Cardiotocograms Dataset Analysis

Machine Learning Algorithms for Binary Classification

Candidate: Edoardo Fantolino s286008

Professors: Francesco Vaccarino,
Mauro Gasparini

Politecnico di Torino

Introduction

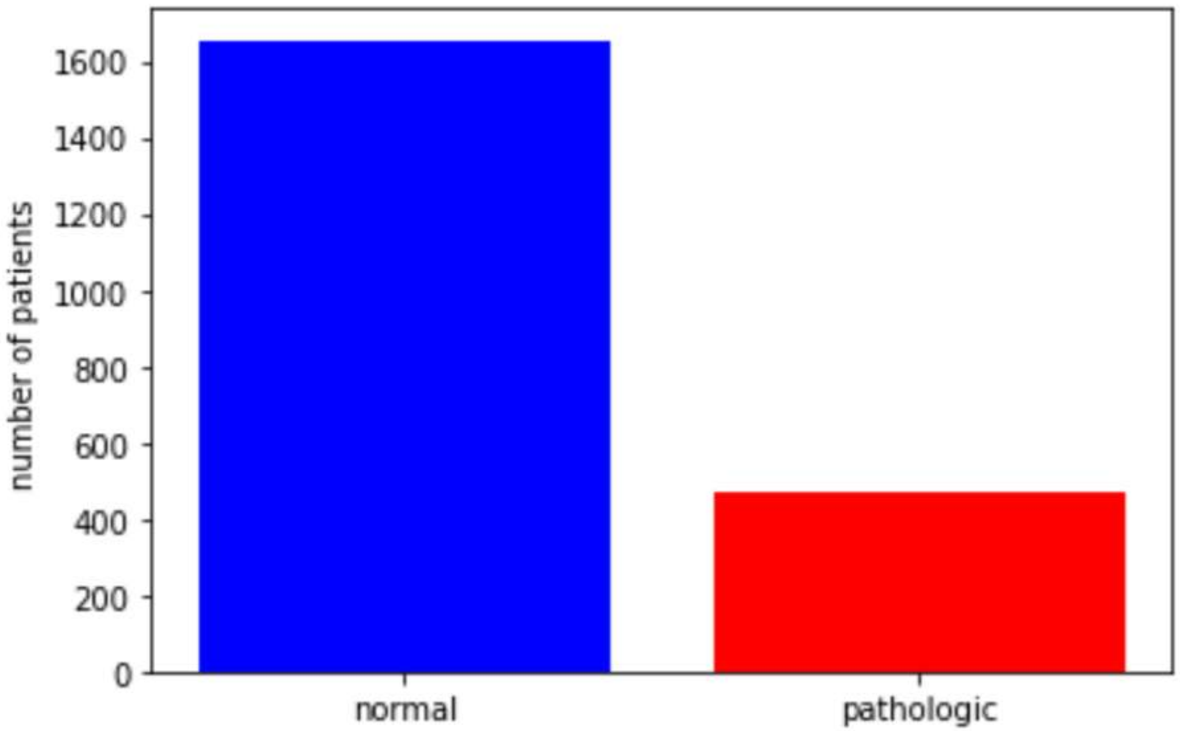
- The aim is to establish the **well being of a baby**.
- Parameters like **heart rates** of the child and the mother uterine **contractions**.
- External, non-invasive technique.

Explore different Machine Learning techniques to predict the conditions of the baby

