IMPLEMENTATION AND PERFORMANCE OF AN OPTIMIZED DATA ANALYSIS USING SUPERVISED DATA MINING

Rasoanaivo Andry[1], Ramafiarisona Hajasoa Malalatiana[2]

[1] *PhD student, TASI, Antananarivo, Madagascar*
[2]*PhD, TASI, ED-STII, Antananarivo, Madagascar*

ABSTRACT

Today, novel techniques for extracting potential and useful information from data are evolving such as the artificial intelligence, machine learning, and data mining. We present in this paper the principle of knowledge discovery in databases and data mining on a large scale of data, and also to know which method is effective between different supervised learning algorithms for classification and prediction, with their precondition optimization. Among these algorithms, which we call " tree bagger " steels precise and powerful, even though the dimension's reduction technique, an effective way to filter the data's relevance, used for a rapid and an unexpansive training from data or retraining from patterns.

Keywords: *Data mining, machine learning, prediction, classification.*

1. INTRODUCTION

Data mining is the extraction of implicit, previously unknown and potentially useful information from a large data set. It uses algorithms based machine learning for extracting information, for the prospect of decision-making. The contribution we provide in this paper is to find the best machine learning algorithm corresponding to a specific bank's data sets and propose a way for optimizing training.

2. PRINCIPLE AND TECHNICS

2.1 Knowledge discovery in databases

It is defined as being a non trivial process in purpose of identifying, in data, a highly comprehensible, valid, novel and potentially useful patterns from databases.

- Process of knowledge discovery in databases

This process aims to transform data into knowledge. This knowledge can be expressed in general concepts' form. The process of knowledge discovery in databases happens with several stages, continuously interrupt by decision-makings by the user.
It summarily requires the preparation of data, the research of patterns, the evaluation of extracted knowledge and their refinement. Then, all repeated in several iterations

- Knowledge discovery in databases tasks

The task represents the goal of the knowledge discovery in databases' process. We can find in practice two high level primary tasks: the prediction and the description.

2.2 Data mining

Data mining is an essential step in the process of knowledge discovery. It is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data [1].

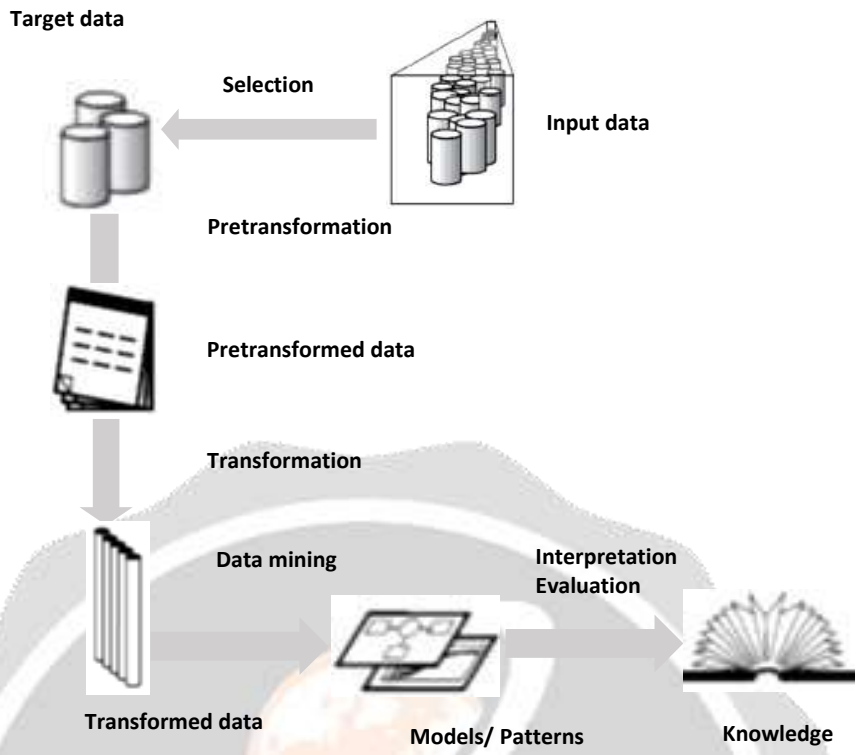

Fig - 1: Knowledge discovery in databases

- Data mining process
  - The different steps involved in the extraction of knowledge from data are:
  - Cleaning of the data.
  - Integrating the data into data warehouses.
  - Selecting task relevant data.
  - Applying data mining to extract patterns.
  - Evaluating the patterns discovered.
  - Presenting the significant patterns to the user [2].

