Combining multiple classifiers: Random Forest

Albert Cardenete Massip

May 2019

1 Introduction

In this second practical delivery for the SEL course we are going to study the Random Forest algorithm, which is a simple rule-based classifier. First we present the algorithm itself, by explaining its main parts and the decision tree classifier chosen (CART). Then, we are going to apply this model to three different datasets and we will compare the results of the predictions and compute an estimation of the importace of each attribute in each of the different experiments.

2 The Random Forest algorithm

Random forests [1] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In these types of algorithms, the generalization error for forests converges asymptotically to a limit as the number of trees in the forest becomes large.

Technically speaking, a random forest would be defined as:

Definition 1 *A* ***random forest*** *is a classifier consisting of a collection of* n_t *tree-structured classifiers* $\{h(\vec{x}, \Theta_k), k = 1, \ldots, n_t\}$*, where the* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\vec{x}$*.*

So, in other words, any classifier consisting of a number of tree-structured classifiers (like for instance ID3 [2], C4.5 [3] or CART [4]), in which we pick a random selection, either in the training samples or the attributes, and in which the final prediction is based on the vote of the most popular class, can be considered as a random forest.

Although this freedom might yield into a large number of possible implementations of a random forest, when Breiman proposed this algorithm [1], he considered a random selection of features, along with bagging of instances. Bagging consists in training each of the trees in the random forest with a random selection of the training instances, without replacement, to finally get the same number of instances to train. On the other hand, the random selection of features would be done in every splitting node of each tree. He did that in order to reduce the correlation between trees in the forest, because that is an upper bound of the generalization error in a random forest.

One can see in the algorithm 1, the basic implementation of a random forest.

Algorithm 1 Random Forest Algorithm

Input $\mathcal{D}, n_t, n_{\text{features}}$ ▷ Dataset, number of trees and number of features considered
forest ← {} ▷ Final trees obtained
for $i = 1, \ldots, n_t$ **do**
    $\mathcal{D}_2$ ← Bagging($\mathcal{D}$)
    forest ← [forest, DecisionTree($\mathcal{D}_2$,n_{features})] ▷ DecisionTree might be any classifier
end for

On the other hand, we need a tree-structured classifier. In our implementation, we will use CART [4]. This classifier works by finding for each of the considered attributes, the splitting that maximizes the reduction of the Gini index. On continuous variables, it considers midpoint between each pair of sorted adjacent values. On discrete attributes, it examines the partitions resulting from all possible subsets of the possible values of that attribute. In this implementation, no pruning was implemented, as in the original paper.

Algorithm 2 CART Algorithm

Input $\mathcal{D}, n_{\text{features}}$ ▷ Dataset and number of features considered
if all_instances_are_classified($\mathcal{D}$) **then**
    tree ← DecisionTree.add_final_class_node($\mathcal{D}$)
    return tree
else
    attributes_reduced ← reduced_random_attributes($\mathcal{D}, n_{\text{features}}$)
    rule ← best_splitting_point($\mathcal{D}$,attributes_reduced) ▷ Split to maximize the reduction of Gini
    $\mathcal{D}_1, \mathcal{D}_2$ ← split_according_to_rule(rule,$\mathcal{D}$)
    if rule **is None then** ▷ If no rule could be found
        tree ← DecisionTree.add_final_class_node_from_mode($\mathcal{D}$)
        return tree
    end if
    tree_1 ← DecisionTree.add_node(rule[0]).add_child(CART($\mathcal{D}_1, n_{\text{features}}$))
    tree_2 ← DecisionTree.add_node(rule[1]).add_child(CART($\mathcal{D}_2, n_{\text{features}}$))
    return tree_1, tree_2
end if

Finally, in order to get a prediction of a new instance, we will predict the result for each of the trees in the forest. The final predicted value will be the most voted from all the predicted values. In case of ties, we will randomly select one of the results with most votes.

3 Experimental Setup

In order to get an estimation of the performance of our resulting model in every dataset studied, we will perform a k-fold cross validation by splitting the data. Due to time restrictions, we will perform a 3-fold cross validation.

In addition to that, we will also compute an estimation of the importance of the features of the dataset. This estimation will consist in counting the number of times in which we split the tree by a given attribute. This is because if we can split the tree with a different number of attributes, we will choose more times the most discriminating attributes.

The datasets studied in this work will be extracted from the UCI ML repository[5], which are:

- **Contraceptive Method:** Containing 1473 instances with 5 categorical attributes, 4 numerical attributes and 3 classes.
- **Nursery:** Containing 12960 instances with 8 categorical attributes and 5 classes.
- **Voting records:** Containing 435 instances with 16 categorical attributes and 2 classes.