Data Overview

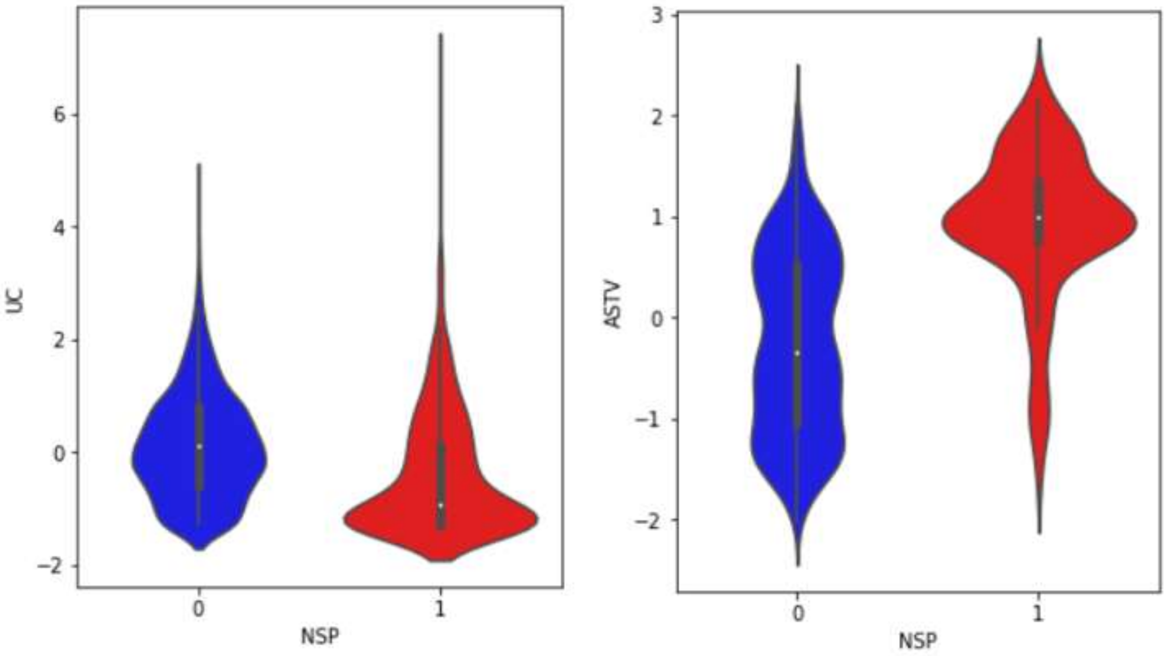
- Heart Beats (FHR) per minute
- Fetal Movement per second
- Uterine Contractions per second
- Minimum of FHR histogram
- Maximum of FHR histogram
- …

From Multiclass to Binary Classification

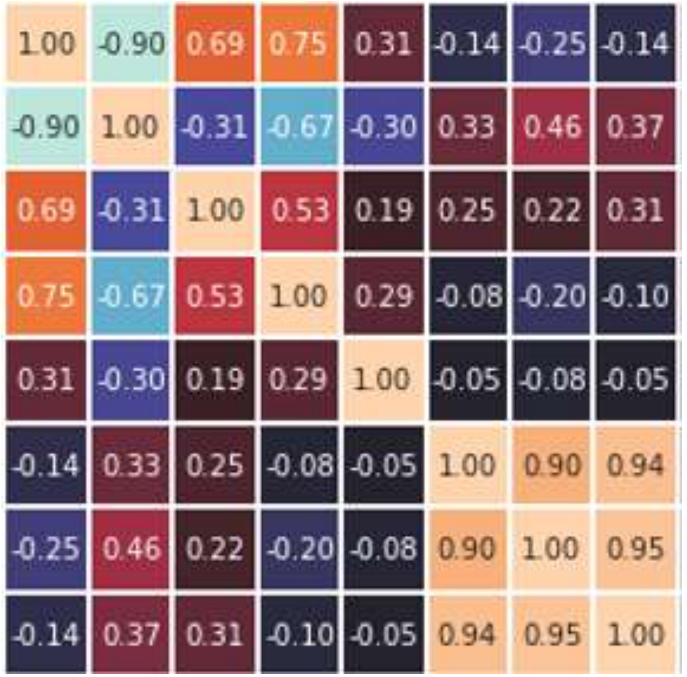


Data Exploration

Feature Distribution with Violin Plots



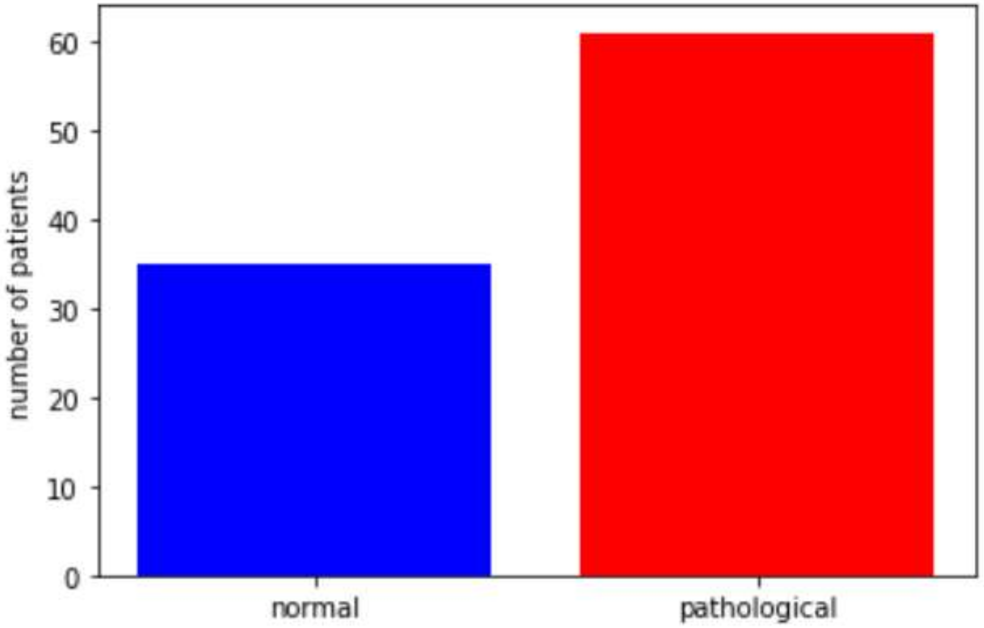
Feature Correlation with Pearson Coefficients

