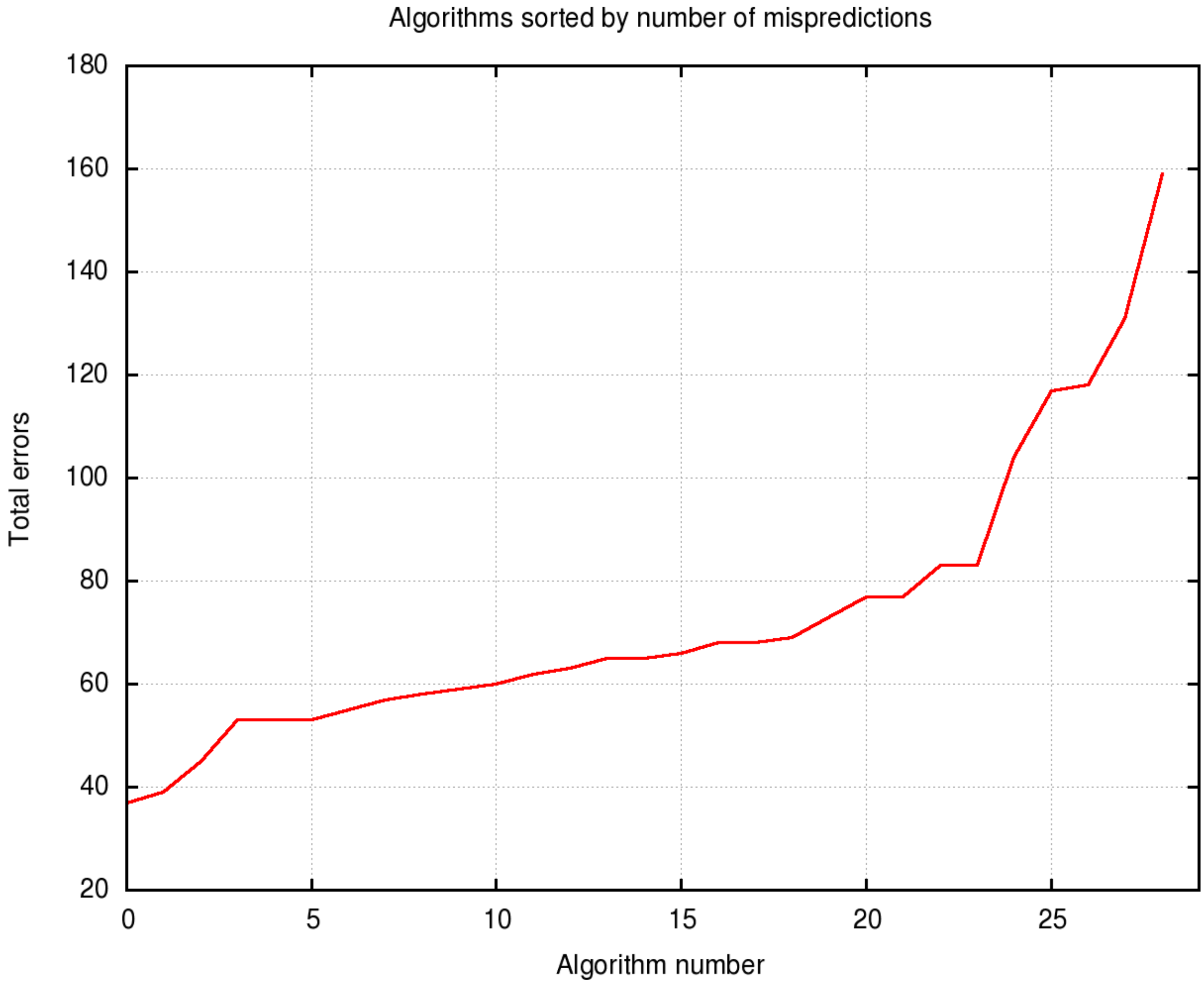
# Classification by decision tree induction over a medical database