For each one of the datasets, we are going to perform experiments with a different number of trees in the forest and number of random attributes considered at each node split:

- Number of trees = 50, Number of attributes = 1.
- Number of trees = 50, Number of attributes = 3.
- Number of trees = 50, Number of attributes = int(log(M)+1).
- Number of trees = 50, Number of attributes = sqrt(M).
- Number of trees = 100, Number of attributes = 1.
- Number of trees = 100, Number of attributes = 3.
- Number of trees = 100, Number of attributes = int(log(M)+1).
- Number of trees = 100, Number of attributes = sqrt(M).

where M is the number of total attributes in each dataset.

4 Results and discussion

4.1 Voting dataset

In this first dataset analyzed, we can see the results of the 3-fold cross validation in the table 1. In these results we can see that best results are achieved when we are considering more than one random feature per node when we split the tree. Although, when we consider 100 trees having one or more number of random features is not statically significant.

n_t	n_{features}	Accuracy	Time [s]
50	1	0.931±0.017	31.11±0.54
50	3	0.9586±0.0098	37.24±0.95
50	$\text{int}(\log(M)+1) = 5$	0.9563±0.0033	39.43±0.71
50	$\sqrt{M} = 4$	0.9540±0.0033	39.08±0.43
100	1	0.933±0.027	62.98±0.80
100	3	0.9540±0.0033	74.0±1.9
100	$\text{int}(\log(M)+1) = 5$	0.9563±0.0033	77.95±0.61
100	$\sqrt{M} = 4$	0.9517±0.0001	75.6±1.9

Table 1: Results of the evaluation of the 3-fold cross validation for the Random Forest with the Voting dataset.

On the other hand, we see that the algorithm scales linearly in the number of trees, and considering a larger number of features does not drastically increase the time needed to induce the random forest.

Finally, we can see the results of the importance on the attributes in the section A.1. The overall results for this dataset has been definitely positive. As a comparison, the best result achieved in the previous delivery with the RULES algorithm, we achieved an accuracy of 0.951, just slightly smaller.

4.2 Nursery dataset

Secondly, we have performed the same experiments for the nursery dataset. The results can be seen in the table 2. In this second dataset we can clearly see that considering more than one feature per splitting node increases the performance of the classifier with statistical significance. We get from $\sim 90\%$ to $\sim 99\%$. We can even see that the best results are achieved with 4 features (the strategy of the logarithm of the number of features), with either 50 or 100 trees in the forest.

n_t	n_{features}	Accuracy	Time [s]
50	1	0.8997±0.0091	854±158
50	3	0.9921±0.0016	851±18
50	$\text{int}(\log(M) + 1) = 4$	0.99630±0.00068	848.5±3.4
50	$\sqrt{M} = 3$	0.9921±0.0016	851±18
100	1	0.908±0.011	1443.6±5.2
100	3	0.99321±0.00048	1689.3±4.3
100	$\text{int}(\log(M) + 1) = 4$	0.99660±0.00022	1700.1±4.0
100	$\sqrt{M} = 3$	0.99321±0.00048	1689.3±4.3

Table 2: Results of the evaluation of the 3-fold cross validation for the Random Forest with the Nursery dataset.

On the other hand, we see that the training times for this dataset are much larger than in the previous experiment. This is due to the larger amount of instances (nearly 20 times more instances than the previous case). As a comparison with the previous delivery, with the RULES algorithm we achieved a 97.5% accuracy and lasted 169 seconds. Clearly we see an improvement on the performance of the algorithm, although this comes at a price of more training time.

The results of the importancy of this dataset in each experiment can be seen in the section A.2.

4.3 Contraceptive dataset

The results for this final experiments can be seen in the table 3. As in the previous case, we can see that the better results are achieved with more than one feature considered per node. This results are also statistically significant. Although that, the accuracies for this dataset are not so good. We do not even see an improvement with a higher number of trees in the forest. This might be due to the fact that we are near the possible maximum achievable result in this dataset for the proposed random forest method.

n_t	n_{features}	Accuracy	Time [s]
50	1	0.5105±0.0019	67.6±1.1
50	3	0.5214±0.0033	190.5±4.4
50	$\text{int}(\log(M)+1)=4$	0.513±0.014	747±28
50	$\sqrt{M}=3$	0.5214±0.0033	190.5±4.4
100	1	0.505±0.015	138.9±5.8
100	3	0.522±0.011	391.3±3.6
100	$\text{int}(\log(M)+1)=4$	0.5241±0.0035	487.0±3.7
100	$\sqrt{M}=3$	0.522±0.011	391.3±3.6

Table 3: Results of the evaluation of the 3-fold cross validation for the Random Forest with the Contraceptive dataset.

