HEARTBEAT CLASSIFICATION

GUIDO GAGLIARDI

MATTEO DESSI'

HEARTBEAT CLASSIFICATION

The aim of this project is to use Machine Learning, using Python, to build several models that will be able to predict the presence of an arythmia from an ECG signal.

- The ML models will classify each beat constituting an ECG signal into one of the four most important macro-classis of arythmia: N, VEB, SVEB and F.
- 5 different models were trained and tested and their performance compared
- The data employed belongs entirely to the **MIT-BH database**

DATASET

EXPLANATION OF THE MIT-BIH DB

DATASET:
MIT-BIH DATABASE

The **MIT-BIH Arrhythmia Database** contains **48 half-hour excerpts of two-channel ambulatory ECG recordings**, obtained from **47 subjects** studied by the BIH Arrhythmia Laboratory between 1975 and 1979.

Each ECG record contains a variable number of beats, these usually vary patient by patient, and are refered to one txt file containing the information about the beats.

Each beat has two files associated to it:

- Txt file, containing in particular the position of the peak and the class to which the beat has been labeled

- CSV file, formed by two columns:
  - Value of the MLII (modified-lead II)
  - Value of one of V1, V2, V4 or V5

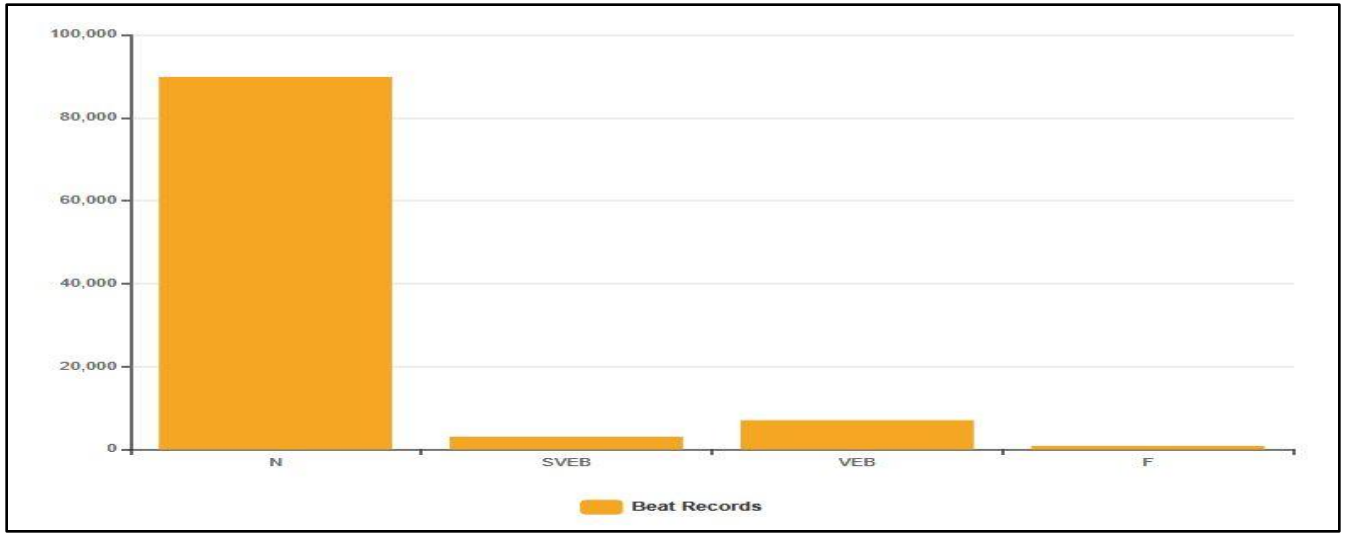
DATASET: CLASS DIAGRAM

Each beat is sampled from the database using the information of its peak and taking a certain number of values from the ECG wave by the left and the right of the peak.

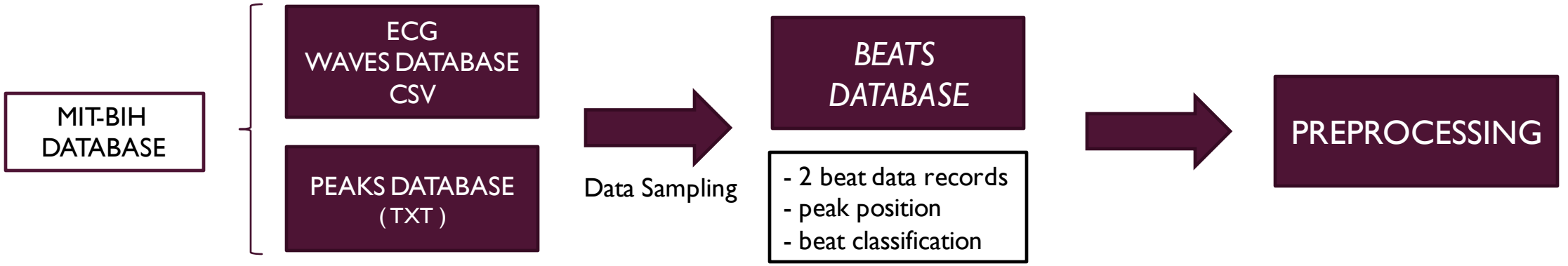
Each beat it is sampled in both of the signal stored in the CSV columns (MLII and V1) and stored again in 2 vector of 360 values representing the y values of the ECG wave.

The dataset is highly imbalanced:

- 90% of the beats belong to class N
- 3% belong to class SVEB
- 6% belong to class VEB
- 1% belong to class F
- The Q AAMI classes have been ignored given their basically non-existance



PREPROCESSING



Some preprocessing has been performed on the dataset before the training, in order to remove noise and clean data.

The procedure consists in performing a baseline removal from both of the 2 beat records:

- Two median filters of 200 ms and 600 ms are applied to obtain a baseline
- This baseline is substracted to the original signal, obtaining the baseline corrected ecg signal

Furthemore a normalization step of the data has been provided before the classification step.

OVERSAMPLING

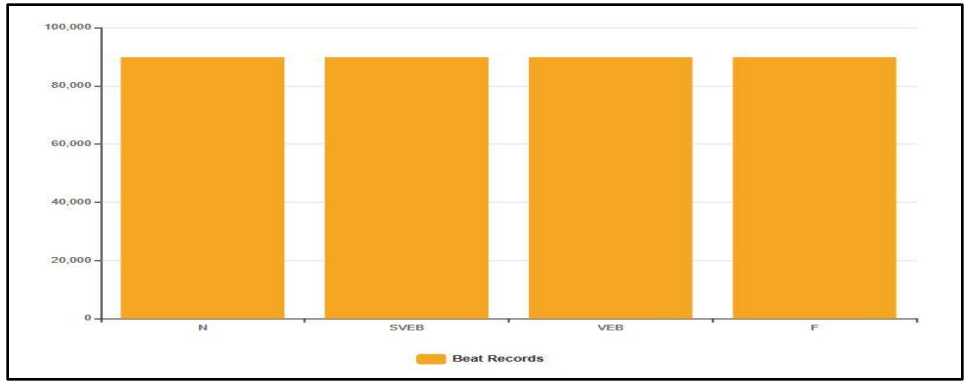
Given the high unbalance of the database, a step of oversampling for the minority classes is necessary

PREPROCESSING → OVERSAMPLING

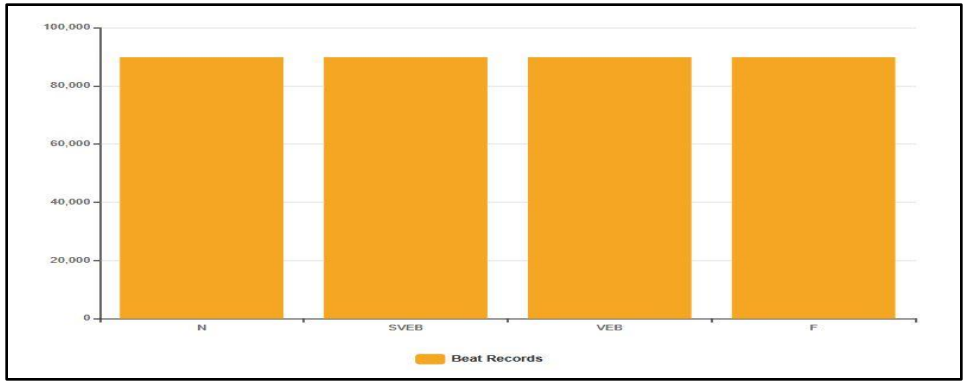
Two metods:

- Artifical Oversample: Taking a point from the segment uniting a point in the dataset and one of its k-neighbors
- Random Oversample: Picking samples at random with replacement

Artificial Oversample



Random Oversample



CLASSIFICATION

MODELS TRAINING USING BEAT AS DATA INPUT

CLASSIFICATION: MODELS

5 Classification Models are tested for the Arythmia Classification Task

- CART - (C4.5 implementation in python scikit-learn)
- Naive Bayesian
- K-NN
- SVM
- RandomForest

Since that each beat has 2 rapresentation sampled by different ECG waves (MLII and V1) in a first step each kind of beat is used to train a different classifier and their performances compared.