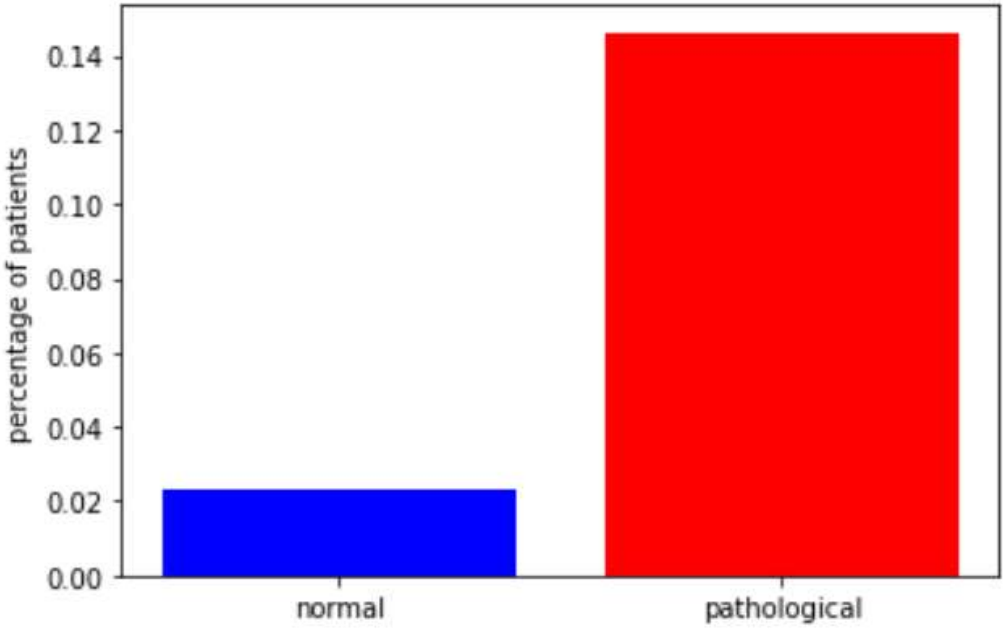


1.00	-0.90	0.69	0.75	0.31	-0.14	-0.25	-0.14
-0.90	1.00	-0.31	-0.67	-0.30	0.33	0.46	0.37
0.69	-0.31	1.00	0.53	0.19	0.25	0.22	0.31
0.75	-0.67	0.53	1.00	0.29	-0.08	-0.20	-0.10
0.31	-0.30	0.19	0.29	1.00	-0.05	-0.08	-0.05
-0.14	0.33	0.25	-0.08	-0.05	1.00	0.90	0.94
-0.25	0.46	0.22	-0.20	-0.08	0.90	1.00	0.95
-0.14	0.37	0.31	-0.10	-0.05	0.94	0.95	1.00