Regarding the times, we can see that it takes a long time to compute the results. If we compare it with the first dataset, we can see that we have ~ 3.5 times more instances, and it takes ~ 2 times more to train. The importance of the attributes in each of the experiments can be seen in the section A.3.

Although the bad results, this is better than random guessing, as we have 3 possible outcome classes. In addition, this is a well known difficult dataset, as might include classes which are labeled incorrectly, due to the nature of how this dataset was built, in the Indonesian society of the 1980's.

5 How to run

First of all, we will set up our python environment with the required packages. Those are pandas (≥0.23.0), scikit-learn (≥0.20.3) and numpy (≥1.15.4).

In order to run the code, first we will have to enter into the `Practical` folder in a console. Then, without entering into the Source folder, the command to run the code is simply:

```
python Source
```

This command has some flags, which can be seen using the help command:

```
python Source -h
```

The list of all the flags are:

- `--dataset`: The name of the file inside the `Practical/Data` folder ended with `.csv`.
- `--classname`: The name of the target class in the previous csv file. The default is `"Class"`
- `--k`: (in lower letters) The number of folds for the validation process. The deault is 3.
- `--nt`: The number of trees for the random forest. The deault is 100.
- `--f`: The number of random attributes considered at each split in the trees. The deault is 1.

So for example, if we have the dataset "`DS.csv`" inside the Data folder with the target class named "`Cls_attr`", if we would like to perform the validation with 50 trees and 3 features with a 3-fold cross validation, we would write:

```
python Source --dataset DS.csv --classname Cls_attr --nt 50 --f 3
```

6 Conclusions

To conclude, we see that the random forest method to generate an ensemble of tree-based classifiers is a very simple, yet effective model for that have some benefits over the bare decision trees by themselves. We have seen that this ensemble does not suffer from overfitting with an increasing number of trees. We have also seen that considering a reduced number of random attributes, we can achieve results as if we were considering all the possible attributes, with the benefits of reducing the time for all those computations.

Although, due to its computational complexity, it is unfeasible to be applied to very large datasets, but we have seen that by applying this method to a dataset with more than ten thousand instances the time taken to train has been almost linear.