In a second step both of the beats type are used togheter to form a unique data mining flow in order to improve the performances of the clasifiers.

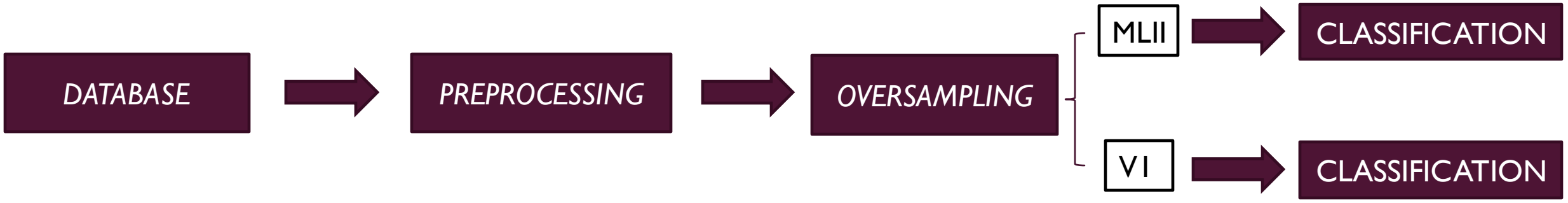
CLASSIFICATION: 6 FOLD VALIDATION

In order to provide relevant data for the classification task a k-fold validation has been performed.

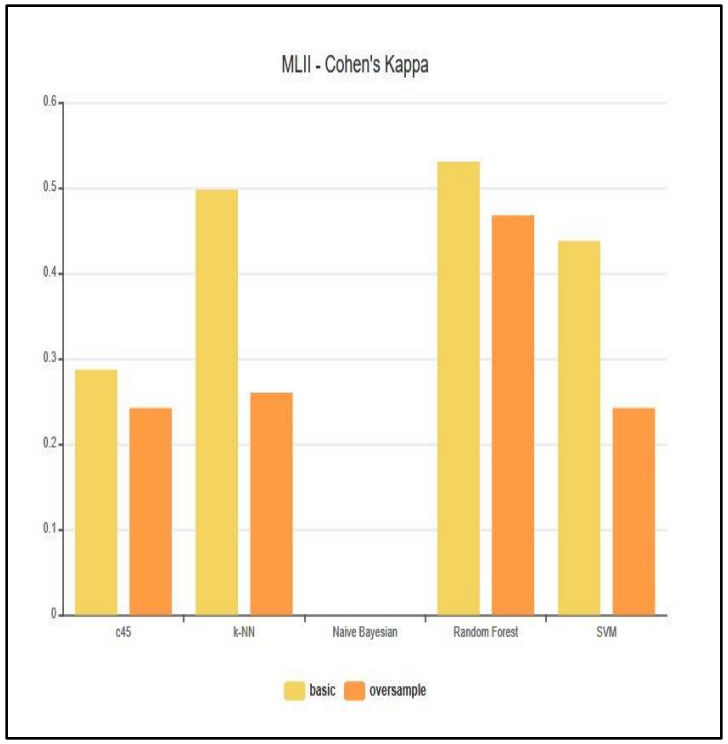
It is decided to use 6-fold while each fold is a patient ECG record (with both of MLII and V1).

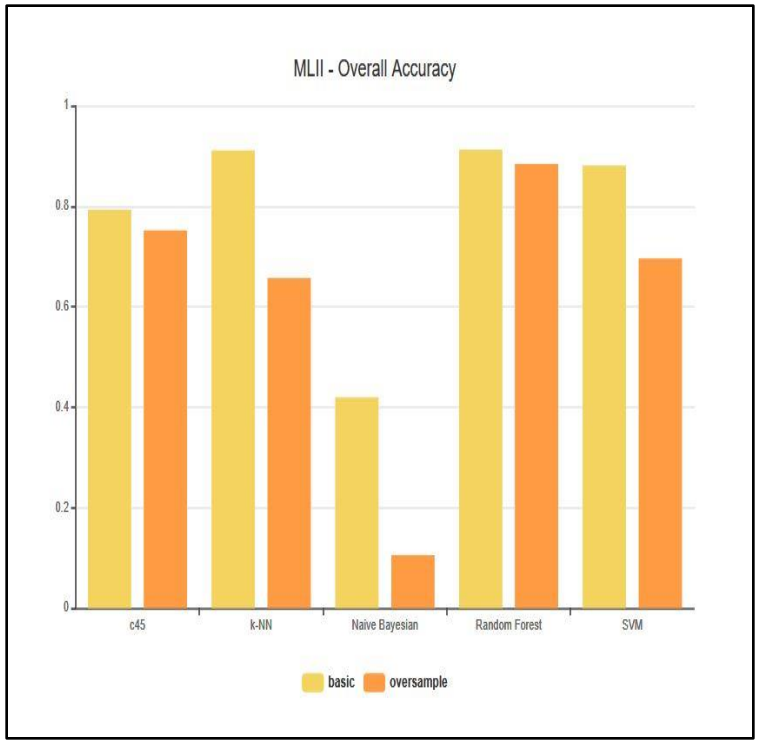
Each of the step in the dataming flow is wrapped in the 6 fold classification method, from the normalization of the data (the scaler is fitted with the training folds and then used on the test fold), to the classification.

