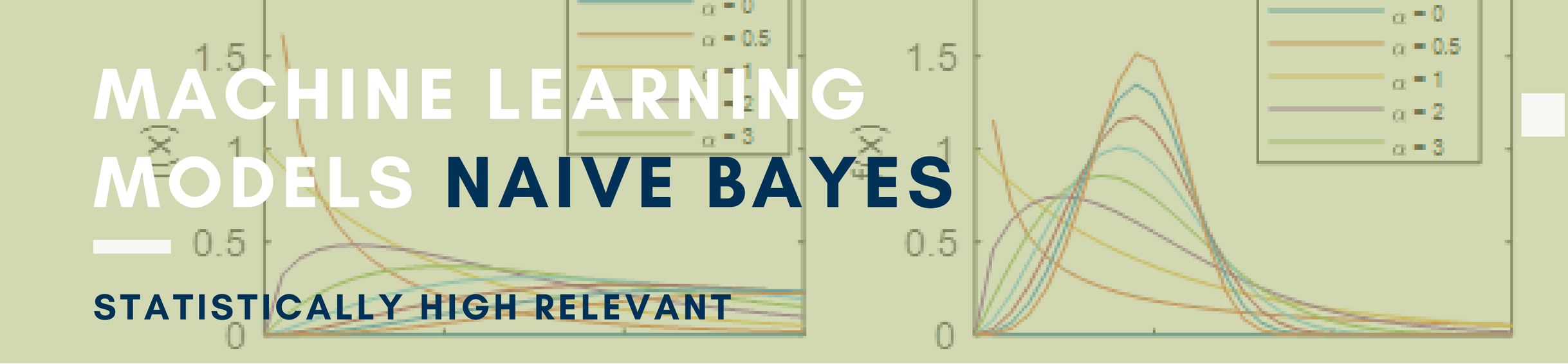
Even if we observe the lack of ground truth, which is a common problem in unsupervised learning methods, ***Clustering*** *still remains one of the most powerful way to organize the data in some meaningful way.*

K-Means Clustering:

- Approximation of the Clustering optimization problem.
- Choose a parameter k such that it's the most suitable for the given problem:
- Clustering by finding minimum distance between the ith-centroid and a sample



MACHINE LEARNING MODELS **NAIVE BAYES**

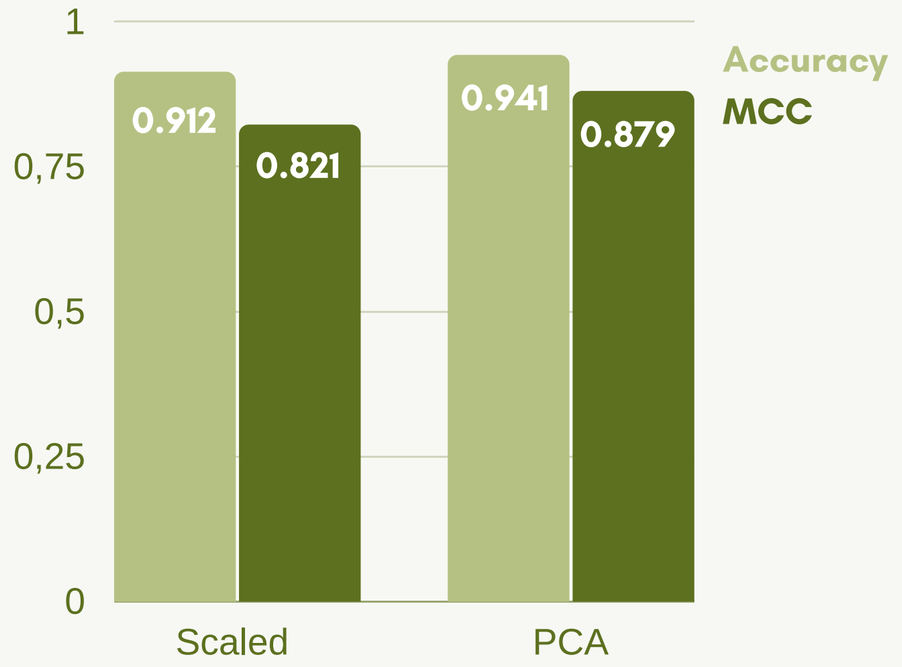


STATISTICALLY HIGH RELEVANT

*The **Naive Bayes classifier** is a classical demonstration of how generative assumptions and parameter estimations simplify the learning process. Sometimes is computationally easier to estimate the parameters of the model than to learn a discriminative predictor*

Gaussian Naive Bayes:

