International Journal of Computer Engineering and Data Science (ISSN: 2737-8543)
Volume 2– Issue 1, January - March 2023

Tree-based algorithms for classification purposes in Health insurance

Fatima EL KASSIMI[1], Ghita HAJRAOUI[2], Jamal ZAHI[3]

[1] PhD, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco
[2] PhD student, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco
[3]Full Professor, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco

E-mail : f.elkassimi@uhp.ac.ma, ghita.hajraoui@gmail.com, zahi71@hotmail.com.

Article history

Received May 10, 2023
Revised May 19, 2023
Accepted Jun 20, 2023
Published Jun 27, 2023

ABSTRACT

Within a heterogeneous insurance portfolio, not all policyholders are equal in terms of risk; some have a riskier profile than others. Therefore, charging the same premium to all may seem unfair. This heterogeneity can be reduced using risk classes (based on risk factors such as gender, age, or other factors). Given this risk classification, the pure premium for each risk class is estimated using a priori techniques. This emphasizes the importance of risk classification in establishing a fair and reasonable rate structure. This paper aims to classify the insured in terms of risk regarding severities costs. To do so, we used Machine Learning algorithms, namely decision trees and random forests.

Keywords: ***Classification, Health insurance, Machine Learning, Decision trees, Random forests.***

I. INTRODUCTION

The insurance company has the social role of creating solidarity among its policyholders. Using statistical learning for underwriting must not lead to extreme discrimination or "risk personalization," which is one of the consequences of introducing BIG DATA in the insurance industry. Indeed, telematics allows real-time pricing leading to a highly sophisticated pricing system that departs from the principle of risk mutuality in actuarial sciences. Subsequently, generating extremely high premiums for individual profiles that do not cause any risk transfer [1]. Through finding a compromise between insured segmentation and risk mutuality [2], the insurer avoids anti-selection, offering efficient coverage involving risk transfer for all policyholders. The insurance company should be held accountable for its transparency, equity, and solidarity pricing models. It is, therefore, vital to read what is between the lines of the statistical learning algorithms and the resulting pricing models.

In this spirit, the present article sets the objective of adjusting the tariff structure into four classes by acting on the severities model and then classifying the insured according to their risk levels. So that a given insured is assigned to a corresponding risk class while considering the principle of solidarity. The idea is to segment our portfolio so that the risks can be regarded as homogeneous within a given category. Statistical learning algorithms based on trees will be used, essentially decision trees and random forests, in order to propose a classification model.

Our paper is organized as follows: we construct our target variable in section 2 and explain the choice of risk classes in detail. Section 3 focuses on the CART decision tree model and emphasizes the method's principle of application, the validation, and the performance of the model, as well as the interpretation of the results. Section 4 will be dedicated to the random forest algorithm. It presents the method, provides its working principle, and then focuses on the importance of the variables in the classification, while the model performance is given towards the end. Section 5 provides a comparative analysis of the two algorithms. Finally, a short conclusion of the article is given.

II. PREPROCESSING AND CONSTRUCTION OF THE DEPENDENT VARIABLE

The aim is to create group members that are as homogeneous as possible regarding risk. Indeed, the quality of segmentation by risk class can be measured according to four essential criteria: homogeneity, equity, incentive character, and feasibility [3]. The equity criterion assumes that there is no bias between predicted and measured risk and that the contributions paid by insureds of a given class should reflect the losses of that same class. Following the homogeneity criterion, risks within a given class are considered homogeneous, so the category cannot be further subdivided.

A. Study of severities per procedure

Our target variable is the average cost since we are interested in the policyholders' induced charges. The study of the severity model has led to fragmenting our target variable into four classes. The first class comprises low-risk policyholders whose average cost is less than 500 Dhs, followed by the less risky ones whose average cost ranges between 500 and 1 000 Dhs. The third class belongs to the so-called risky policyholders, whose average consumption varies between 1 000 and 5 000 Dhs. The fourth class contains very risky policyholders with an average cost generally higher than 5 000 Dhs.

TABLE I. INSUREDS RISK CLASSES

Risk_S	Risk class	Severity per procedure	Insureds number
R1	Low-risk	< 500	90 215
R2	Less risky	$\geq 500, < 1000$	3 460
R3	Risky	$\geq 1000, < 5000$	2535
R4	Very risky	$\geq 5000 < 280001$	735

In light of what has been discussed, the next step is to create a variable that we will refer to as "Risk_S" using severities data. It is a matter of assigning each of the intervals to a risk class (R1, R2, R3, and R4), where R1 corresponds to low-risk individuals, R2 to those with a medium level of risk, R3 to highly risked insured, and R4 to extremely attempted individuals.