A Results of the importance for each dataset

A.1 Voting dataset

```
1.NT=50, F=1
----------------------------------------------------------------------------
Final accuracy: 0.9310344827586207±0.016893032708849492
Final time: 31.114462455113728±0.5371877835146983
Final importance metrics:
duty-free-exports : 0.07307205304733684 ± 0.008288071217891924
physician-fee-freeze : 0.07075275350957862 ± 0.004030515068658654
immigration : 0.06768069331608124 ± 0.00263380758679202
export-administration-act-south-africa : 0.06494737229778941 ± 0.014066902035457886
synfuels-corporation-cutback : 0.06486731034078776 ± 0.0011866188203290189
water-project-cost-sharing : 0.06428578167505904 ± 0.002940121628745765
superfund-right-to-sue : 0.06378460593591989 ± 0.0038648643331167147
education-spending : 0.063548077217659 ± 0.0043922087412895
handicapped-infants : 0.06291600457292269 ± 0.0005322894732385195
crime : 0.061884980439132674 ± 0.00781896658962691
adoption-of-the-budget-resolution : 0.06073157764937511 ± 0.003153214384494541
mx-missile : 0.059853321744678036 ± 0.005927638712850627
aid-to-nicaraguan-contras : 0.05907699912047842 ± 0.003794641127029862
el-salvador-aid : 0.05807096290317459 ± 0.004894767074712223
religious-groups-in-schools : 0.054222871055065534 ± 0.004431007525182455
anti-satellite-test-ban : 0.05030463517496117 ± 0.0015451971140357278
----------------------------------------------------------------------------

2.NT=50, F=3
----------------------------------------------------------------------------
Final accuracy: 0.9586206896551724±0.00975319698188338
Final time: 37.237003326416016±0.9506731770241289
Final importance metrics:
physician-fee-freeze : 0.11904629080416405 ± 0.004449829252501403
synfuels-corporation-cutback : 0.08537156052738266 ± 0.012920951080863106
adoption-of-the-budget-resolution : 0.07724546627917804 ± 0.010921665237166568
education-spending : 0.07097616807375655 ± 0.012996937558544742
export-administration-act-south-africa : 0.06724093815732923 ± 0.0023233423284393785
water-project-cost-sharing : 0.06536426207126149 ± 0.006512491886009798
```

```
superfund-right-to-sue : 0.062329089499919256 ± 0.004938578719952131
immigration : 0.062248605214636855 ± 0.006827017104741535
handicapped-infants : 0.06108348459209901 ± 0.00432423653424819
duty-free-exports : 0.060666052015920846 ± 0.0033989453172263385
anti-satellite-test-ban : 0.05118907893650626 ± 0.005231755300836574
mx-missile : 0.048592699494568146 ± 0.007336764181508953
aid-to-nicaraguan-contras : 0.04451070209759971 ± 0.010047068364586206
crime : 0.04403904012119871 ± 0.00422332354814216
el-salvador-aid : 0.041958631112978484 ± 0.0031891720122336304
religious-groups-in-schools : 0.03813793100150067 ± 0.004020801347277375
-----------------------------------------------------------------------------

3.NT=50, F=int(logM+1)=5
-----------------------------------------------------------------------------
Final accuracy: 0.9563218390804598±0.003251065660627811
Final time: 39.4273419380188±0.7146127147943576
Final importance metrics:
physician-fee-freeze : 0.1321283570690525 ± 0.005313014458368772
synfuels-corporation-cutback : 0.09074890785794583 ± 0.010520426966500062
adoption-of-the-budget-resolution : 0.09028378227539398 ± 0.011254293885518778
export-administration-act-south-africa : 0.07503742309341667 ± 0.0043332464602113124
education-spending : 0.07406379044659366 ± 0.01132253994101719
immigration : 0.06972882272809926 ± 0.003968651270346114
duty-free-exports : 0.060081414621927974 ± 0.012337932584227771
water-project-cost-sharing : 0.05959855450182341 ± 0.010830595583001583
handicapped-infants : 0.05603291765436205 ± 0.0013797492499674736
superfund-right-to-sue : 0.05553917404195424 ± 0.0041327562073286295
anti-satellite-test-ban : 0.048224649106671624 ± 0.009077714052925559
aid-to-nicaraguan-contras : 0.04130721733763632 ± 0.0036962794681499023
mx-missile : 0.04091578436192634 ± 0.007744796367403746
crime : 0.03879797148840158 ± 0.005849431723903606
el-salvador-aid : 0.03664216502459904 ± 0.01037383523086258
religious-groups-in-schools : 0.03086906839019553 ± 0.007112716610662741 6
-----------------------------------------------------------------------------

4.NT=50, F=sqrt(16)=4
-----------------------------------------------------------------------------
Final accuracy: 0.9540229885057472±0.003251065660627811
Final time: 39.0816384156545±0.4261886087157176
Final importance metrics:
physician-fee-freeze : 0.12382550358512256 ± 0.00847930744713427
adoption-of-the-budget-resolution : 0.0909330217629205 ± 0.006623753075004073
synfuels-corporation-cutback : 0.08690015757353077 ± 0.013082896924664909
education-spending : 0.0667885662364166 ± 0.012542107436506692
export-administration-act-south-africa : 0.0640269107501698 ± 0.005093624016533393
superfund-right-to-sue : 0.06381787367534782 ± 0.0015764361227732025
immigration : 0.06315474151688756 ± 0.01103495274431208
water-project-cost-sharing : 0.06261016873511129 ± 0.010037072710721224
duty-free-exports : 0.06051882132225248 ± 0.008607546704706514
handicapped-infants : 0.059351997507675204 ± 0.005275374272505139
anti-satellite-test-ban : 0.04689056374537689 ± 0.008085805556937196
```

el-salvador-aid : 0.04681981467294999 ± 0.003949033140772897
aid-to-nicaraguan-contras : 0.04380503084376325 ± 0.010366180086630452
crime : 0.043430662436721834 ± 0.004610599279551086
mx-missile : 0.042610654805681186 ± 0.005050115416957348
religious-groups-in-schools : 0.03451551083007226 ± 0.007211014640184737