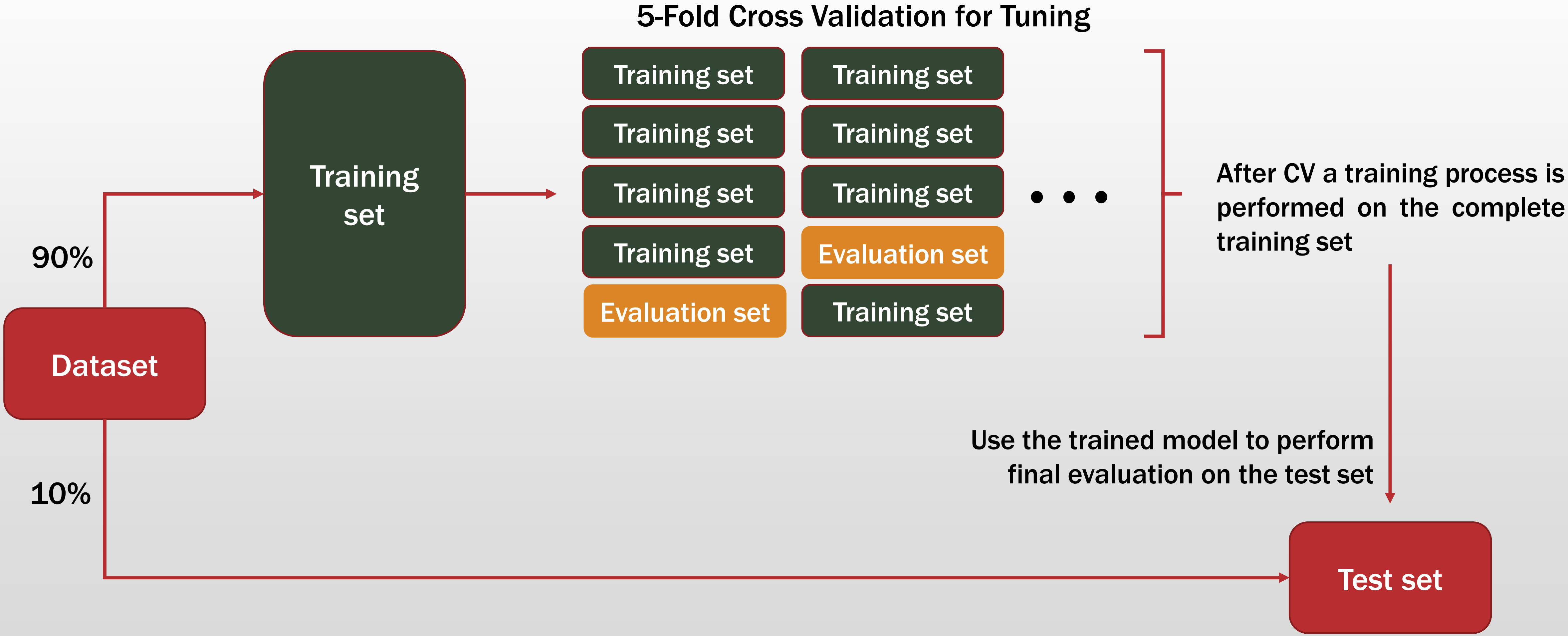
Outliers Detection: Isolation Forest

The power of Isolation Forest lies in the fact that anomalies are **easier to separate** and **few** w.r.t. inliers points