B. Dataset presentation

Our portfolio is managed by a Morrocan mutual health insurance company that is part of the private sector and governed by the National Social Security Fund. The database contains information on 98 000 health insurance claims observed during 2019. The post-processed database has 96 540 rows and 20 variables, out of which six features will be used as classification features in our model.

The table below shows the features that may explain the insured claims experience, including those who do not pay. Each line refers to a single insured. Among its characteristics are the following:

TABLE II. FEATURES SELECTION

Feature	Modalities
Age range	T1 : [0,10[, t2 : [10,20[, t3 : [20,30[, . . ., t7 : [60,70[, t8 : 70 and plus,
Gender	M: male, F: female
Presence of Chronic disease	Y: yes, N: no
PEC_TYP_ BEN	A: insured him/herself; C: insured spouse; E: Insured son or daughter.
Catégorie_Socioprofession	C: single; M: married; D: divorced; V: Widowed.
Nature of the care consumed CAT_LIB.	*Act_dent, surgical procedure, biology, card_inter, rad_vas, cardiovascular, consultation, ..., expl_rad, hospitalization, explore, oncology, rad_interv, radiology,*

C. Cross-validation

K-fold cross-validation is a mechanism that allows us to build the best possible model from an initial dataset and to test its predictive power on a test sample. It randomly divides the training data into k=5 equal dimensions called folds. These folds are created so that no two folds contain the same data points. In other words, they are designed so there is no correlation between them. At each iteration, one of the folds is left out as a validation set, and the model is run on the remaining four folds [4]. Thus, the validation error of the decision tree and random forests is the average of the validation errors made over the five configurations. The “F1-score” evaluation metric will then be calculated for the left out fold.

III. CART DECISION TREES

The CART algorithm has the advantage of being simple to read. This algorithm will be used to diagnose the degree of risk associated with the insureds of our portfolio in the four risk groups detailed in Table 2. Indeed, decision trees are well adapted to the case of a polytomous endogenous variable: "Decision trees give rise to very versatile methods allowing to treat the case of regression, bi-class or multi-class classification similarly or to mix quantitative and qualitative explanatory variables" [5].

A. Presentation and principle of the CART algorithm

The acronym CART stands for "Classification And Regression Trees." It is a statistical algorithm introduced by [6] initially designed for complicated problems involving a series of decisions. Over the years, CART has become a common practice for classification and regression problems. Unlike other algorithms where the constructed classifier is a black box (e.g., random forests and neural networks), the decision tree is commonly represented as readable successive decisions or events. The singularity of the algorithm lies mainly in its simplicity.

B. Tree Construction

The discrimination problem we are trying to solve can be summarized in two steps. First, we select the most discriminating predictors in terms of risk, and second, we construct a rule for assigning the future insureds to one of the four categories. There are many packages in R to build CART decision trees. We have chosen the reference package rpart designed by Therneau in 2009.

The tree construction is done, in general, after dividing the total sample into a learning sample and a test sample whose purpose is to test the learning of the tree. It should be noted that for all the models built throughout this work, we have kept the same approach. The learning base comprises 80% of the initial database, while the test base comprises the rest, i.e., 20% of our database. The comprehensive database of insureds was used to obtain a learning sample of 96 935 insureds by drawing lots. Figure 1 shows the resulting tree.

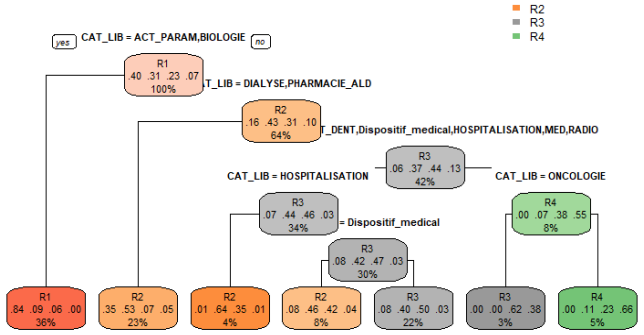

Figure 1 Insureds classification tree by risk class

The present classification tree possesses 7 terminal nodes, each of which is assigned to one of the four risk classes: a node is assigned to a given class if the proportion of insureds belonging to that class is the highest in that node. We thus define an assignment rule associated with this tree. The predictor that best separates the insured according to their risk levels (our target variable) is the category of consumed care noted "CAT_LIB," corresponding to the care item consumed. This variable not only best classifies the insured but, on the same occasion, is the most discriminating feature. Indeed, it is present in each separation. No other variable except for this was used in classification, which reflects its weight in explaining the severities model and, consequently, its influence on the level of risk.