- Data mining tasks

Data mining deals with the kind of patterns that can be mined. On the kind of data's basis to be mined, there are two categories of functions involved in Data Mining: the first is descriptive, and the second is Predictive.

2.3 Machine learning

Machine learning is a field in computer science where existing data are used to predict future data. It is a technique that figures out the "model" out of "data" [3]-[4].
Then, the data that machine learning uses in the modeling process are called training data or training sets [3].

- Principle of machine learning

The main objective of machine learning is to extract and exploit automatically knowledge in a data file. This appears as a model of training for a future usage, prediction or classification of classes

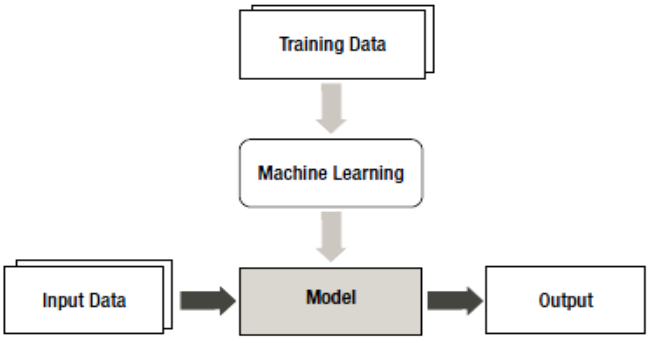

Fig - 2: Machine learning process

- Training

A system mapping an input to an output needs training. Then, the objective of training or learning is to minimize theoretical error between correct output and output from the model for the same input, as the following formulate:

$$ET = \frac{1}{card(\Omega)} \sum_{\Omega} \Delta[Y, \hat{f}(X, \tilde{\alpha})] \quad (1)$$

With:

- ET is the theoretical error,
- Ω is the sample of population,
- $\Delta[.] = \begin{cases} 1, if & Y \neq \hat{f}(X, \tilde{\alpha}) \\ 0, if & Y = \hat{f}(X, \tilde{\alpha}) \end{cases}$

- Types of machine learning

Many different types of machine learning techniques have been developed. Those can be classified into three types depending on the training method mainly: supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning

In this paper, we are focusing on this first case. In beginning, supervised learning includes both “regression” and “classification” algorithms.
In supervised learning, each training sets should consist of input and correct output pairs. The correct output is what the model is supposed to produce for the given input.
Then, supervised learning utilizes two types of variables, mainly: input variables noted X, and output variable noted Y. Therefore, the purpose of supervised learning is to build a classification function which is written as following:

$$Y = f(X, \alpha) \quad (2)$$

Generally, this occurs in two stages: the training of the model on a training set, and test of the model on a test set.