What happen to the dataset?
General Overview



After CV a training process is performed on the complete training set

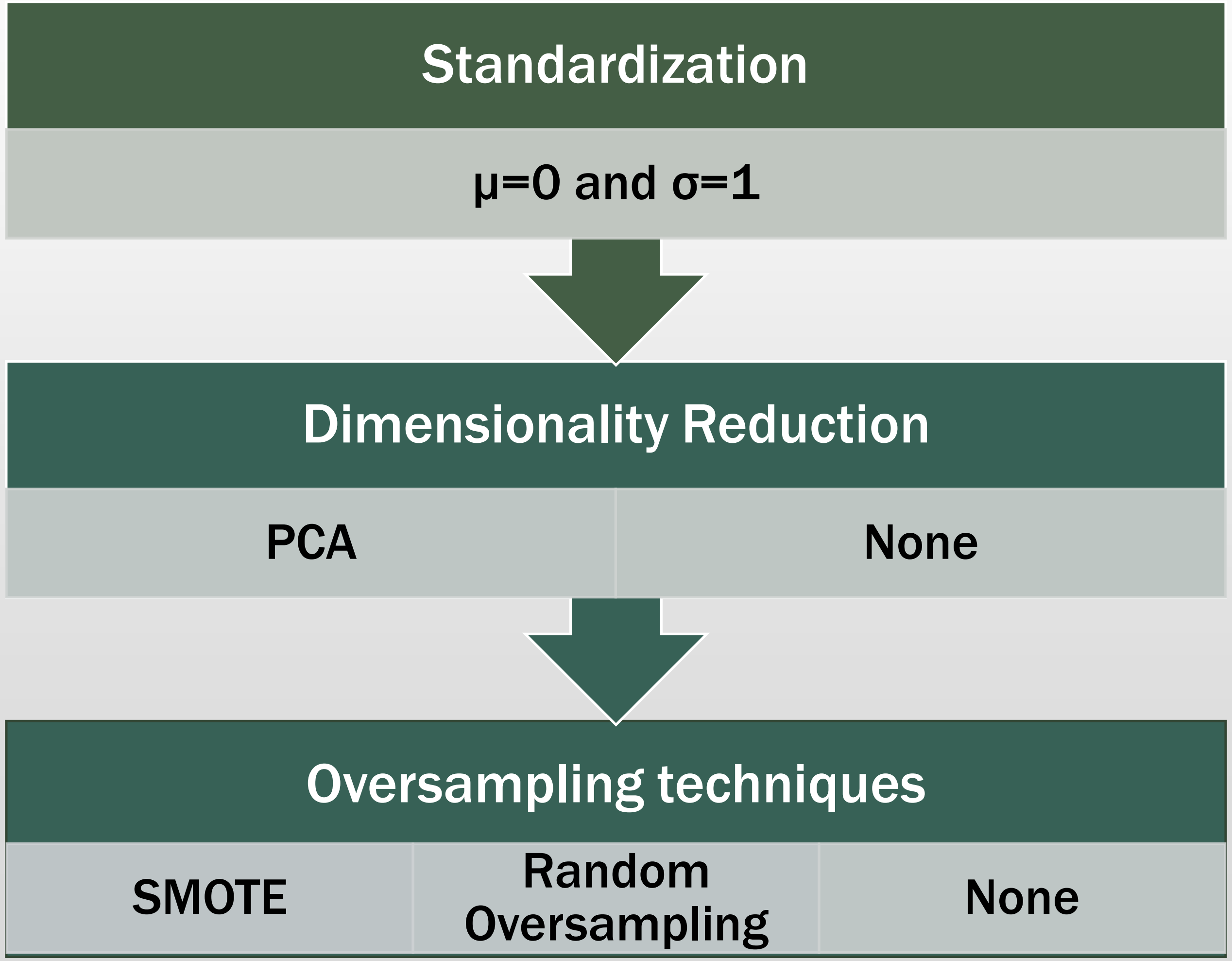
Use the trained model to perform final evaluation on the test set

Blindly feed the data to the algorithm? NO!
Data Preparation

Dare centered in 0 with a standard deviation equal to 1

Reduce the number of features of our dataset keeping as much as information as possible

Balance the dataset by adding samples to the minority class



Standardization

μ=0 and σ=1

Dimensionality Reduction

PCA | None

Oversampling techniques

SMOTE | Random Oversampling | None

Metrics

Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

F1 Score

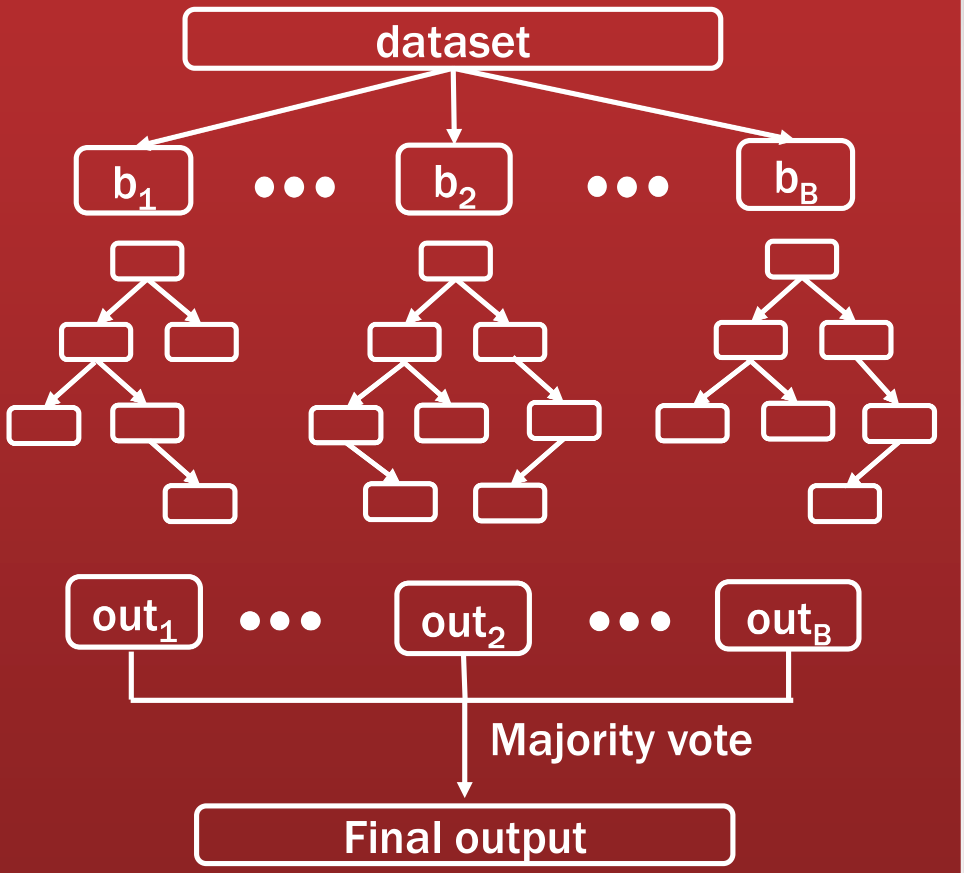
F1 Score is an evaluation metric that in the medical field is more reliable with respect to the most common metric: accuracy.