C. *Interpretation and assigning rules*

- Insureds who consume paramedical and biological procedures make up 36% of all insureds, constituting a pure node; these latter belong to the first risk class, corresponding to the low risked class;
- Insureds who use hemodialysis procedures or have a long-term condition belong to the second category, "R2", and are therefore moderately risky. Almost 23% of insureds;
- Insureds who have been hospitalized during the year belong to the second class "R2", i.e., 4% of insureds;
- Insureds consuming medical devices belong to the second class "R2", i.e., over 8% of insureds;
- Insureds who have been hospitalized during the year, those consuming dental, radiology, and medication procedures, belong to the third "R3" class, i.e., 22% of insureds;
- Insureds suffering from cancer whose consumption consists mainly of oncological care belong to the third class "R3", i.e., 3% of insureds;
- Insureds having consumed interventional radiology, vascular radiology, exploratory radiology, nuclear radiology, ANAP procedures, exploratory procedures or have been subject to surgery, belong to the riskiest category "R4", and constitute over 5% of the total number of insureds;

TABLE III. INSUREDS PROFILES PER RISK CLASS

	Class 1	**Class 2**	**Class 3**	**Class 4**
	< 500	$\geq 500, < 1000$	$\geq 1000, < 5000$	≥ 5000
Insured profiles	***36%***	***35%***	***25%***	***5%***
	- Paramedical procedure - Biological procedure	- Hemodialysis - With a chronic disease - Hospitalization - Medical device	-Oncology	- Dental procedures -Radiology procedures - Medication

Source: Authors conception

D. *Confusion Matrix*

Confusion matrices have the advantage of being simple to read and analyze. In addition, they facilitate the interpretation of statistical data with a simple visualization of the results offering the opportunity to diagnose errors, thus allowing for trend identification, which can help reconfigure the parameters.

These matrices are used to measure the performance of classification models. They mainly consist of a table that summarizes the prediction results on a classification problem, illustrating the class distribution of the model predictions and the actual labels. Each row of the table corresponds to an actual class, while each column corresponds to the predicted class instances. In addition, the confusion matrix allows to perform some calculations necessary to evaluate the performance of the model, essentially the:

- Number of positives well classified noted VP, indicating the occurrence of a positive prediction;
- The number of positives classified as negatives noted VN, indicating a negative prediction;
- The number of negatives classified positive noted FP, specifying an incorrect positive prediction;
- The number of well-classified negatives noted FN, indicating an incorrect negative prediction;

These quantities allow us not only to identify the errors made but also to know their types. These data will be used afterwards to calculate the metrics for evaluating the model's performance, which will be discussed in the following section.

TABLE IV. CONFUSION MATRIX

	R1	**R2**	**R3**	**R4**
R1	664	73	51	3
R2	188	419	145	29
R3	46	194	285	30
R4	2	6	26	85

As mentioned before, it is the observations positioned on the diagonal that the algorithm has correctly predicted. So here we have: 664 insureds having been classified as belonging to the R1 class out of a total of 900 insureds, 188 insureds belonging to this class have been misclassified to the second "R2"; nevertheless, we can say that the algorithm manages to classify the insureds of this category correctly. For R2 class,

419 insureds out of 692 were accurately identified, with almost 194 policyholders misclassified to "R3". Unfortunately, the algorithm does not predict this class accurately compared to the first. For the third class, 285 out of 507 insureds were correctly identified, with 145 being falsely assigned to the "R2" class when they belonged to the third. The algorithm seems confused between these two classes and performs less well on their subject—the fourth class, which in turn, contains 85 out of 147 insureds. Finally, the overall true positives (TPs), the sum of the values on the diagonal, amounts to 1 453.

To evaluate the performance of our model concerning the assignment of insureds to the different risk classes, it is necessary to analyze the performance metrics derived from the confusion matrix.