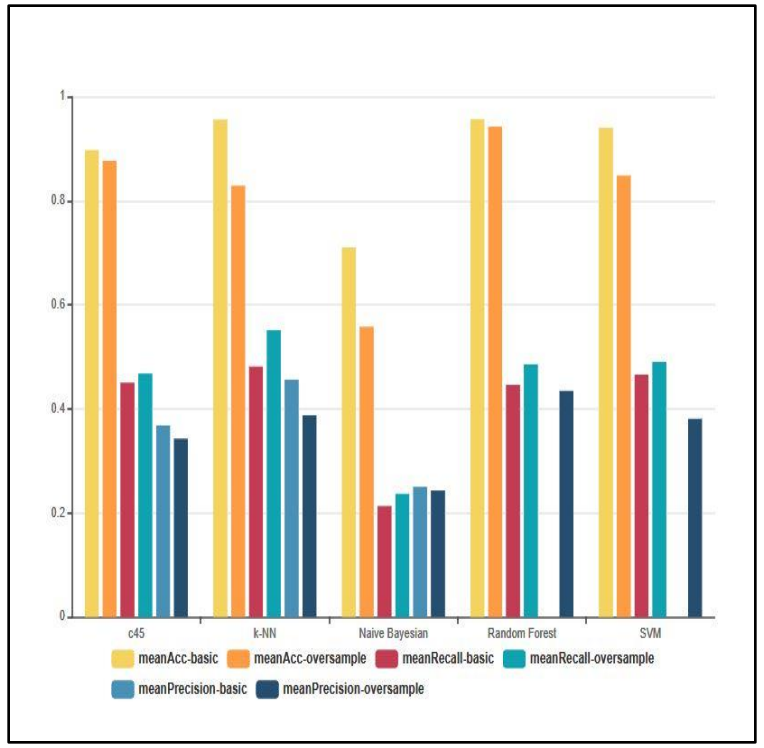
CLASSIFICATION: DATA MINING FLOW



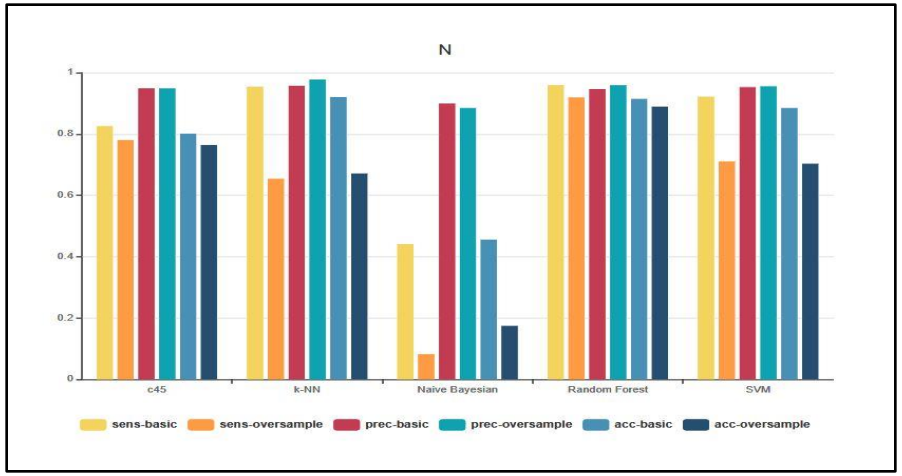
CLASSIFICATION:
I - MODELS EVALUATION WITH BEAT AS DATA - MLII

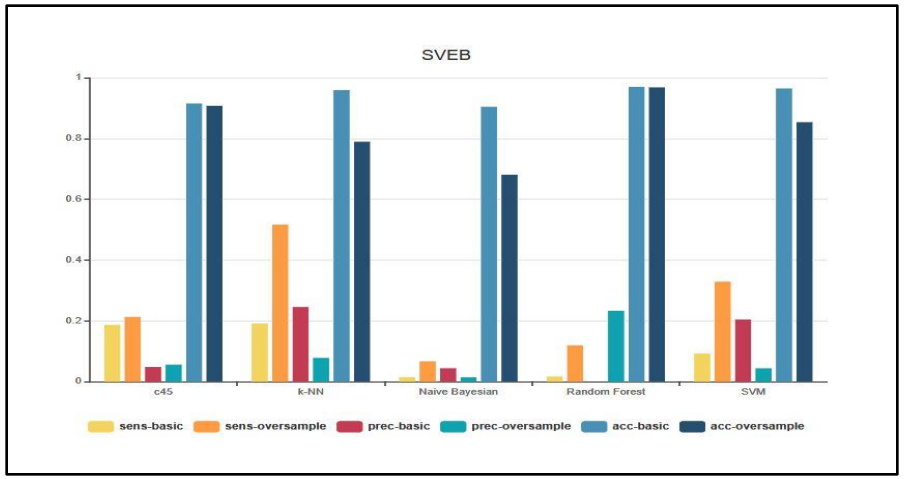


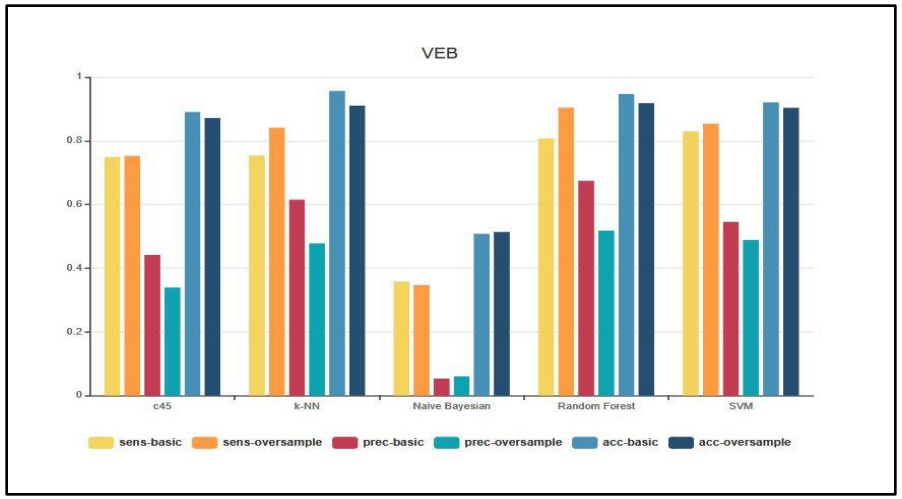


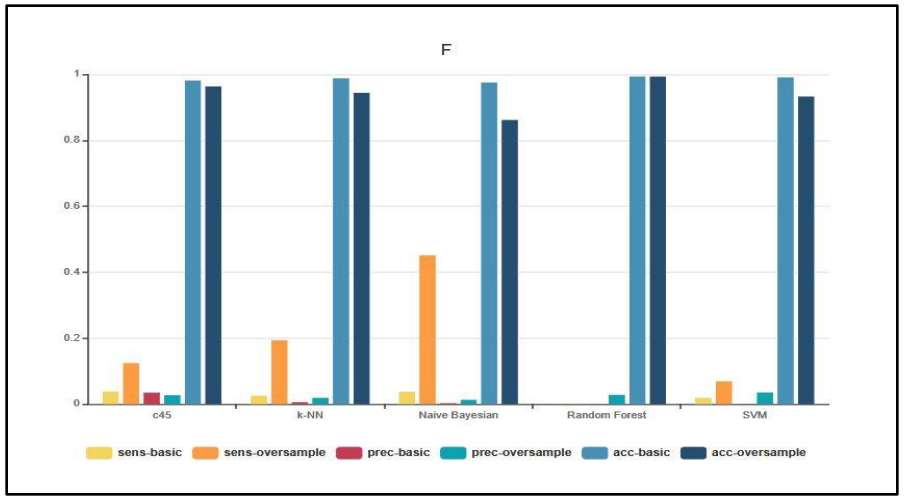


CLASSIFICATION:
I - MODELS EVALUATION WITH BEAT AS DATA - MLII

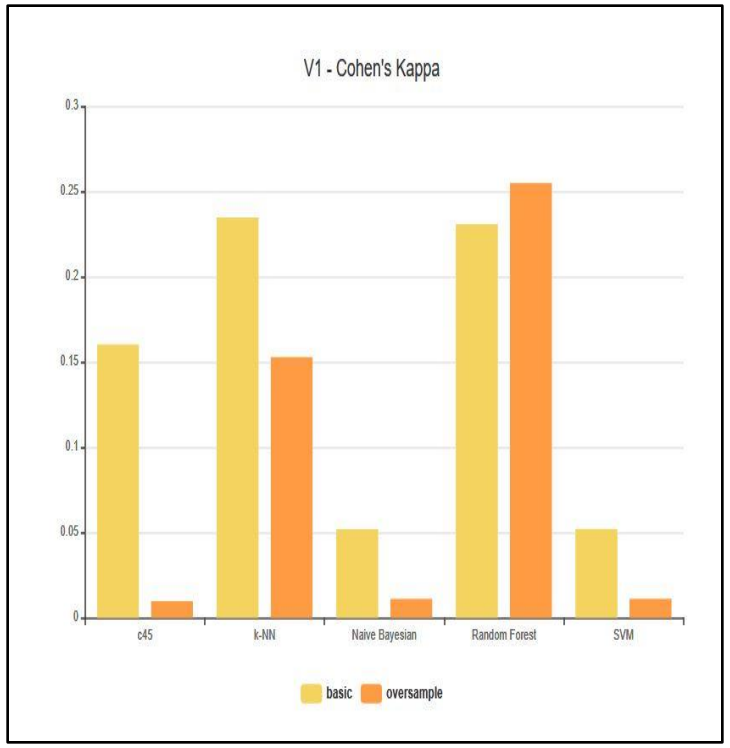


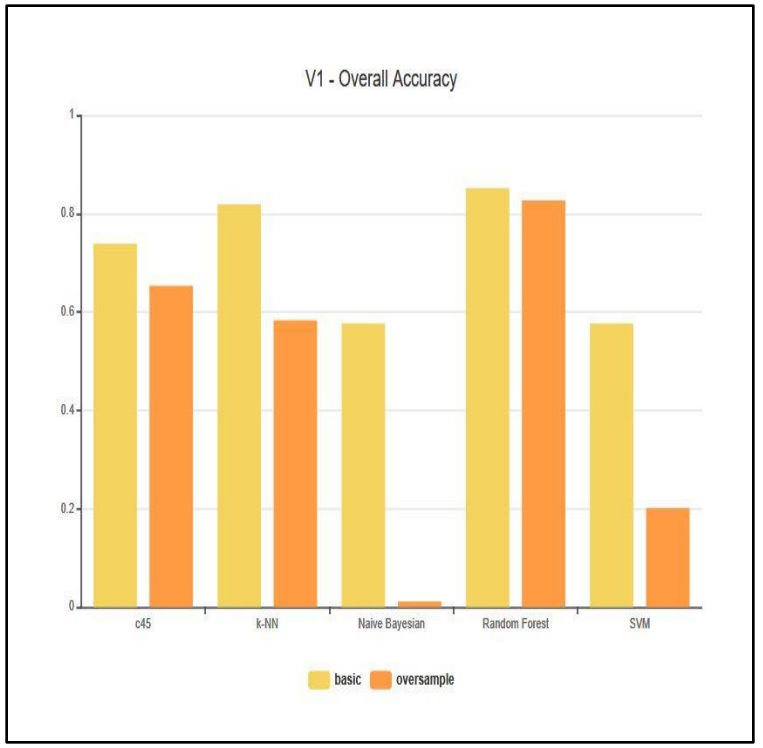


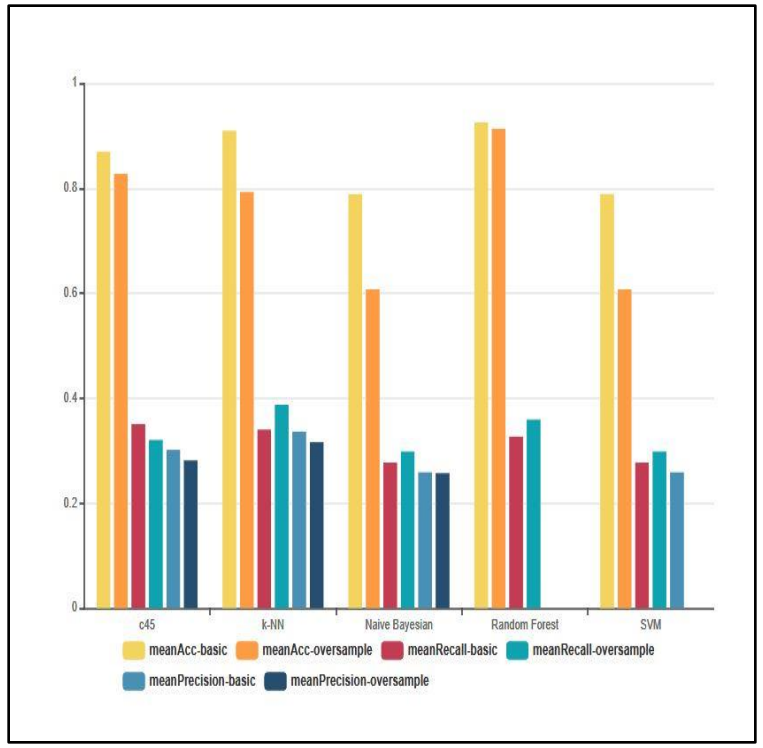




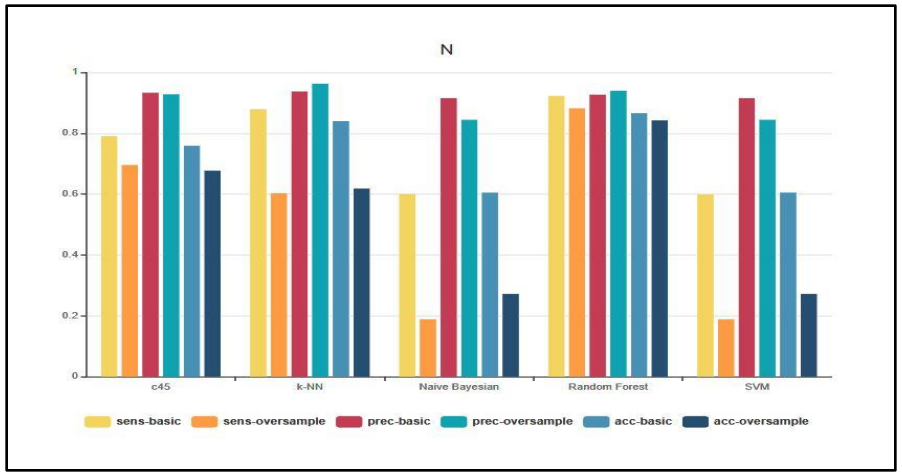
CLASSIFICATION:
I - MODELS EVALUATION WITH BEAT AS DATA - V1