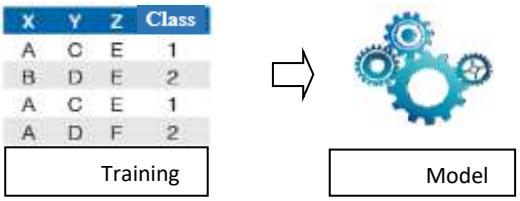

Fig - 3: Training of the model on a training set

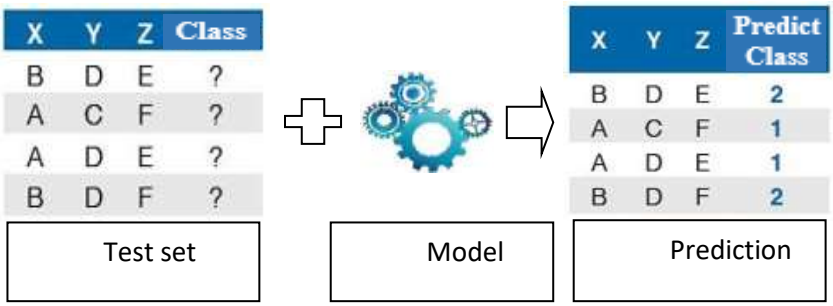

Fig - 4: Test of the model on a test set

- Unsupervised learning

Unsupervised learning does not utilize training sets. It is often used to discover patterns in data for which there is no "right" answer. Clustering algorithms are generally examples of unsupervised learning [2].

- Reinforcement learning

It is based on the concept of how the output should be changed as input changes. The algorithm learns from its evolving environment [3].

3. SIMULATION WITH MATLAB

The aim of this work is to find the best machine learning algorithm, for our data mining prediction and classification, corresponding to a specific bank's data sets and propose solution for optimizing training.
During the experiments, we are choosing to use those supervised machine learning algorithms, such: Generalized Linear Model, Discriminant Analysis, k-Nearest Neighbors, Naive Bayes, Support Vector Machines, Decision Trees, Ensemble learning: Tree Bagger.
These methods are compared, in performance, using confusion matrix, graphs, ROC curve or "Receiver Operating Characteristics" and AUC or "Air Under Curve".
And finally, we were tried to improve training, using a model's reduction technique and then test it with the better algorithms' finding before.

3.1 Data sets

The data sets are related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product, bank term deposit, would be 'yes' or not 'no' subscribed [5]. Its own characteristics are illustrating in Table. I. as following.

Table. I. Characteristics of the data set « Bank-full.csv »

Data Set:	Multivariate	Number of instances:	45211
Attribute:	Real	Number of attributes:	17
Associated Tasks:	Classification	Missing values?	N/A

3.2 Illustrations

Let's start with "data visualization". This process helps to visualize the values of predict variable "y", results of the clients' subscription prediction.

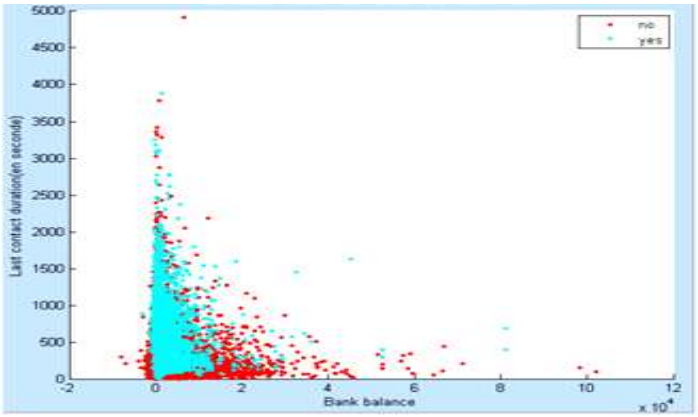

Fig - 5: Clients' subscription prediction

For the preparation of the data, we were using, through this stage, a cross validation technique which is called "hould out". The following Fig. 6. shows the repartition of the data sets: the whole data, training set, and test set.

	Value	Count	Percent
1	no	39922	88.3015
2	yes	5289	11.6985

	Value	Count	Percent
1	no	23940	88.2516
2	yes	3187	11.7484

	Value	Count	Percent
1	no	15982	88.3765
2	yes	2102	11.6235

Fig - 6: Data preparation: (a) initial data (b) training set (c) test set

3.3 Evaluations

- Histogram

The following Fig. 7. illustrates the performance comparison between classifiers, using supervised algorithms, and illustrates that the ensemble learning named "tree bagger" is the most powerful algorithm among them.
To evaluate algorithms robustness, confusion matrix and ROC curve are also the most useful metrics, Fig. 8., Fig. 9.