E. Model performance

When evaluating the performance of statistical learning models, the choice of an evaluation metric is essential. Indeed, metrics generally allow for an in-depth comparison of the actual instances to those predicted by the model. Moreover, they make it possible to interpret the predicted probabilities associated with these classes. In this paragraph, we will analyze four metrics that allow us to learn about our model's performance: Accuracy, Precision, Recall, and F1-score.

The evaluation of the performance of a classifier requires the analysis of the confusion matrix.

TABLE V. MODEL EVALUATION METRICS

	R1	R2	R3	R4
Precision	0.84	0.57	0.51	0.71
Recall	0.74	0.60	0.56	0.58
F1	0.78	0.57	0.54	0.64
Prevalence	0.40	0.30	0.22	0.06
Detection rate	0.29	0.19	0.13	0.04
Detection Prevalence	0.35	0.35	0.25	0.05
Balanced Accuracy	0.82	0.69	0.70	0.78

$$\textbf{Accuracy} = \frac{VN+VP}{VN+FP+VP+FN}$$

Accuracy: this allows us to measure the proportion of good predictions compared to the total predictions. The operation is simple: The number of good predictions is divided by the total number of predictions. The weakness of this metric lies in not indicating the model's strengths and weaknesses. This metric cannot be used when dealing with unbalanced classes; it can be misleading because a model with a high Accuracy is not necessarily good. Data scientists refer to this as the "High Accuracy Paradox," which states that predictive models with a given level of Accuracy can have robust predictive power compared to those with a higher Accuracy.

We will therefore be interested in the precision

$$\textbf{Precision} = \frac{VP}{VP+FP}$$

Precision: refers to the number of insureds correctly assigned to a given class relative to the total number of insureds predicted to be in that class. In the case of "R1", this is the number of times an insured was classified as low risk versus the number of times they were misclassified. This metric measures the so-called first-order error, which is the cost of false positives that the algorithm has detected by mistake. For our model, we reached an accuracy of 84% for the first class, 54 and 51, respectively, for the categories "R2" and "R3," which is relatively poor. In contrast, "R4" has a level of 71%. We are also interested in the Recall:

$$\textbf{Recall} = \frac{VP}{VP+FN}$$

Recall: refers to the number of members correctly assigned to their class compared to the total number of insureds in that class, i.e., the total of true positives. It corresponds to the number of times insureds belonging to a class were correctly classified compared to the times they should have been assigned to a given category.

To consider a compromise between these two last metrics, we can calculate the "F-measure," also known as the F1-Score.

$$\textbf{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F-score: subtly combines precision and Recall. It is more interesting than Accuracy because the number of true negatives is not considered. And in cases of unbalanced classes. The F1-Score metric is just the harmonic Mean of Recall and precision.

IV. RANDOM FORESTS

A. Presentation and principle

Generally, more efficient than simple decision trees, the Random Forests algorithm, developed in the early 2000s by [7], is a set of binary decision trees in which randomness has been introduced. It is one of the most widely used supervised learning methods, particularly for its effectiveness in making robust predictions. However, this method also introduces interaction relations between the predictors and the dependent variable during estimation [8]. Thus, one of the specificities of this algorithm is that it makes it possible to detect the importance of the predictors in estimating the dependent variable, which is why we are interested in this algorithm. Indeed, measuring feature relevance, particularly "CAT_LIB" in the framework of Random Forests, will allow us to extend the analysis made with the CART method.

The robustness of Random Forests and their performance in everything related to regression and classification makes this algorithm multidisciplinary. As a result, this method has become widespread in different fields, such as marketing [9], chemoinformatics [10], and object recognition [11], to name a few.

There are several models of decision trees, each of which uses a particular method or heuristic for choosing the splitting modality. In the context of random forests, [7] proposes to use

the CART model trees that select predictors according to the GINI index. These random forests can be intrinsically capable of handling multi-class problems [12].

Random forests are used in classification and regression frameworks; with only a slight parameterization, one can switch from one framework to another. In addition, they have an excellent predictive performance, in terms of the generalization error, of course, in both cases [5].