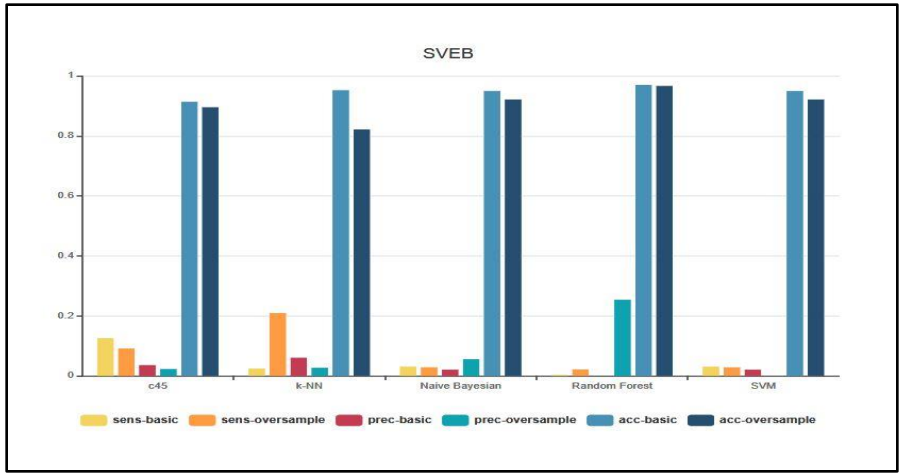


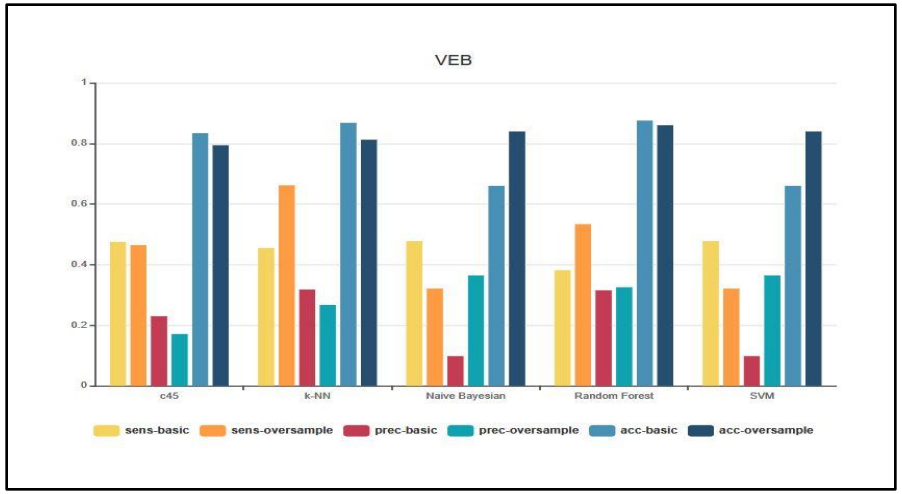


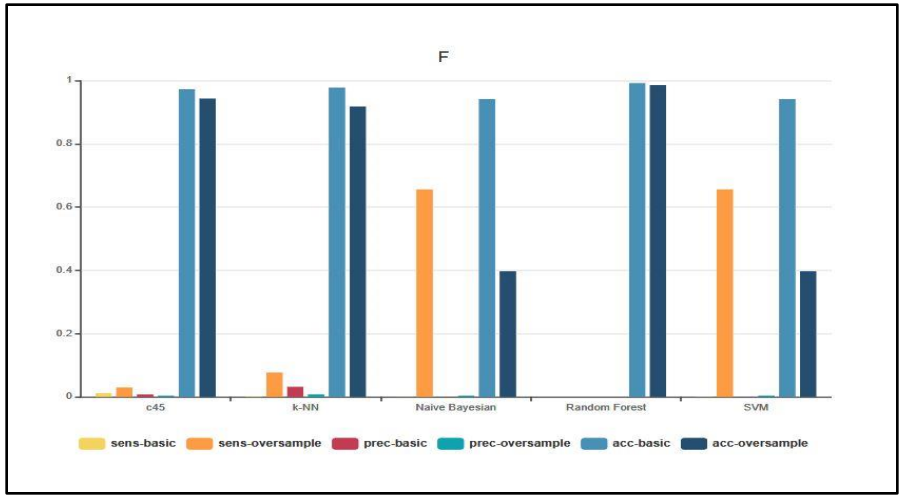


CLASSIFICATION:
I - MODELS EVALUATION WITH BEAT AS DATA - V1









CLASSIFICATION: VOTING MODEL

In the second step we try to improve the performance of the classifications merging the previus steps.

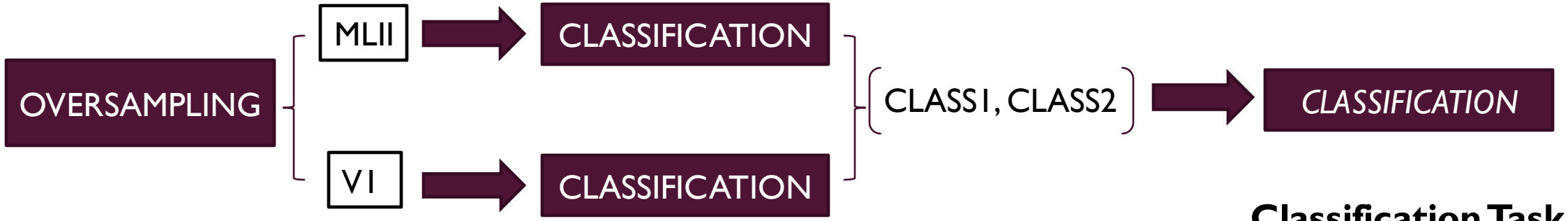
Two kinds of approaches could be possible at this point:

- Merge the two beats data and train a single model
- Train two different models with the 2 kinds of beats and then perform a voting strategies on the results.