*Roberto Innocente – inno@sissa.it*

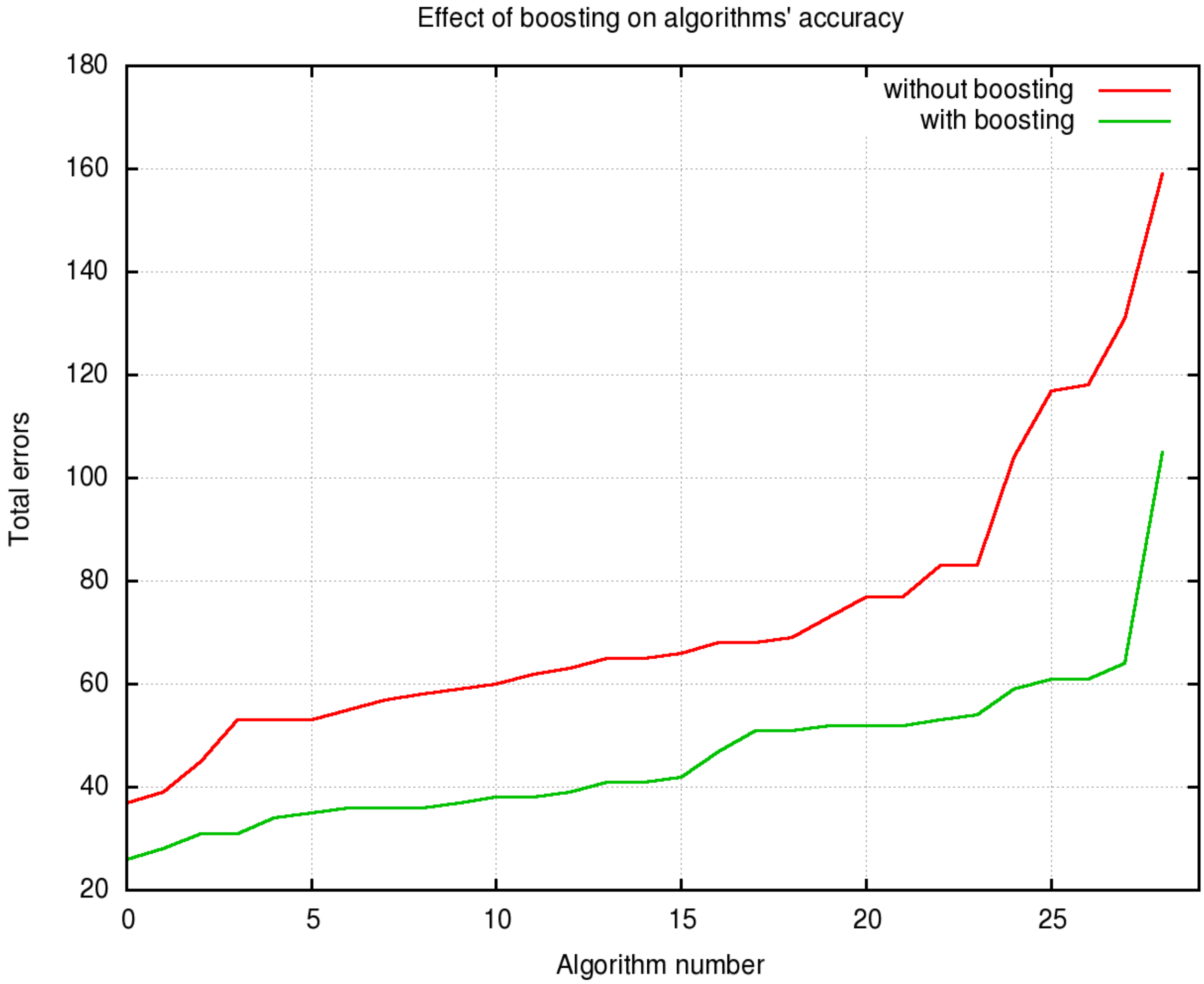
Classification by decision tree induction is a well known technique in machine learning [1],[2].
Many different methods to grow decision trees have been described[3],[4].
In this second phase of the project on a database of 868 patients admitted in a stroke unit at the Trieste hospital, we have tried 28 different attribute selection measures (14 regular and 14 1-level lookahead versions of the same) registering different accuracies with a minimum of 37 errors (~= 4.2%) , respect to the expert provided territorial classification of the strokes.




Some of them produced acceptable accuracy, but were not adherent to current clinical practice.
We applied 2 successive steps to increase accuracy and clinical acceptability of the results : *boosting* and *majority voting*.