B. *Generalization error: Out-Of-Bag (OOB) error*

Out-Of-Bag (OOB) error means "outside the bootstrap." Random Forests always give an estimate of the generalization error of the algorithm. The approach to computing this error is as follows: Consider a member defined by $(X_i ; Y_i)$ the training sample, where X_i represents its characteristics (the explanatory variables available to us: age group, sex, ALD risk,...) while Y_i represents its risk class (R1 R2, R3, R4), and consider the set of trees obtained on the bootstrap samples not including this member, i.e., for which this member is "Out-Of-Bag." We subsequently join only the predictions of these trees to make our estimate Y_i. This is done this way for the training sample data set, after which we calculate the error rate, which corresponds to the OOB error of the Random Forests algorithm. This is because the predicted algorithm has yet to encounter these members.

TABLE VI. CONFUSION MATRIX AND "OUT OF BAG" ERROR

	R1	R2	R3	R4	Class Error
R1	2714	740	145	1	0.25
R2	324	1713	700	31	0.38
R3	207	705	1030	86	0.49
R4	11	120	170	287	0.51

The confusion matrix above gives the error rate associated with each risk class. For instance, the error rate for "R1" is 24%, so we can say that by classifying an insured taken randomly from the learning base, we risk making a wrong assignment in 24 cases out of 100. This rate is 38% for "R2" and 49% for "R3". In comparison, it reaches its maximum for the class "R4" with an error rate of about 51%, specifying that by randomly classifying an insured drawn from the learning base, we risk in 51% of the cases to misclassify him. The RF algorithm cannot distinguish between the last two classes, and they should be merged to obtain better results.

C. *Parameterization of the Random Forests*

To build our random forest algorithm, we propose to use the package "randomForest" introduced on the open-source R by [13] et [14]. This package has two essential parameters, it is, of course, the parameters 'mtry' and 'ntree':

- "*mtry*" corresponds to the number of predictors chosen by chance at the nodes of the trees. The latter can take 1 to *p* values, where p is the set of variables in the learning base. Generally, it is set by default to the threshold p=3 when it is a regression and to √p when it is a classification.
- "*ntree*" refers to the number of trees in the forest. Where 500 is the default setting. The optimal choice can be made after many experiments.

The "randomForest" package also allows one to search for the number of variables needed for each division.

D. *Importance of predictors*

Unlike the CART algorithm, the final output is not a tree. Random forests are part of the so-called black box models, whose major drawback is the absence of a direct interpretation of the results. Nevertheless, intermediate results can be analysed to improve the algorithm's prediction. It is thus possible to see the importance of each variable, which is a powerful tool for RF, a tool that has even been used in many works just for variable selection. Some of these works include [12], who present a variable selection approach based on this tool, [15] present a procedure for selecting essential variables using random forests.

In this spirit, we will analyze each variable's contribution to the constitution of the model, an illustration of which is given in Figure 2 and Table 8. These intermediate results will be used to improve the classification of the members of our portfolio. Nevertheless, first, the algorithm allows us to visualize the feature importance and, thus, their contributions to the classification.

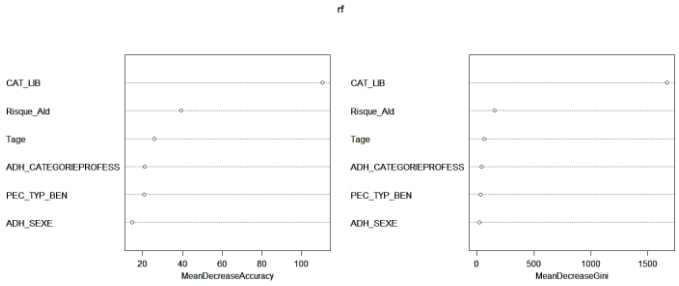

Figure 2 Features importance in the classification using Random forest

Thus, we can see that the random forests rank the variable CAT_LIB as the most critical variable in the classification, consistent with the results observed in the CART algorithm. Figure 2 reinforces this finding, which goes to confirm the inescapable effect of the nature of the consumed item of care on determining the insured's risk class. The variables sex of the insured "ADH_sexe" and type of beneficiary "PEC_TYP_BEN" are less relevant in constructing the random forest learning algorithm. Therefore, regardless of beneficiary sex or type, it is the nature of the consumed care, for the most part, and the state of health, in part, that decides the level of risk.