5.NT=100, F=1

Final accuracy: 0.9333333333333332±0.026611119316759142
Final time: 62.980676809946694±0.8028995074556691
Final importance metrics:
physician-fee-freeze : 0.07049266165184802 ± 0.0052761964183893235
water-project-cost-sharing : 0.07025996317017853 ± 0.0014682284391661208
duty-free-exports : 0.06832997898306034 ± 0.005761817131625796
education-spending : 0.06756799999682046 ± 0.001836722262734179
handicapped-infants : 0.06741852385771568 ± 0.004478982607860192
adoption-of-the-budget-resolution : 0.06596292409700338 ± 0.004415164755910913
superfund-right-to-sue : 0.0654693622794127 ± 0.00043507394673074085
mx-missile : 0.06472926308768537 ± 0.007499279656482866
export-administration-act-south-africa : 0.06325512853473371 ± 0.003994676131291229
immigration : 0.061748309846149235 ± 0.001656789670625696
synfuels-corporation-cutback : 0.061177063716854836 ± 0.0039289380762645865
aid-to-nicaraguan-contras : 0.0580808465960666 ± 0.0023776877124171295
anti-satellite-test-ban : 0.05760462359451454 ± 0.0035629158535826905
el-salvador-aid : 0.05733549751898204 ± 0.0038154742945511397
crime : 0.052488139669483035 ± 0.002925742353959081
religious-groups-in-schools : 0.048079713399491464 ± 0.001648467935103804

6.NT=100, F=3

Final accuracy: 0.9540229885057472±0.003251065660627811
Final time: 73.96685338020325±1.937628187813939
Final importance metrics:
physician-fee-freeze : 0.1142898947306159 ± 0.0050303511931602055
synfuels-corporation-cutback : 0.08339527798222446 ± 0.00283761120520831
adoption-of-the-budget-resolution : 0.07942528273209247 ± 0.0083732216302085
education-spending : 0.07343868293954449 ± 0.01097585818729006
export-administration-act-south-africa : 0.06720026051834745 ± 0.002381162610434184
immigration : 0.06620714656976791 ± 0.003754330929047702
water-project-cost-sharing : 0.06216640917760232 ± 0.003088381073992618
handicapped-infants : 0.061585367532862034 ± 0.004523426021898273
superfund-right-to-sue : 0.060770475409465084 ± 0.005567887686900882
duty-free-exports : 0.05738769711144085 ± 0.0011034814633468651
anti-satellite-test-ban : 0.052159878857411884 ± 0.005394452430187691
mx-missile : 0.051709389735950684 ± 0.0019694122353057213
crime : 0.04920471440266069 ± 0.0009466130636448217
aid-to-nicaraguan-contras : 0.04759825743183125 ± 0.0036696194206569015
el-salvador-aid : 0.03895590130092743 ± 0.0018844258373708431
religious-groups-in-schools : 0.03450536356725505 ± 0.003184214111689194

7.NT=100, F=int(logM+1)=5

Final accuracy: 0.9563218390804598±0.003251065660627811
Final time: 77.94768762588501±0.6073242223208424
Final importance metrics:
physician-fee-freeze : 0.13417338406937643 ± 0.007965821682543346
adoption-of-the-budget-resolution : 0.09013372955131793 ± 0.008604092052423518
synfuels-corporation-cutback : 0.09011500545739011 ± 0.011074572702583232
education-spending : 0.0750572713596719 ± 0.01811984044661329
export-administration-act-south-africa : 0.0729665666837046 ± 0.0037948201986814853
water-project-cost-sharing : 0.06621793144912459 ± 0.006363730969653996
immigration : 0.06462203360531371 ± 0.0048393996797516554
superfund-right-to-sue : 0.060410444486984884 ± 0.006688749958541732
duty-free-exports : 0.05745530568028225 ± 0.010188018345817892
handicapped-infants : 0.05122672894578149 ± 0.003080114507782187
mx-missile : 0.04609049867469391 ± 0.0034222281603990563
anti-satellite-test-ban : 0.04590462891666907 ± 0.0043038091665424045
aid-to-nicaraguan-contras : 0.042409799718280454 ± 0.0069741286334713895
el-salvador-aid : 0.03981057300010556 ± 0.004940812362701602
crime : 0.03564530213725332 ± 0.008739481325830549
religious-groups-in-schools : 0.02776079626404979 ± 0.0036609480489277787

8.NT=100, F=sqrt(16)=4

Final accuracy: 0.9517241379310345±0.0
Final time: 75.63278865814209±1.9360579989363906
Final importance metrics:
physician-fee-freeze : 0.11799961675791122 ± 0.00347696471469505
synfuels-corporation-cutback : 0.08970334441722655 ± 0.0038189276290184807
adoption-of-the-budget-resolution : 0.08415794358242291 ± 0.012707180561328439
education-spending : 0.07484064887652037 ± 0.008466334641027268
export-administration-act-south-africa : 0.0678905006380205 ± 0.005636416485512847
water-project-cost-sharing : 0.06587386571618882 ± 0.0020295845693313235
immigration : 0.06294993364831089 ± 0.0026788251203514256
duty-free-exports : 0.06146605333912344 ± 0.011291413743442651
handicapped-infants : 0.05830351971538491 ± 0.0023934425712562922
superfund-right-to-sue : 0.05764086842202604 ± 0.0035527784607872846
anti-satellite-test-ban : 0.0517290750783951 ± 0.01191898920128554
mx-missile : 0.04594929238185266 ± 0.005556456131128142
aid-to-nicaraguan-contras : 0.04379664627348748 ± 0.005140512871164276
crime : 0.04257779388848307 ± 0.006901563220203473
el-salvador-aid : 0.040108715093932885 ± 0.007280436650286833
religious-groups-in-schools : 0.03501218217071315 ± 0.002245382328270528