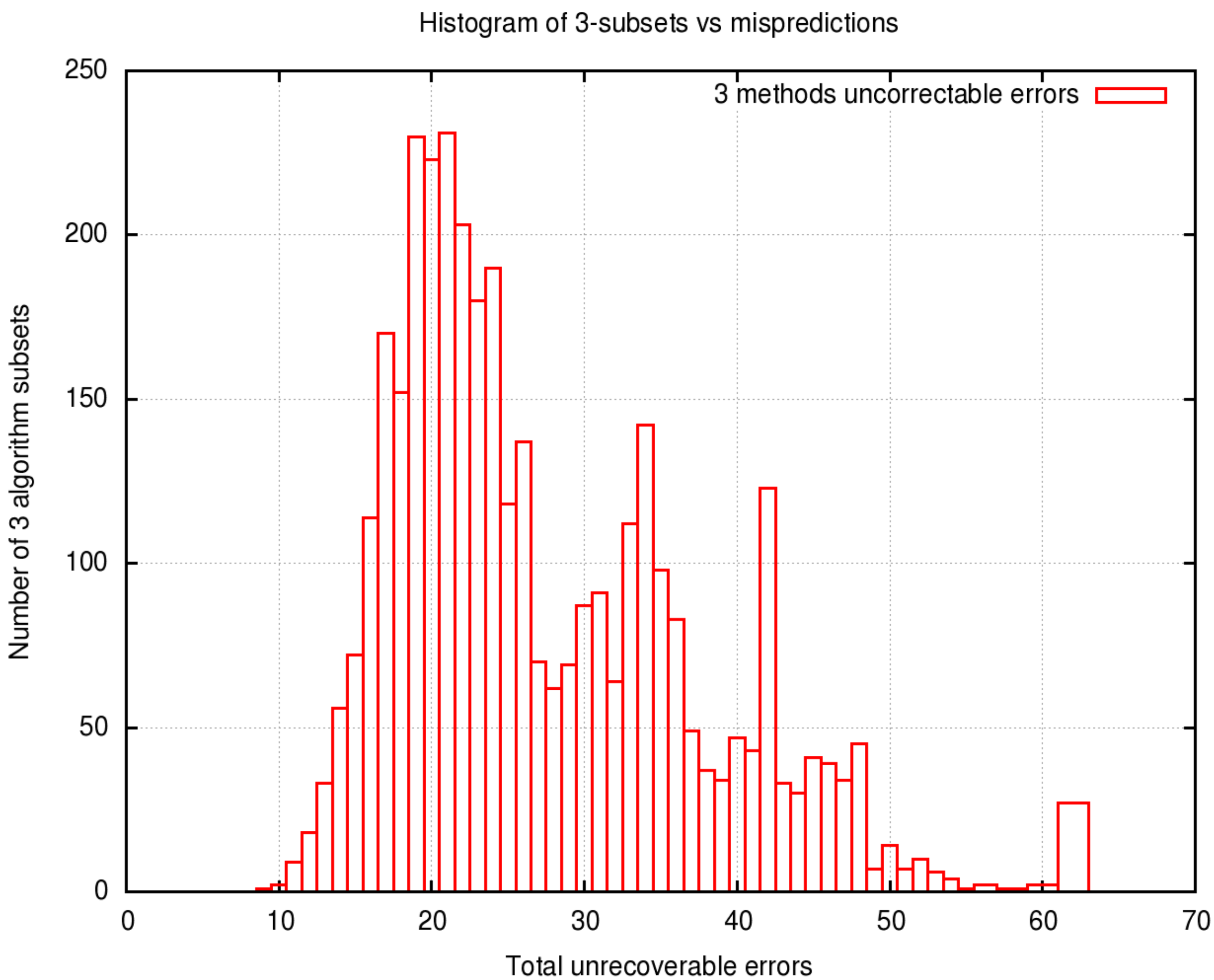
## Boosting

Boosting is a general way to increase classifier accuracy : a weight is assigned to each training example and successively the weights are updated increasing those of the mispredicted examples [5].
We obtained a general increase of accuracy using boosting, with a minimum of 26 errors ( ~= 3%), still the problem was the adherence to clinical practice of the best methods.