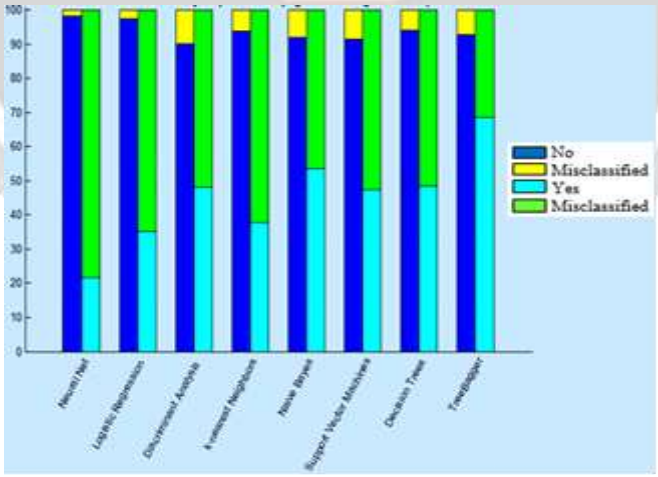

Fig - 7: Result of the classifications comparison

- Confusion matrix

Fig. 8. illustrates also the whole confusion matrix's accurate corresponding to each classifier. And it justifies that tree bagger algorithm "C_tb" steels more accurate.

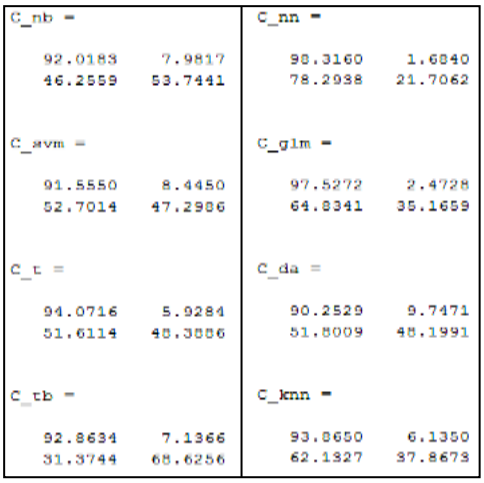

Fig - 8: Result of evaluation: confusion matrix

- ROC curve of tree bagger

In ROC curve, the learner is trying to select samples of test instances that have high proportion of positives. Then, ROC curve plot "true positive" rate on the vertical axis against "false positive" rate on the horizontal axis. With RVP is the true positive rate and RFP is the false positive rate, as describing under:

$$RVP = \frac{True \quad Positive}{True \quad Positive + False \quad Negative} \tag{3}$$

$$RFP = \frac{False \quad Positive}{True \quad Negative + False \quad Positive} \tag{4}$$

Fig - 9: Shows that process is nearly from a perfect prediction

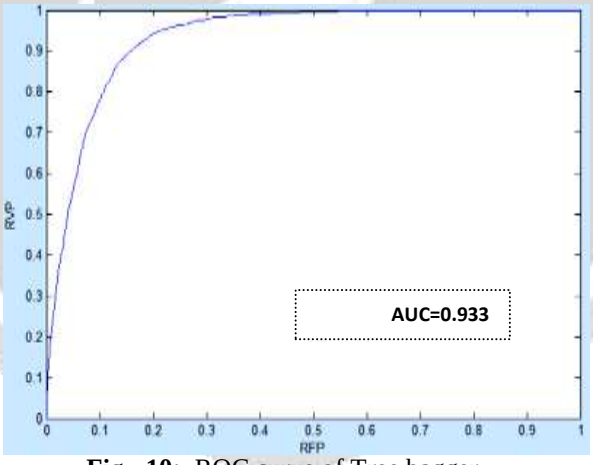

Fig - 10: ROC curve of Tree bagger

3.4 Simplify model

- Features' reduction

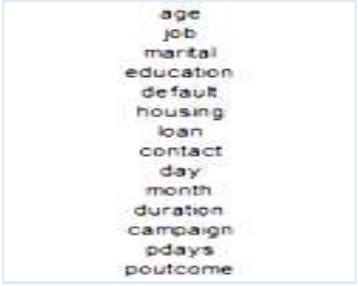

Fig - 11: The list of the remaining features after simplification

To simplify model, a feature selection technique is needed. Fig.10. illustrates the result of the model's simplification. Before, there was 16 features, but after, 2 are removed. Then, dimension of the model was simplifying.