A.2 Nursery Dataset

1.NT=50, F=1

```
--------------------------------------------------------------------------------
Final accuracy: 0.8996913580246915±0.009088576007238617
Final time: 854.259694258372±157.88931852192513
Final importance metrics:
health : 0.16433670311101556 ± 0.0006340590613236211
has_nurs : 0.14805557230455876 ± 0.010554798122806894
form : 0.13584248080964298 ± 0.0029712146304975997
children : 0.1326142182344863 ± 0.002007943447729405
parents : 0.1153823474092718 ± 0.005747452784909825
social : 0.11217920690619741 ± 0.003875188929665872
housing : 0.11155101882194753 ± 0.005691450139300264
finance : 0.08003845240287967 ± 0.004964466618731991
--------------------------------------------------------------------------------

2.NT=50, F=3
--------------------------------------------------------------------------------
Final accuracy: 0.9921296296296296±0.0016368212527466321
Final time: 850.775496562322±17.795949275638023
Final importance metrics:
form : 0.18148083006011895 ± 0.004130263014942239
children : 0.15994396629768168 ± 0.0039032260651462093
has_nurs : 0.13033958682091532 ± 0.0030256206081896455
social : 0.11959617142689155 ± 0.0064929879381502515
housing : 0.11674198266723745 ± 0.00779441785918439
health : 0.1046317284351818 ± 0.012300370024978047
finance : 0.10113040188150563 ± 0.002878303053751981
parents : 0.08613533241046763 ± 0.006458639564083289
--------------------------------------------------------------------------------

3.NT=50, F=int(logM+1)=4
--------------------------------------------------------------------------------
Final accuracy: 0.9962962962962963±0.0006814630298092743
Final time: 848.5244580109915±3.4366979033932745
Final importance metrics:
form : 0.1992134107697742 ± 0.0011994264832483235
children : 0.17962796703071557 ± 0.0037012029046899543
housing : 0.142725651592899 ± 0.004646324561641313
social : 0.1195638766010021 ± 0.0029319054499981137
finance : 0.1122812067983392 ± 0.0019960386862363025
has_nurs : 0.10934889841481298 ± 0.008008220374910318
health : 0.07009071579625943 ± 0.006554472454185352
parents : 0.06714827299619756 ± 0.0031958918662745397
--------------------------------------------------------------------------------

4.NT=50, F=sqrt(8)=3
--------------------------------------------------------------------------------
Final accuracy: 0.9921296296296296±0.0016368212527466321
Final time: 850.775496562322±17.795949275638023
Final importance metrics:
form : 0.18148083006011895 ± 0.004130263014942239
children : 0.15994396629768168 ± 0.0039032260651462093
```

```
has_nurs : 0.13033958682091532 ± 0.0030256206081896455
social : 0.11959617142689155 ± 0.0064929879381502515
housing : 0.11674198266723745 ± 0.00779441785918439
health : 0.1046317284351818 ± 0.012300370024978047
finance : 0.10113040188150563 ± 0.002878303053751981
parents : 0.08613533241046763 ± 0.006458639564083289
-----------------------------------------------------------------------------

5.NT=100, F=1
-----------------------------------------------------------------------------
Final accuracy: 0.9080246913580247±0.01082834203190333
Final time: 1443.5894656181335±5.20321076373738
Final importance metrics:
health : 0.1584288586949241 ± 0.0031294599729817358
has_nurs : 0.1569074145588511 ± 0.002726794084004306
form : 0.1317939968031162 ± 0.0033129308330585015
children : 0.12982250224804287 ± 0.005227486292251625
parents : 0.11932448763649851 ± 0.0036419883150257755
housing : 0.1167302190212328 ± 0.0015240682171377834
social : 0.11198243716232938 ± 0.003294258608266424
finance : 0.07501008387500503 ± 0.003858579496413041
-----------------------------------------------------------------------------

6.NT=100, F=3
-----------------------------------------------------------------------------
Final accuracy: 0.9932098765432098±0.00047564922862416173
Final time: 1689.3294450441997±4.27443187963344
Final importance metrics:
form : 0.1770916048574034 ± 0.0010922224586077905
children : 0.1605220521903434 ± 0.004599826478981661
has_nurs : 0.12696980075139733 ± 0.0023792023383498815
social : 0.1189995512080399 ± 0.0018661380508464152
housing : 0.1171252716157254 ± 0.004561601590233211
health : 0.11433210647055331 ± 0.005186552496951549
finance : 0.10008541522449677 ± 0.004173934981950158
parents : 0.08487419768204045 ± 0.005985224624868464
-----------------------------------------------------------------------------

7.NT=100, F=int(logM+1)=4
-----------------------------------------------------------------------------
Final accuracy: 0.9966049382716049±0.00021824283369951605
Final time: 1700.1369864940643±3.98366982628215
Final importance metrics:
form : 0.20034777330439535 ± 0.0010430840430055814
children : 0.18330387498919495 ± 0.004409645589269729
housing : 0.13597106410903056 ± 0.001657352423488458
social : 0.11949347271149309 ± 0.005089089751596262
finance : 0.11124581824273024 ± 0.002157336936343101
has_nurs : 0.10575619831256576 ± 0.0011204934961384712
parents : 0.07644508886689814 ± 0.0038176364048026607
health : 0.06743670946369194 ± 0.006275811707482475
```

```
--------------------------------------------------------------------------------

8.NT=100, F=sqrt(8)=3
--------------------------------------------------------------------------------
Final accuracy: 0.9932098765432098±0.00047564922862416173
Final time: 1689.3294450441997±4.27443187963344
Final importance metrics:
form : 0.1770916048574034 ± 0.0010922224586077905
children : 0.1605220521903434 ± 0.004599826478981661
has_nurs : 0.12696980075139733 ± 0.0023792023383498815
social : 0.1189995512080399 ± 0.0018661380508464152
housing : 0.1171252716157254 ± 0.004561601590233211
health : 0.11433210647055331 ± 0.005186552496951549
finance : 0.10008541522449677 ± 0.004173934981950158
parents : 0.08487419768204045 ± 0.005985224624868464
--------------------------------------------------------------------------------
```