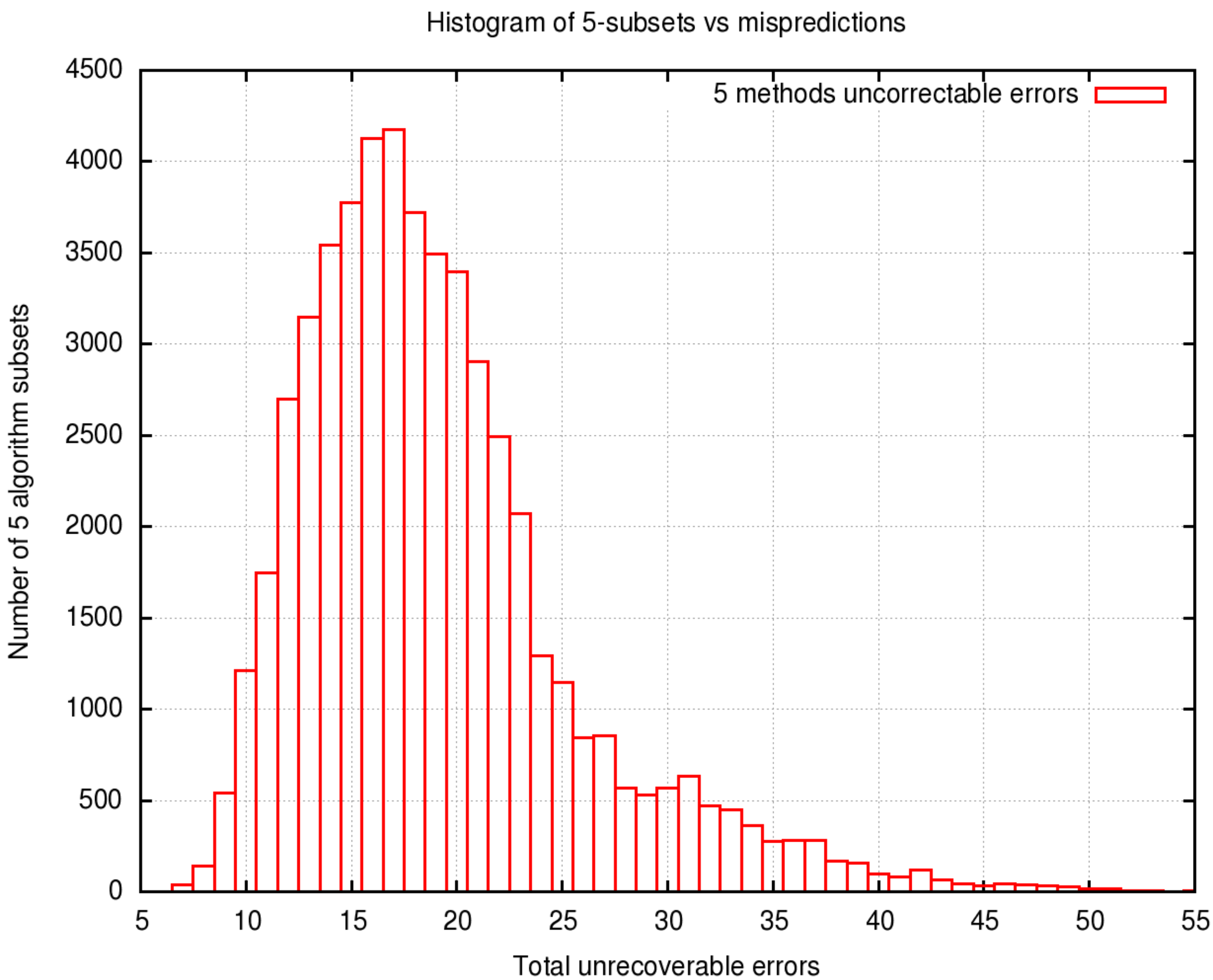


## Majority voting

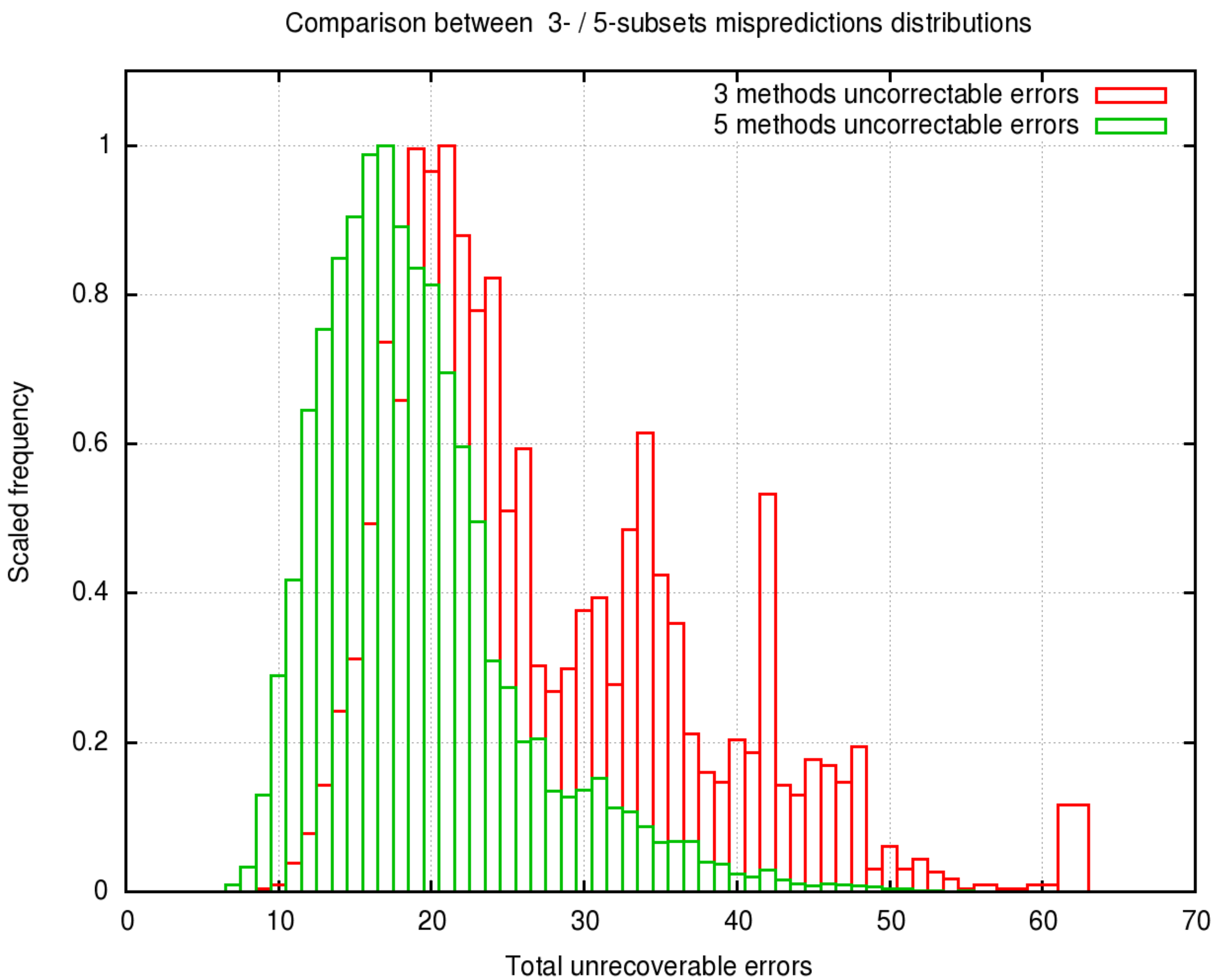
Majority voting is also a general technique to increase learners accuracy :
*2n+1* learners are used and in this way we can correct all the errors made by at most *n* learners [6].
We tried subsets of 3 methods out of the 28 boosted methods at a time, and we were able to decrease the number of errors down to 9 (~= 1%), with some possibility to choose between different sets accepting a slightly superior number of total errors. Using majority voting on 3 methods we can correct all the errors made by only 1 method.




We tried then to use majority voting on 5 methods out of the 28.
In this way all single and double errors can be corrected, 3,4 and 5 simultaneous errors are uncorrectable.
We had in this case a minimum of 6 (~=0.69%) uncorrectable mispredictions.