- Gene descriptors' behaviour modelled as Gaussian RV
- Using MLE to find estimates
- Naive assumption: all the features are independent variables



RANDOM FOREST

DIVIDI ET IMPERA

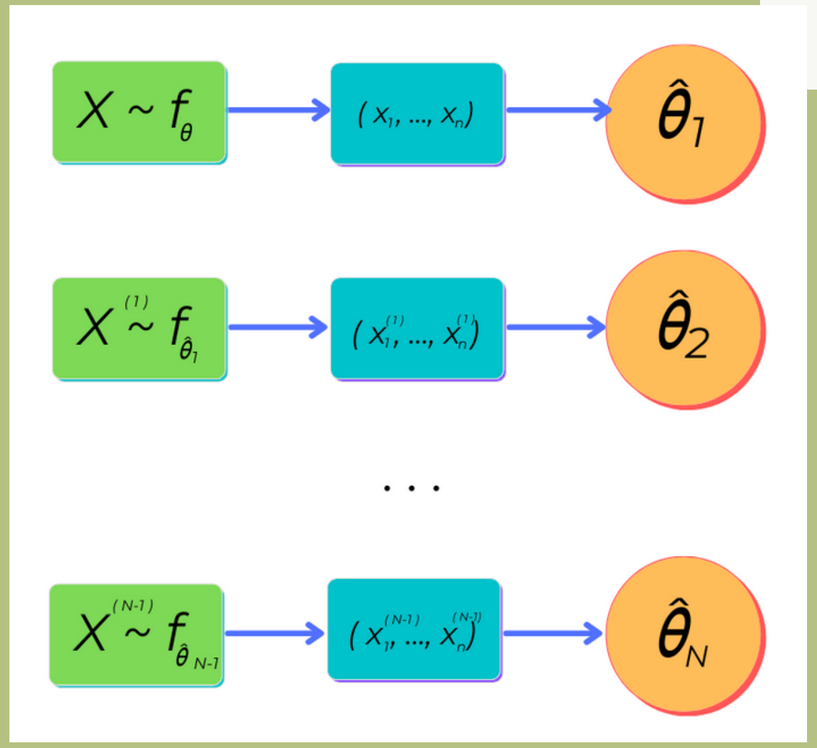
Random forest is a classifier consisting of an ensemble of trees (a collection of decision trees) where each final prediction is based on the majority vote (averaging for Regression Tasks) over the predictions among each group of trees.

The problem is focused in finding a way to create an ensamble of trees which guarantees. the **decorrelation** among them. This is possible by exploiting the **Bootstrap** technique, which may prevent the same subset of informations to be processed again, and a fine **Hyperparameters Tuning** process.

BOOTSTRAP

RESAMPLING IS THE KEY

Bootstrap goal is to repeatedly select (N times) a random sample **with replacement** of the training set X and fit the trees with this updated data



That's why Bootstrap sampling is a **de-correlated way** to approach each tree family: by showing them different training sets everytime, it decreases the overall variance, yet maintaining the same bias.