- To cure Trees overfitting the solution is searched in the “wisdom of the crowd”.

- Random Forest is a supervised ensemble method that exploit bootstrapped aggregation of the training set.

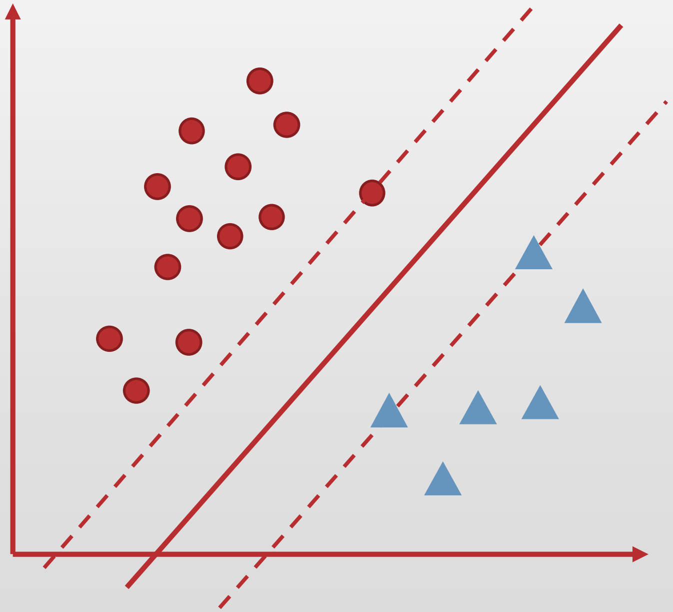
- The performances are enhanced using feature bagging.