In this study the second option is choosed because each beat is represented as a vector of 360 values, so merging 2 beats imply training classifiers with vectors of 720 values as objects. The computational cost would be too high to handle and the traning phase too slow.

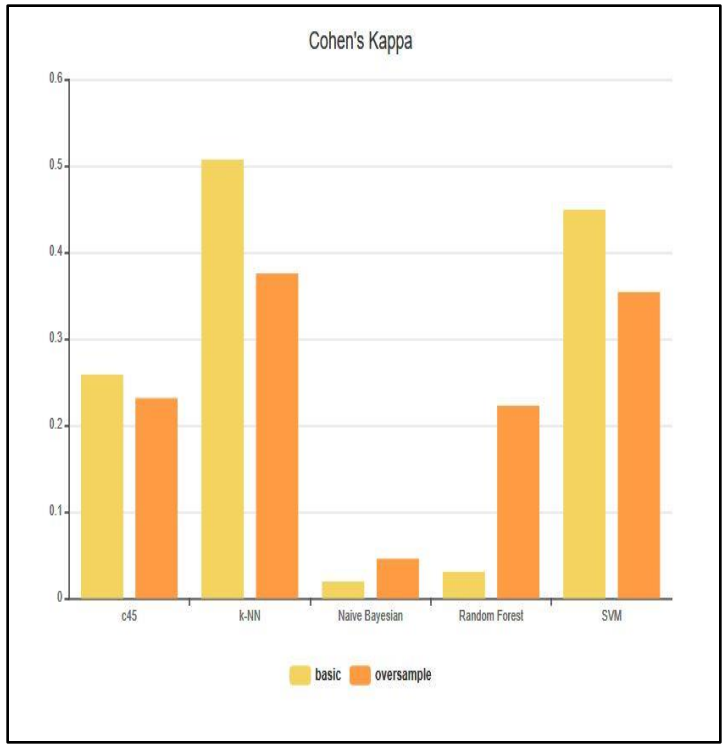
Each classification has been performed on a single beat record: so 2 models for classificator have been evaluated in order to have 2 distinct classifications for beat.

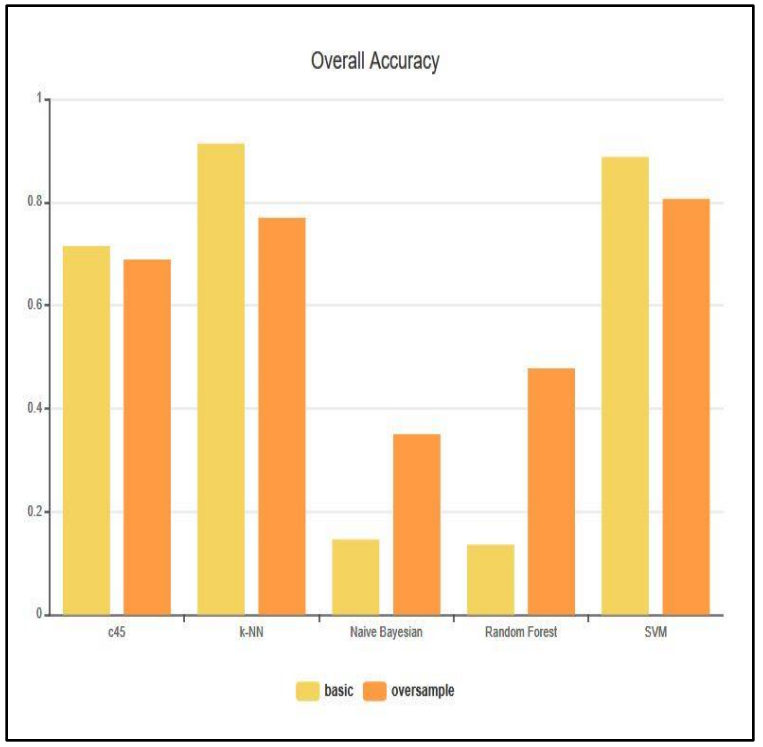
Then a voting step it is performed using a RandomForest decision tree on the classificators results.

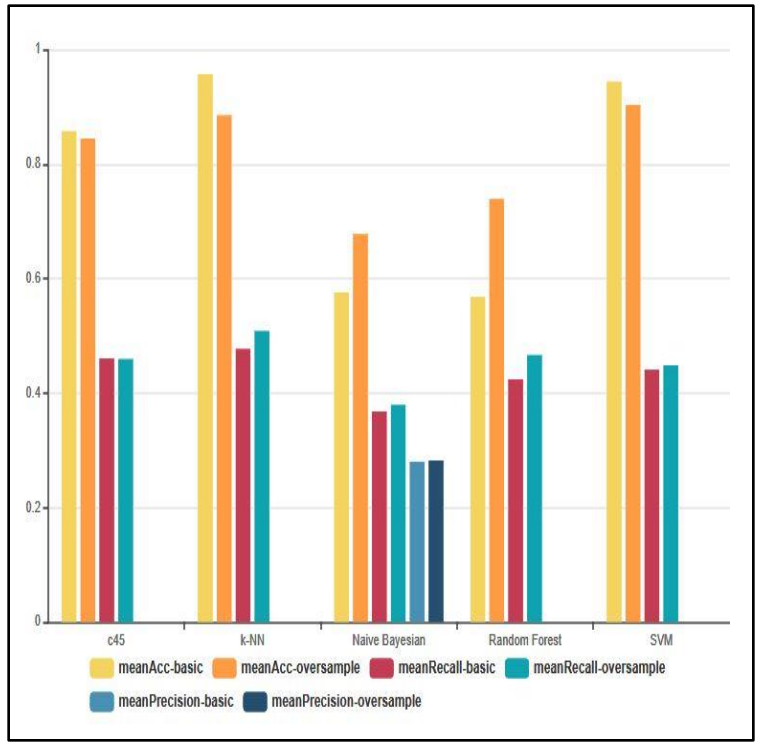


Classification Task

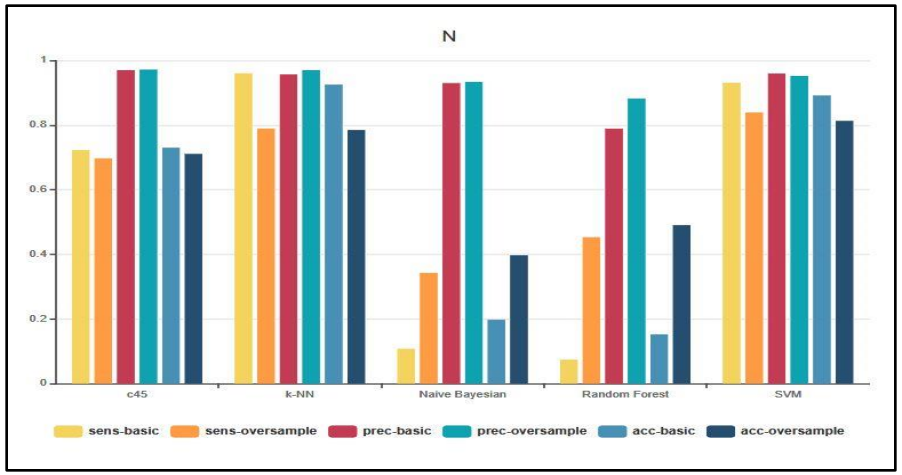
CLASSIFICATION:
II - MODELS EVALUATION WITH BEAT AS DATA (1/2)