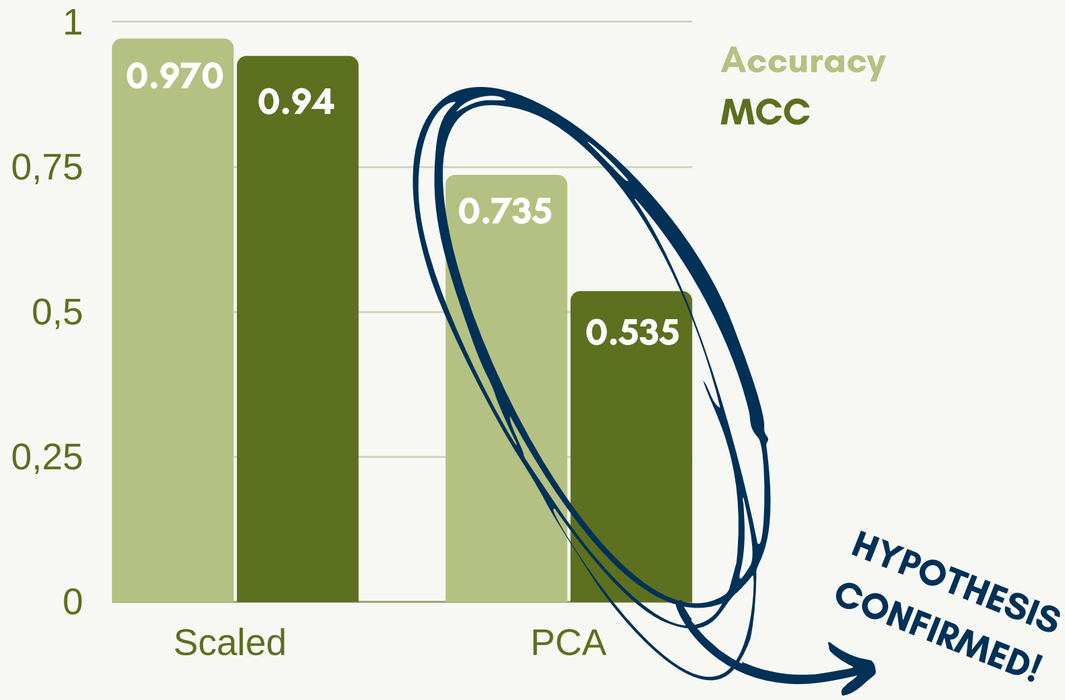
RANDOM FOREST
HYPERPARAMETERS TUNING

STATISTICALLY HIGH RELEVANT

One can be interested in understanding how the PCA reduced set could be useful in our RF application: nevertheless, we need to take in mind that each child tree better improves its learning by studying original (or at least scaled) features. This is done thanks to a ***GridSearch*** *above all proposed configurations.*

Hyperparameters to be tuned - BEST:

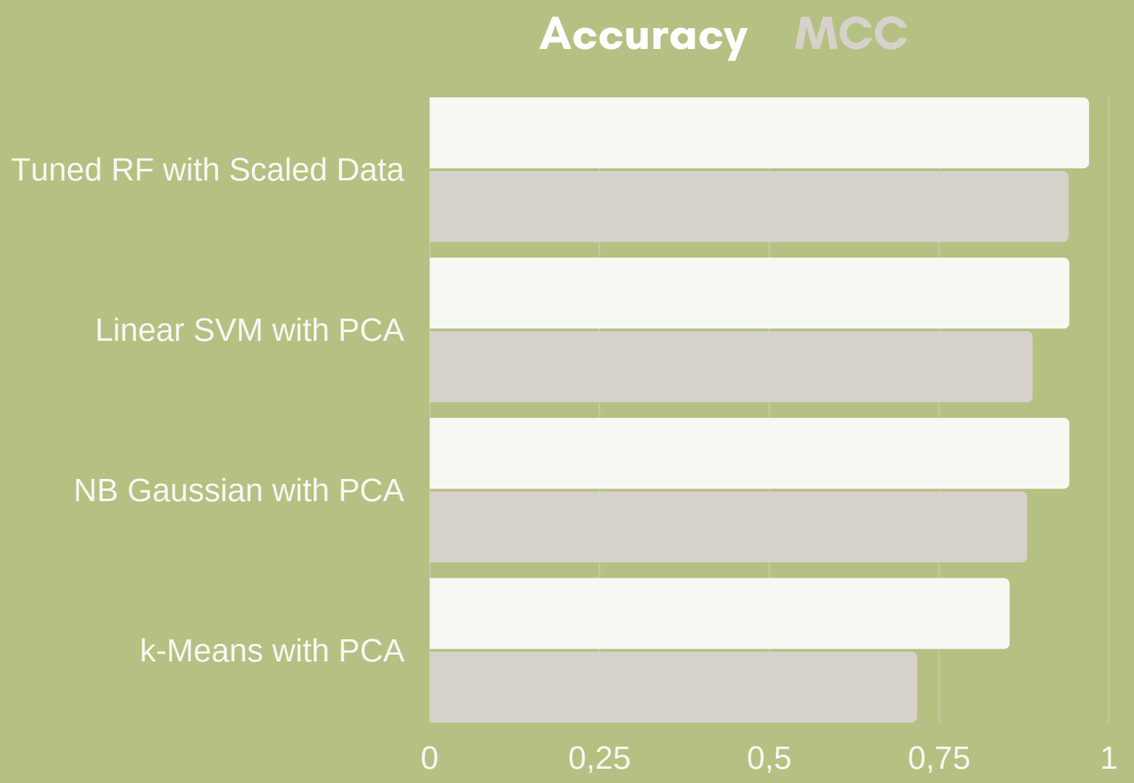
- num of estimators - **400**
- max features - **0.6**
- max depth - **3**
- min samples split - **12**
- min samples leaf - **2**
- bootstrap - **True**