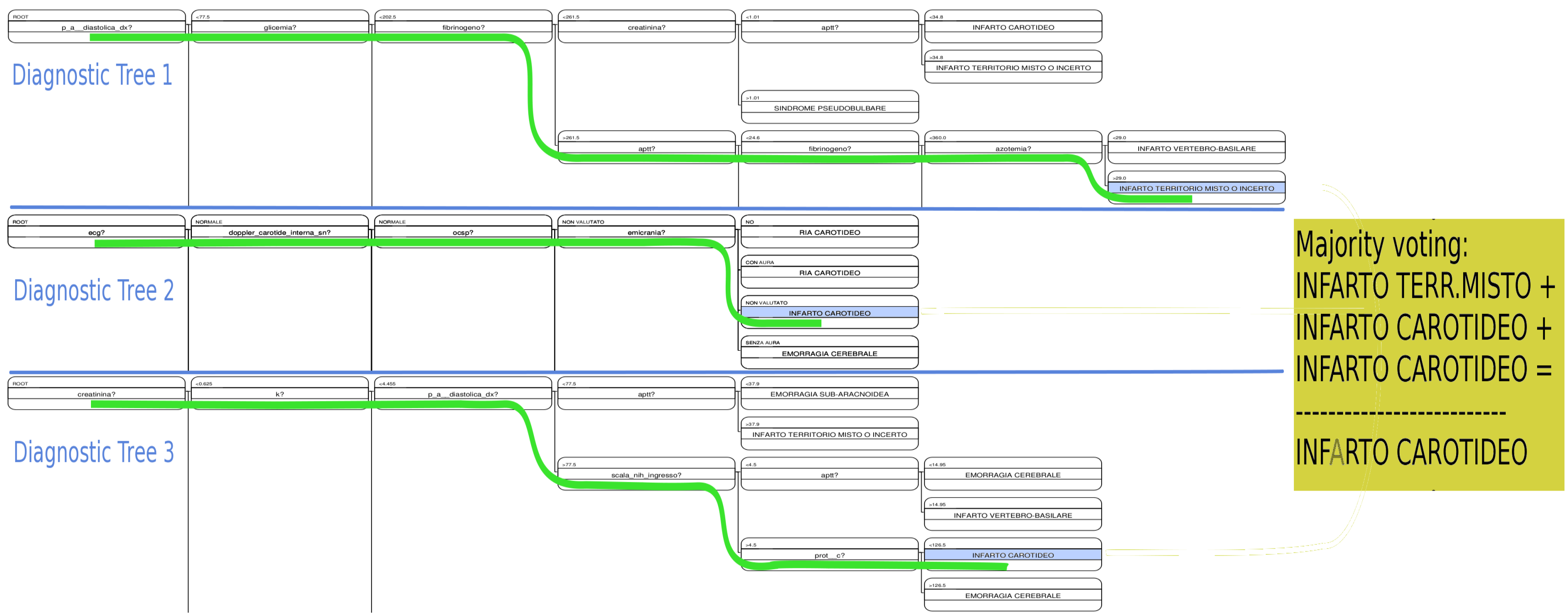



The distribution of uncorrectable errors shifted left again a bit but we think the burden of using 5 different trees is not justifiable.




Bibliography:

[1] J.R.Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986
[2] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann,1993
[3] W.L.Buntine and T.Niblett, A further comparison of splitting rules for decision-tree induction, Machine Learning, 8:75-85, 1992
[4] S.K.Murthy, Automatic construction of decision trees from data: A multidsciplinary survey. Data Mining and knowledge Discovery, 2:345-389,1998
[5] Y.Freund and R.E.Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55:119-139,1997
[6] J.Kittler. Combining classifiers: A theoretical framework. Pattern analysis and Applications, 1998



Diagnostic Tree 1
ROOT p_a__diastolica_dx?
<77.5 glicemia?
<202.5 fibrinogeno?
<261.5 creatinina?
<1.01 aptt?
<34.8 INFARTO CAROTIDEO
>34.8 INFARTO TERRITORIO MISTO O INCERTO
>1.01 SINDROME PSEUDOBULBARE
>261.5 aptt?
<24.6 fibrinogeno?
<360.0 azotemia?
<29.0 INFARTO VERTEBRO-BASILARE
>29.0 INFARTO TERRITORIO MISTO O INCERTO
Diagnostic Tree 2
ROOT ecg?
NORMALE doppler_carotide_interna_sn?
NORMALE ocsp?
NON VALUTATO emicrania?
NO RIA CAROTIDEO
CON AURA RIA CAROTIDEO
NON VALUTATO INFARTO CAROTIDEO
SENZA AURA EMORRAGIA CEREBRALE
Diagnostic Tree 3
ROOT creatinina?
<0.625 k?
<4.455 p_a__diastolica_dx?
<77.5 aptt?
<37.9 EMORRAGIA SUB-ARACNOIDEA
>37.9 INFARTO TERRITORIO MISTO O INCERTO
>77.5 scala_nih_ingresso?
<4.5 aptt?
<14.95 EMORRAGIA CEREBRALE
>14.95 INFARTO VERTEBRO-BASILARE
>4.5 prot__c?
<126.5 INFARTO CAROTIDEO
>126.5 EMORRAGIA CEREBRALE
Majority voting:
INFARTO TERR.MISTO +
INFARTO CAROTIDEO +
INFARTO CAROTIDEO =
------------------------
INFARTO CAROTIDEO