- Evaluations of tree bagger's classification

For checking the robustness of our improving methods, we were classified our model with the last powerful classifier. In Fig. 11., Fig. 12. we find that in spite of the simplification, classification's accurate is always performed.

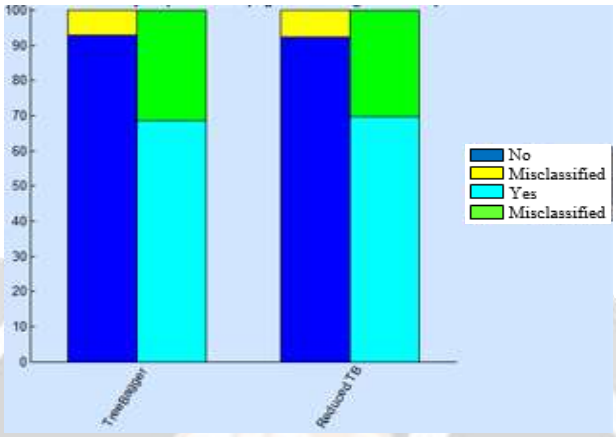

Fig - 12: Evaluation with simplified histogram model

```
C_tb =                          C_tb_r =

   92.8634    7.1366               92.3438    7.6562
   31.3744   68.6256               30.3318   69.6682
```

Fig - 13: Evaluation with confusion matrix

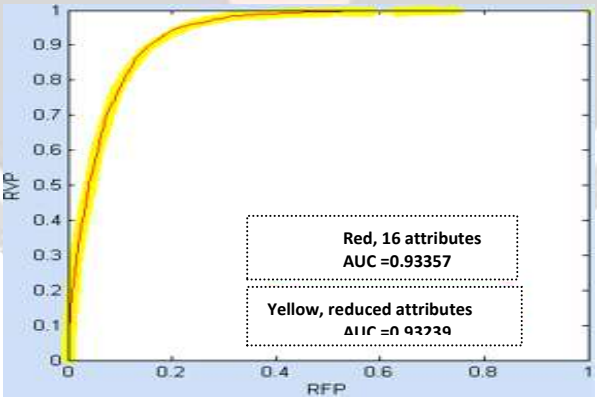

Fig - 13: Comparison of ROC curves

5. CONCLUSION

To summary, classification algorithms of data mining have been successfully applied and compared. However, the accuracy of such methods differ according to the classifier. Identifying the best algorithm is consequently a challenging task. In this study, we have concluded that ensemble learning as tree bagger outperformed with our data sets. After, the next contribution is to provide a reduction's dimension technique, using feature selection, simplifying models without deteriorating classification performance.

6. REFERENCES

[1] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, « Data Mining, Practical Machine Learning Tools and Techniques », ELSEVIER, vol. 4, 2017, pp. 3-88.
[2] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, K. C. Lichtendal, « Data mining for business analytics », WILEY, 2018, pp. 3-147.
[3] P. Kim, « MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence », Apress, 2017, pp. 1-18.
[4] M. Paluszek, S. Thomas, « MATLAB Machine learning », Apress, 2017, pp. 1-85.
[5] S. Moro, P. Cortez, P. Rita, « A Data-Driven Approach to Predict the Success of Bank Telemarketing », Decision Support Systems, Elsevier, 2014, pp. 22-31.
[6] M. R. Amini, « Apprentissage machine, de la théorie à la pratique », Eyrolles, 2015, pp. 63-93.
[7] R. Kumar, « Machine Learning and Cognition in Enterprises », Apress, 2017, pp. 27-65.