A.3 Contraceptive dataset

```
1.NT=50, F=1
--------------------------------------------------------------------------------
Final accuracy: 0.5105227427019687±0.0019201813474176747
Final time: 67.55975699424744±1.1162524283847606
Final importance metrics:
age : 0.16613675107547254 ± 0.0032565156151225976
childs : 0.15314420391452543 ± 0.004602204218124627
living : 0.1247912868819887 ± 0.0019459828135522277
w_education : 0.12401988673769854 ± 0.0020564437797700817
occupation : 0.11805239889468427 ± 0.005694766759487758
h_education : 0.1174686203721843 ± 0.0058529777773537114
working : 0.07732258103663496 ± 0.0034084660997204907
media : 0.06018626882112784 ± 0.0005757130542648576
religion : 0.058878002265683405 ± 0.00020680847723009613
--------------------------------------------------------------------------------

2.NT=50, F=3
--------------------------------------------------------------------------------
Final accuracy: 0.5213849287169042±0.0033258516534734332
Final time: 190.5179315408071±4.391603046712774
Final importance metrics:
age : 0.26377314794371304 ± 0.0023928047792060782
childs : 0.18983499138888701 ± 0.0011923001943283994
living : 0.11756911584441987 ± 0.005507547481105673
occupation : 0.10809140783350131 ± 0.0031984604243066096
w_education : 0.09431103879082837 ± 0.0028197884554986004
h_education : 0.08698617530200452 ± 0.0037521063106929968
working : 0.06677809113230983 ± 0.0030617978370063516
religion : 0.048246713028418875 ± 0.0019520445981006968
media : 0.024409318735917156 ± 0.0014219860712212052
--------------------------------------------------------------------------------
```

```
3.NT=50, F=int(logM+1)=4
----------------------------------------------------------------------------
Final accuracy: 0.5132382892057027±0.013611634525297667
Final time: 746.6161858240763±28.053016381566323
Final importance metrics:
age : 0.2946435186093404 ± 0.00538832391239139
childs : 0.18100036731252797 ± 0.005318717824202917
living : 0.1225848293276784 ± 0.007316025320947619
occupation : 0.10465791831573855 ± 0.0015005531354228985
h_education : 0.08378072130638621 ± 0.0025178648551507227
w_education : 0.07855418165760335 ± 0.0050880799124732035
working : 0.0653410288541333 ± 0.001234574117698299
religion : 0.04828393225262245 ± 0.004428205038038121
media : 0.02115350236396937 ± 0.0021461072781136696
----------------------------------------------------------------------------

4.NT=50, F=sqrt(10)=3
----------------------------------------------------------------------------
Final accuracy: 0.5213849287169042±0.0033258516534734332
Final time: 190.5179315408071±4.391603046712774
Final importance metrics:
age : 0.26377314794371304 ± 0.0023928047792060782
childs : 0.18983499138888701 ± 0.0011923001943283994
living : 0.11756911584441987 ± 0.005507547481105673
occupation : 0.10809140783350131 ± 0.0031984604243066096
w_education : 0.09431103879082837 ± 0.0028197884554986004
h_education : 0.08698617530200452 ± 0.0037521063106929968
working : 0.06677809113230983 ± 0.0030617978370063516
religion : 0.048246713028418875 ± 0.0019520445981006968
media : 0.024409318735917156 ± 0.0014219860712212052
----------------------------------------------------------------------------

5.NT=100, F=1
----------------------------------------------------------------------------
Final accuracy: 0.505091649694501±0.015240966952236005