Random Forest



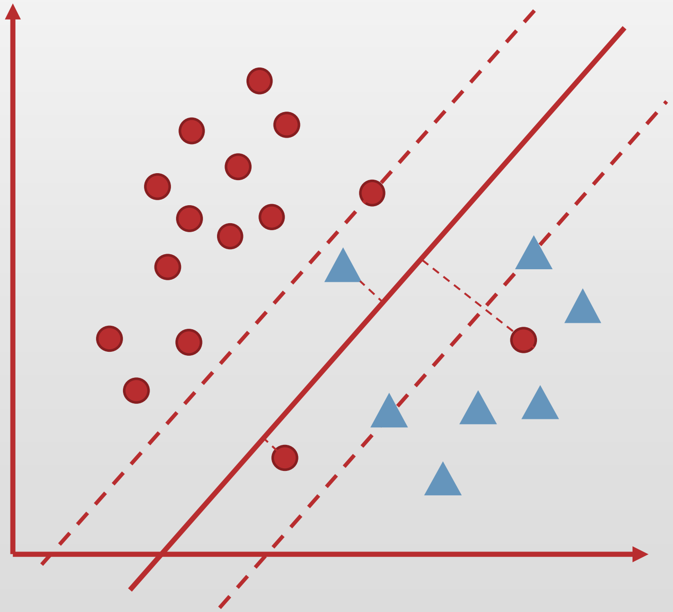
Support Vector Machine

Hard-SVM



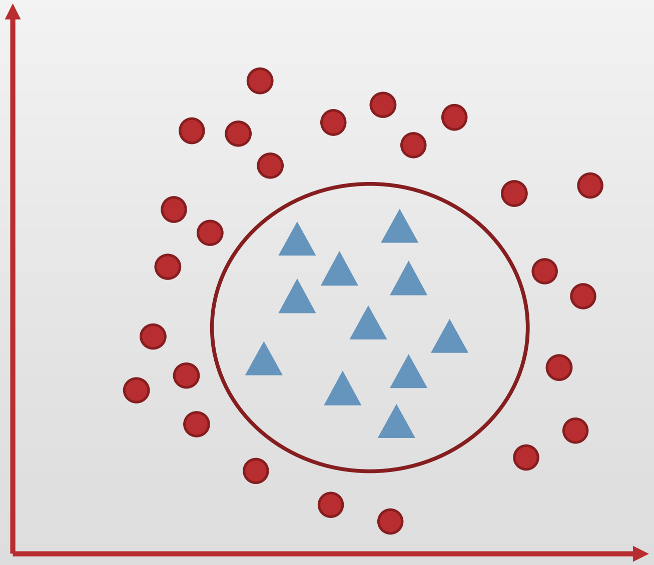
Linear decision boundary maximizing the margin

Soft-SVM



Trade off between margin maximization and classification error

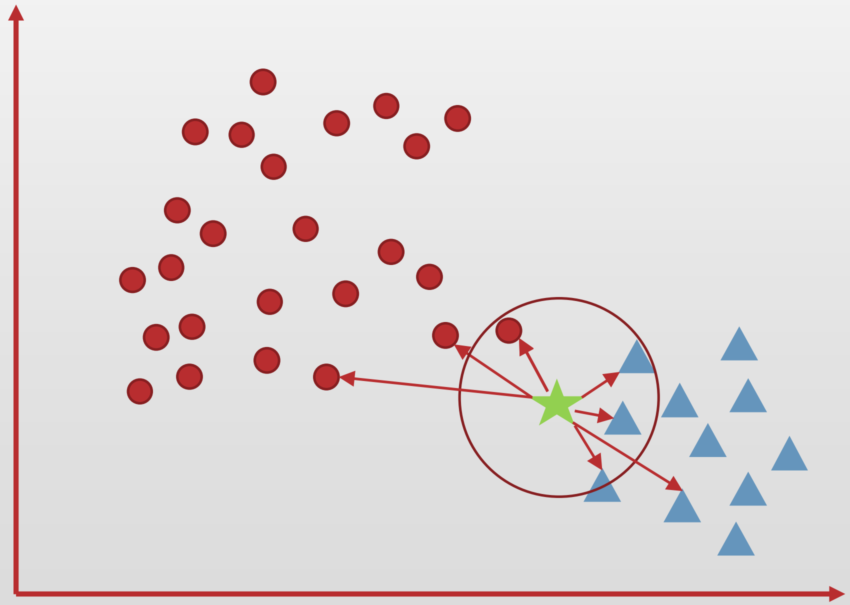
Kernel-SVM



Making the non-linear linear exploiting the Kernel functions

K-Nearest Neighbors

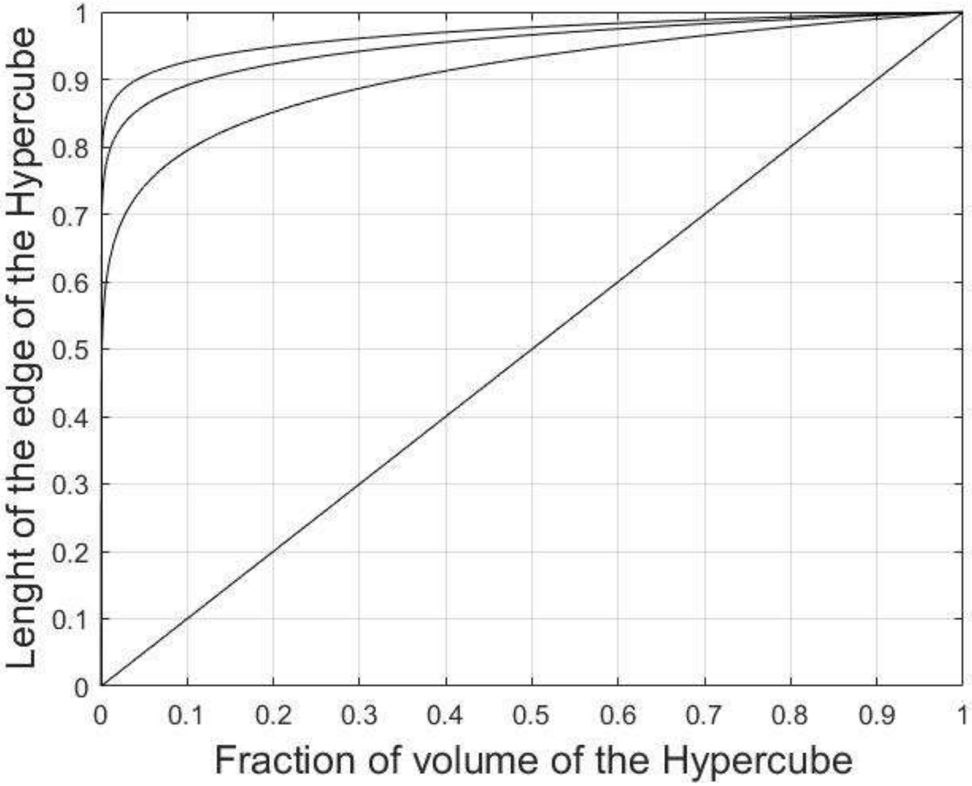
- Simplest family of algorithm in the Machine Learning field
- Based on storing information about Training Set
- The class of a new instance is decided looking at the neighbors