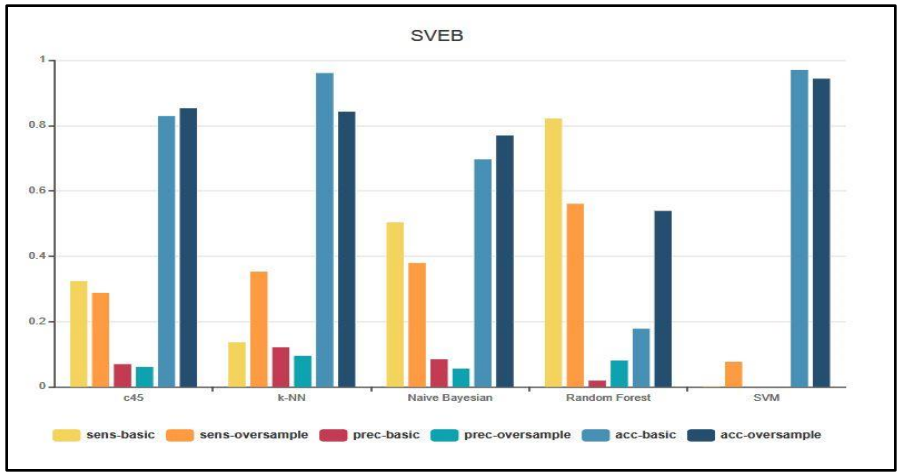


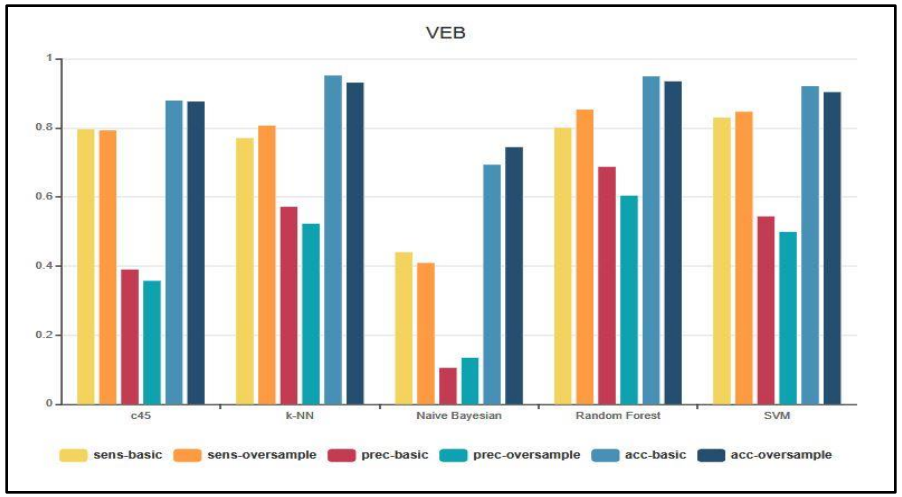


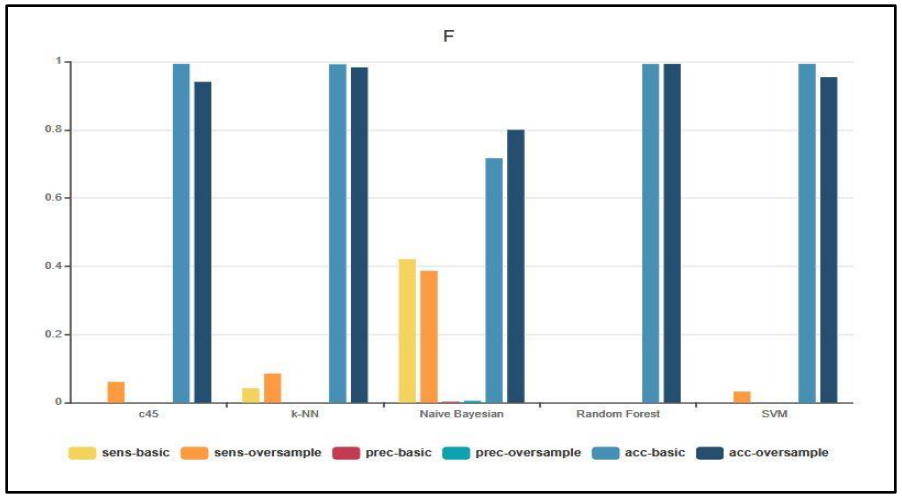


CLASSIFICATOIN:
II - MODELS EVALUATION WITH BEAT AS DATA (2/2)









FEATURES SELECTION

EXTRACTION, SELECTION AND REDUCTION OF THE FEATURES FROM BEATS DATA

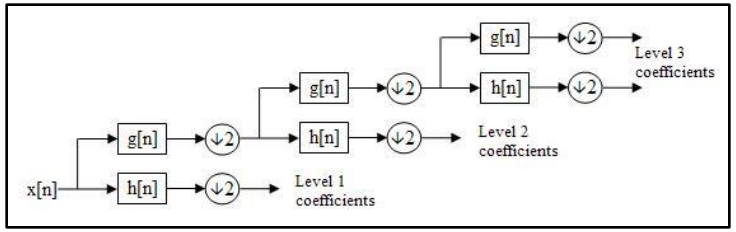
FEATURE SELECTION: EXTRACTION

In order to improve the classification results, a features selection step is provided, to avoid the curse of dimensionality and to send more significant data to the classifiers.

Both the MLII and V1 data were used to construct each features, so the previus voting strategies has been removed.

Wavelets:

The coefficient of detail and aproximation in the level 3 of the decomposition, related to the 0-45 Hz frequency where most of the energy of the signal is located



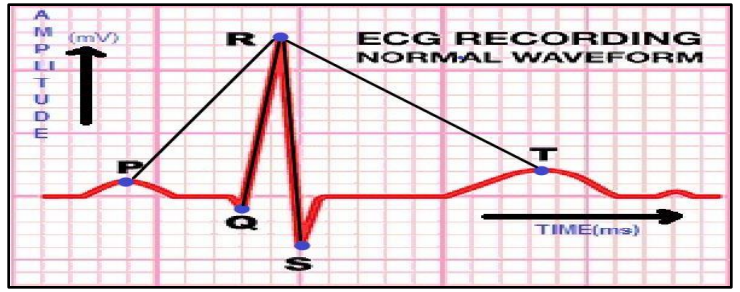
HOS:

The union of the kurtosis and skewness value of each of the 5 intervals dividing a beat

Morphological Descriptor:

Euclidean distance between the R-peak and this four points in each beat:

- Max (beat [0,40])
- Min (beat [75,85])
- Min (beat [95,105])
- Max (beat [150,180])



FEATURE SELECTION: EXTRACTION

Hbf5:

Coefficients, computed on each beat, of the Hermite series of degree 3, 4 and 5 that are the lest fit square

RR intervals:

A descriptor formed by 4 values:

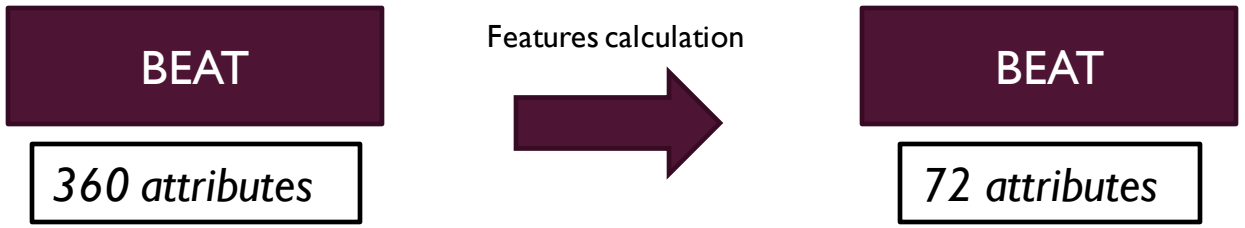
- Pre-RR: distance between the actual beat and the previous one
- Post-RR: distance between the actual beat and the next one
- Local-RR: average of the 10 previous Pre-RR values
- Global-RR: average of the previous Pre-RR values produced in the last 20 minutes

LBP:

For each data point in a beat, a binary code is produced by the comparison of its value with the value of its neighbours. Then, an histogram that contains the frequency of each binary pattern is built

FEATURE SELECTION: EXTRACTION

For each beat there are calculate a number of **72 values** that represent the sum of all the values calculated for each features.
So **from 360 attributes for beat, with the features calculation it is passed to 72 values**.



A feature selection phase is used in order to decrease again the number of data necessary to do the classification.
Two features selection method are used:

- **Info-gain**
- **F-value**

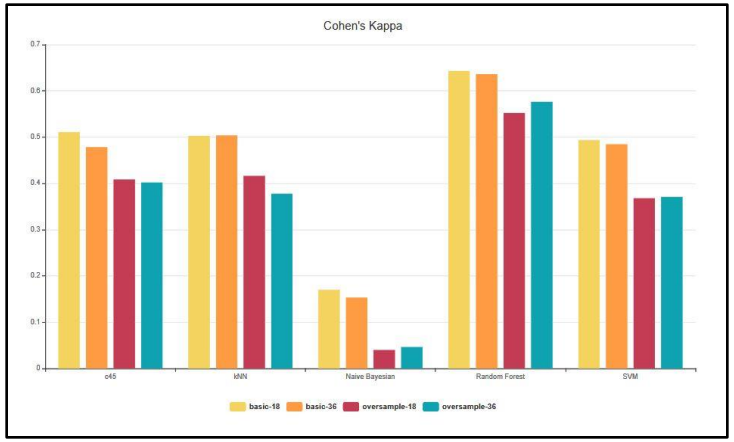
For each method a classification model is trained, using each time different number of features in order to find the best solution.