Final time: 138.94044725100198±5.783009827984033
Final importance metrics:
age : 0.16460926286817648 ± 0.005524110531484962
childs : 0.16113212465756166 ± 0.004479782072516018
living : 0.12400892807430054 ± 0.003952046730688576
w_education : 0.12181559521547615 ± 0.005022358004439373
occupation : 0.11799592892297513 ± 0.006522165158826066
h_education : 0.11250587967867844 ± 0.0022572495557476448
working : 0.07401567077167995 ± 0.0015313258178437264
religion : 0.06414300557925055 ± 0.0012201823880795561
media : 0.05977360423190107 ± 0.0010028402376224088
----------------------------------------------------------------------------

6.NT=100, F=3
----------------------------------------------------------------------------
Final accuracy: 0.5220638153428377±0.010819672403604354
```

```
Final time: 391.33057459195453±3.5647629222050647
Final importance metrics:
age : 0.26637832313707094 ± 0.002072672492439324
childs : 0.1906576109121836 ± 0.0023254864446232157
living : 0.11945595275157776 ± 0.006705745178258568
occupation : 0.10787184375795132 ± 0.004611464695820834
h_education : 0.08884354764377451 ± 0.0023648256457763617
w_education : 0.08635257342276381 ± 0.00038921778519904356
working : 0.06732490388106314 ± 0.001669806513219574
religion : 0.04861296837394513 ± 0.0011685533980543316
media : 0.02450227611966974 ± 0.002802624686576131
-----------------------------------------------------------------------------

7.NT=100, F=int(logM+1)=4
-----------------------------------------------------------------------------
Final accuracy: 0.5241004752206382±0.0034616561531519287
Final time: 487.02581151326496±3.7061459188550705
Final importance metrics:
age : 0.29290031347097695 ± 0.0072251698482793
childs : 0.18213376293640193 ± 0.002992804615682734
living : 0.12020475468244123 ± 0.002998993623895783
occupation : 0.10977473042250425 ± 0.0010327990816734614
h_education : 0.08416059099286917 ± 0.002059743541179012
w_education : 0.0765951696666145 ± 0.004231577062245836
working : 0.0663464027669226 ± 0.00043596869002908195
religion : 0.04807202116526652 ± 0.0026723496264453687
media : 0.01981225389600279 ± 0.0018369933872329485
-----------------------------------------------------------------------------

8.NT=100, F=sqrt(10)=3
-----------------------------------------------------------------------------
Final accuracy: 0.5220638153428377±0.010819672403604354
Final time: 391.33057459195453±3.5647629222050647
Final importance metrics:
age : 0.26637832313707094 ± 0.002072672492439324
childs : 0.1906576109121836 ± 0.0023254864446232157
living : 0.11945595275157776 ± 0.006705745178258568
occupation : 0.10787184375795132 ± 0.004611464695820834
h_education : 0.08884354764377451 ± 0.0023648256457763617
w_education : 0.08635257342276381 ± 0.00038921778519904356
working : 0.06732490388106314 ± 0.001669806513219574
religion : 0.04861296837394513 ± 0.0011685533980543316
media : 0.02450227611966974 ± 0.002802624686576131
-----------------------------------------------------------------------------
```

References

[1] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[2] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[3] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[4] Breiman, L. (2017). Classification and regression trees. Routledge.

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.