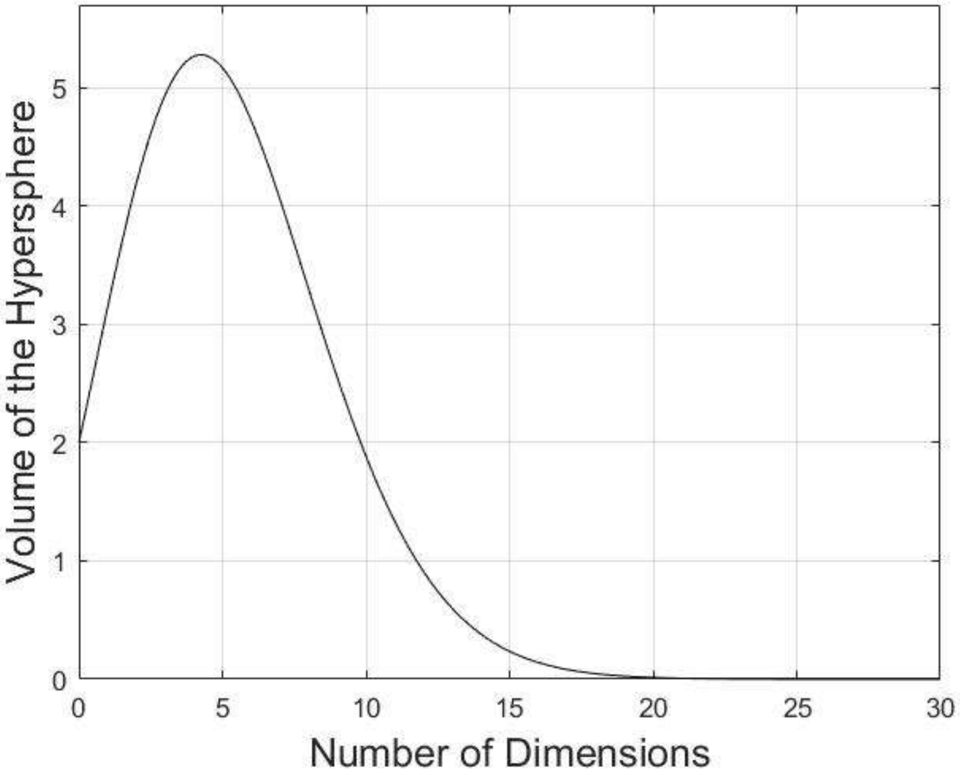


Summary of the results

Model	Pipeline	F1 Score
Kernel SVM	Normalized	91.74
Random Forest	Normalized + SMOTE	89.30
K-Nearest Neighbors	Normalized + PCA + SMOTE	82.35
Soft SVM	Normalized + SMOTE	76.42
Hard SVM	Normalized +SMOTE	74.80

Why PCA is convenient for KNN and not for SMV? Curse of Dimensionality





Conclusion

- Manage imbalance dataset
- Use dimensionality reduction
- Evaluate performances of ML algorithms
- Exploit theory results

It is strongly suggested to use the results of this algorithms as suggestion and not as final verdict. You must integrate the results with the domain knowledge of experts and other sources of documentation regarding the patients.