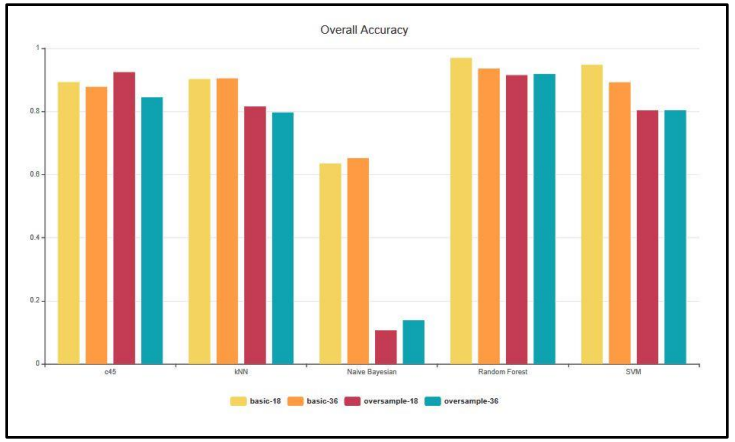
CLASSIFICATION

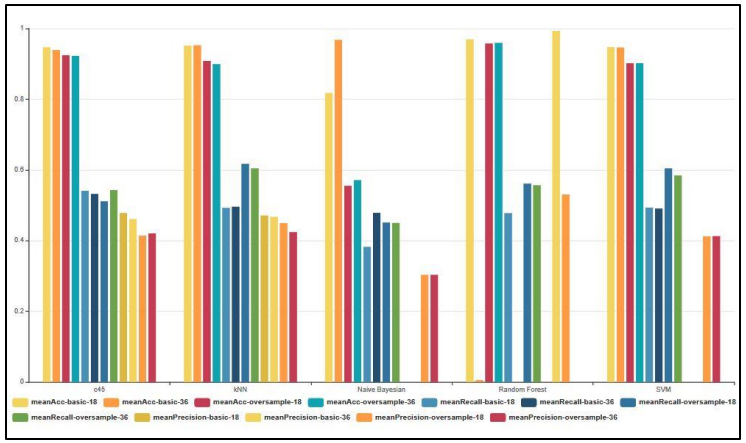
TRAIN MODELS WITH FEATURES AS INPUT DATA

CLASSIFICATION: DATA MINING FLOW

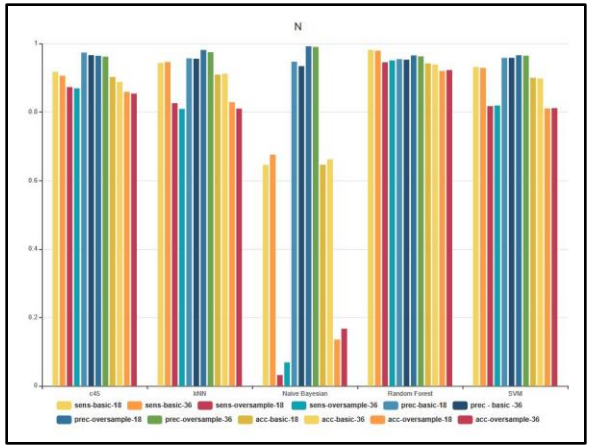
III - MODELS EVALUATION WITH FEATURES (1/2) - INFO GAIN

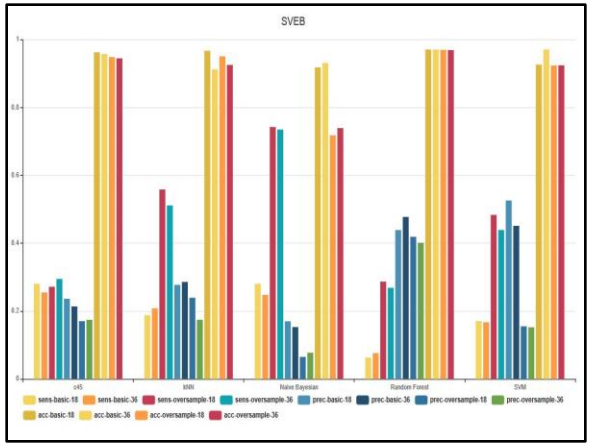


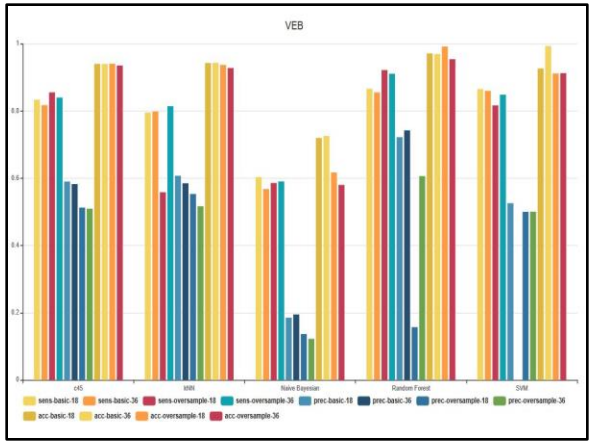


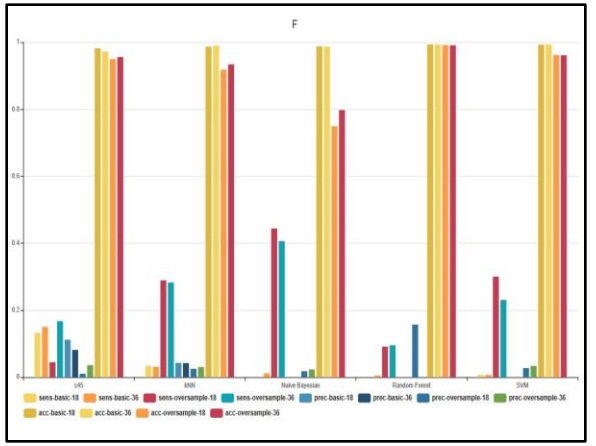


III - MODELS EVALUATION WITH FEATURES (2/2) – INFO GAIN

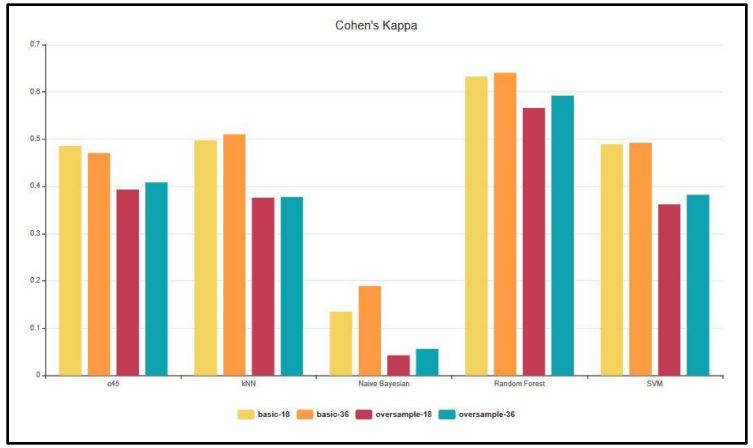


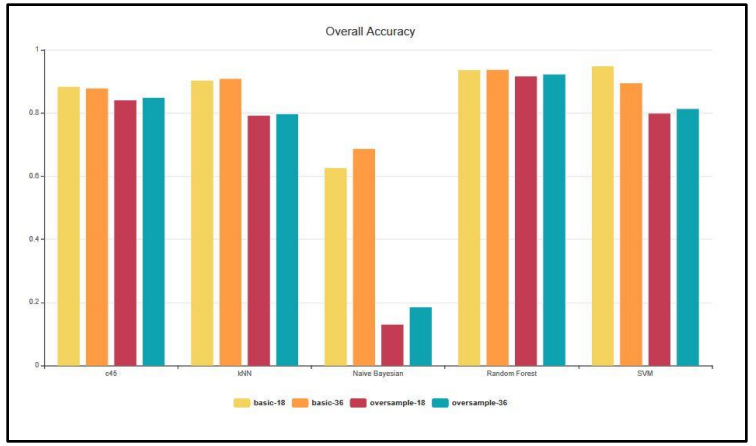


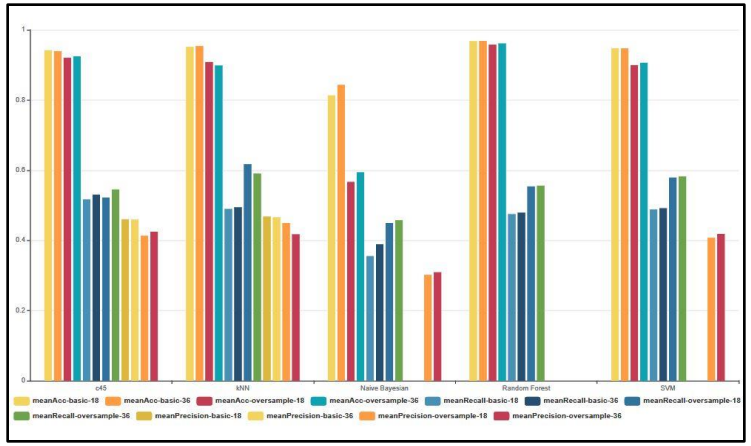




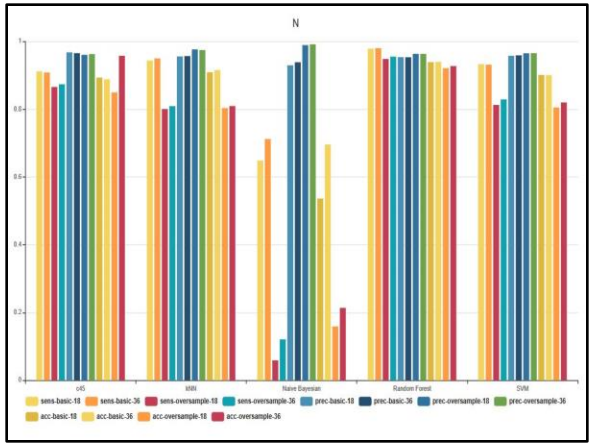
III - MODELS EVALUATION WITH FEATURES (1/2) - F VALUE