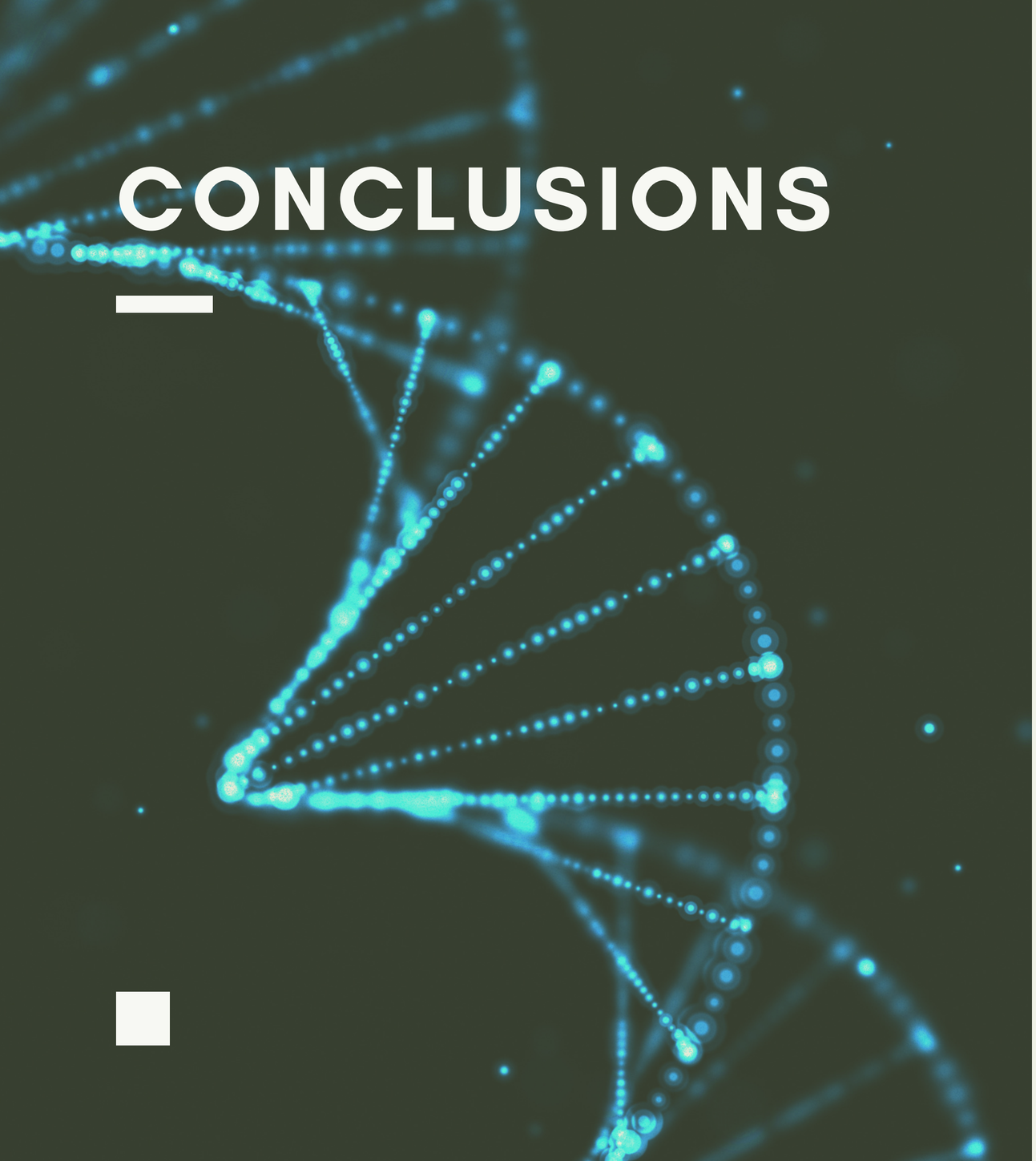
PERFORMANCES RECAP

THE MORE I KNOW, THE BETTER

Overall, the models behaved very well. Among those, Random Forest Classifier with a fine Hyperparameters Tuning on the scaled dataset has achieved the best accuracy and MCC score possible.



Best configuration for each proposed model



CONCLUSIONS

(VERY) BIG DATA

Genomics databases are truly highly data populated folders. For this reason, in further works, may be useful to exploit fine-tuned Random Forest Classifiers to surf across all these infos.

NOT SO STRAIGHTFORWARD

Certainly, against high dimensional data PCA is a very powerful weapon. However, it could be convenient to explore other paths, that could give the best by learning informations through a large set of features.

"IT'S GOOD BECAUSE IT'S DONE THIS WAY"

We grew up by a "trial and error" approach and for cancer-related topics we can't allow to us not to put much effort in avoiding mistakes. Maybe, we can discover new valuable models that aren't so trivial.

THANKS FOR YOUR ATTENTION!

VINCENZO MARCIANO', 282004