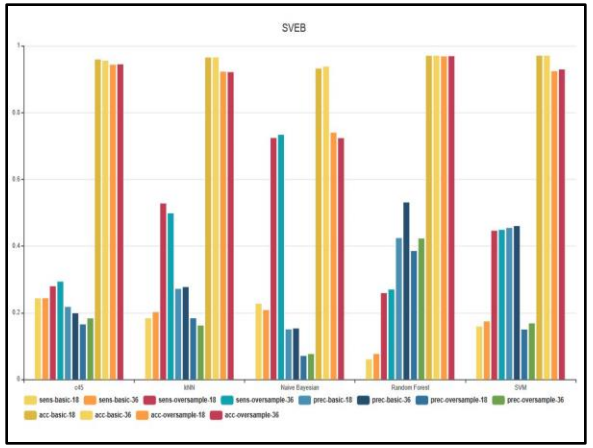


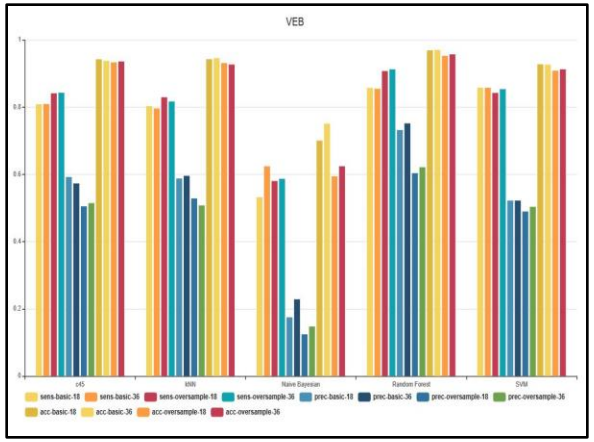


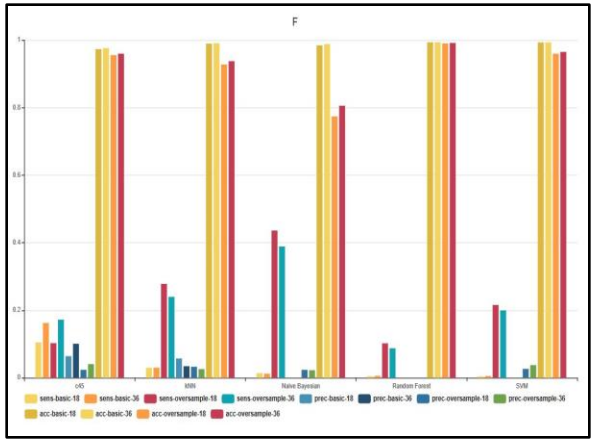


III - MODELS EVALUATION WITH FEATURES (2/2) – F VALUE









CLASSIFIERS EVALUATION

CLASSIFIERS EVALUATION

Based on the results provided by the preavious section the **two classifiers** that has greater results are **k-NN** and **SVM** using **info-gain** as features selection method with **18 features**.

Since that the database is highly unbalanced, we have considered the recall and precision score of each minority-class as parameter, because all the classifiers have good results with the classification of beats that belongs to class N.

In the next step the best classifier is evaluated using a t-test with the F2-score as parameter.

CLASSIFIERS EVALUATION: STUDENT'S T-TEST

F2 score table

	SVM	k-NN
1-fold	0,4067	0,3382
2-fold	0,3230	0,2608
3-fold	0,5596	0,57029
4-fold	0,5729	0,5854
5-fold	0,4623	0,5103
6-fold	0,5268	0,6276
mean	0,4752	0,4821
var	0,001	0,022

- P-value = 0,8054
- T = 0,1504
- T(alpha=0.5/2, d=5) = 2,571

So we can't conclude that the differences between SVM and k-NN are statistically significant.

Moreover k-NN is slower than SVM in predicting the results, for our pourpose we decided to use SVM.

CLASSIFIER EVALUATION: WILCOXON TEST

	SVM	k-NN	di	rank	sign
1-fold	0,4067	0,3382	0,0685	5	+
2-fold	0,3230	0,2608	0,06223	4	+
3-fold	0,5596	0,57029	-0,01069	1	-
4-fold	0,5729	0,5854	-0,0125	2	-
5-fold	0,4623	0,5103	-0,048	3	-
6-fold	0,5268	0,6276	-0,1008	6	-

W+ = 5+4 = 9
W- = 1+2+3+6 = 12
W = min{W+, W-} = 9

k*(k+1)/2 = 21
(W+) + (W-) = 21
Wcrit = 0 ~ 2
k*(k+1)/2 = 21 > 20

Assume that the distribution is normal and evaluate the z-score:
meanw = 10,5
varw = 4,76
z-score = -0,3151
Area = 0,1255*2 = 0,251
P-value = 0,749

CLASSIFIER EVALUATION: VEB VS SVEB(1 / 2)

VEB

	sens svm	prec svm	svm-f2	sens k-nn	prec k-nn	k-nn f2
1-fold	0,814	0,0818	0,2917	0,907	0,1099	0,3701
2-fold	0,0847	0,1136	0,0892	0,1356	0,1151	0,1309
3-fold	0,4732	0,165	0,3445	0,4732	0,1443	0,3250
4-fold	0,7387	0,2033	0,4838	0,6376	0,2478	0,4850
5-fold	0,5714	0,1082	0,3078	0,5262	0,1954	0,3931
6-fold	0,2172	0,2564	0,2240	0,668	0,6206	0,657
		mean	0,2902		**mean**	0,3936
		var	0,0143		**var**	0,0254

- pvalue = 0,19

CLASSIFIER EVALUATION: VEB VS SVEB(2/2)

SVEB

sens svm	prec svm	svm-f2	sens k-nn	prec k-nn	k-nn f2
0,8592	0,1429	0,4290	0,7298	0,3895	0,6212
0,5298	0,0418	0,1588	0,4643	0,0358	0,1368
0,9111	0,9598	0,9204	0,923	1	0,9374
0,9032	0,9032	0,9032	0,9115	0,5199	0,7921
0,73	0,4178	0,6350	0,8627	0,5873	0,7887
0,9647	0,9421	0,9600	0,8773	0,784	0,8569
	mean	0,6677		**mean**	0,6888
	var	0,0869		**var**	0,07

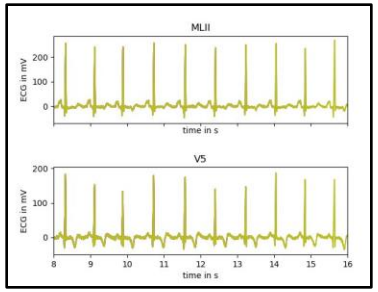
- pvalue = 0,7

PROGRAM

Once the application has started, the user can request to analyze a specific ecg

```
APPLICATION STARTED SUCCESSFULLY
Application commands are:
!quit - quit the application
!ecg_predict - perform arythmia detection on a specific ecg_signal
Enter your command:
```

First, the program will show the ecg wave on both MLII and V5



Then it will compute the features

```
Enter ecg-wave's name: 100
Loading 100...
Computing features: RR, HOS, Wavelet, Our Morph, Lbp and Hbf
RR ...
lbp ...
hbf ...
Wavelets ...
HOS ...
My Descriptor ...
```

And, show at video a summary of the prediction and save into a dedicated file the detailed results

```
Result Prediction ...
96.30118890356671% of the beats have been classified as belonging to the class N
3.6988110964332894% of the beats have been classified as belonging to the class F
The complete prediction has been saved into the file 100result.txt in the directory results
```