The random forest algorithm thus suggests two criteria to evaluate the relevance of the variables in the classification, namely:

1) Average decrease in Accuracy

The Mean Decrease Accuracy [16] is the average decrease in the margin of out-of-bag observations when the variable is randomly perturbed. It consists of an indirect evaluation of the predictors' impact on the classification quality. The more attributes that make the classification of a predictor, the more important this one is. In other words, this criterion expresses the degree of precision the model loses by excluding each variable. Conversely, the lower the precision, the more influential the variable is for the success of the classification.

2) Mean Decrease in Impurity

Also called Mean Decrease Gini, this criterion was introduced by [6]. The CART tree induction method is based on the decrease in heterogeneity or impurity measured by the Gini index. The total decline in heterogeneity related to a variable corresponds to the reduction of the cumulative entropy over the total number of nodes it divides. This reduction is then averaged over all the trees. The importance of a variable in the classification corresponds to the average reduction in impurity that it allows. For each tree, the objective is to reduce the Gini index and thus increase the purity of the resulting nodes. The higher the Mean decreased Gini score, the greater the importance of the variable in the model. The disadvantage of this criterion is that it favors continuous or multi-attribute variables. This is because these types of predictors are selected mainly by the CART model [17],[18],[19].

According to [16], these two criteria are similar in that a predictor's relevance is related to its frequency of occurrence and position in the tree construction.

TABLE VII. EVALUATING FEATURES WEIGHT IN THE CLASSIFICATION USING THE RANDOM FOREST ALGORITHM

	Risk class				Mean Decrease Accuracy	Mean Decrease Gini
	R1	*R2*	*R3*	*R4*		
ADH_SEXE	-7.3	16.4	0.3	7.5	14.9	20.9
Tage	-15.1	19.8	6.3	14.3	25.9	62.8
CAT_LIB	109.7	54.4	57.4	57.7	110.4	1669.4
ADH_CATEGORIEPROFESS	-9.01	12.0	7.3	11.5	20.9	39.5
PEC_TYP_BEN	-5.26	20.0	1.65	10.5	20.73	28.9
Risque_ALD	19.3	28.6	18.2	17.9	39.1	152.9

The classification of our variables according to these two criteria leads to Table 8, where we can see the "CAT_LIB" variable with a very high score, according to these two criteria, respectively, which is not surprising given its discriminatory power demonstrated by the CART algorithm. Then comes the variable "risk_ALD" in second place, even if it did not appear in the divisions of the CART algorithm due to the variable "CAT_LIB" dominance; however, its weight remains enormous, and its contribution remains undeniable in the discrimination of the members. Followed by the variable "Tage," whose contribution remains constant. For the rest of the predictors, their importance remains modest. Nevertheless, the variables "PEC_TYP_BEN" and "ADH_CATEGORIEPROFESS" are well ranked by the Random Forests compared to the "ADH_SEX" variable.

E. Model performance

1) Confusion matrix

TABLE VIII. CONFUSION MATRIX PRIOR TO MODEL CALIBRATION

	R1	**R2**	**R3**	**R4**
R1	671	193	35	1
R2	83	433	174	2
R3	51	171	260	25
R4	3	30	37	77

From the present confusion matrix and the one in Table 5, we can see a minimal improvement in the F-score due to the dominance of a single variable. It was impossible to further improve the model, as the parameters were already well configured, and an F1-score of 62% is acceptable.

2) Computational analysis of model performance

TABLE IX. CLASSIFICATION MODELS PERFORMANCE

	Machine Learning Algorithm	
	CART	***Random Forest***
F1-score Average	0,6325	0,6200

The performances of the machine learning models regarding insured classification are gathered in Table 10. It should be noted that the degradation of the F1-score associated with the Random Forest amounts to 1.2% compared to the CART model. However, it is limited if we notice that the F1 scores are almost the same magnitude.

V. CONCLUSION

Statistical learning algorithms were used to ensure both the profiling and classification of the portfolio members studied by risk class. In addition, CART decision trees are compared to Random Forest. These two learning techniques were used to classify the members given their characteristics; each insured was assigned to a risk class according to his characteristics. As a result, four risk classes were defined, each corresponding to a different level of risk, ranging from the low-risk class to the one with a high risk of causing considerable future losses. This classification allowed us to: first, segment the insured in terms of risk while determining the profile associated with each risk class, thanks in particular to the CART algorithm; second, the features importance tool of the Random Forest algorithm allowed us to confirm the relevance of the care consumed feature in the determination of the level of risk inherent to the members. Thirdly, this classification will make it possible to match each risk class with an equivalent tariff class so that members of a risk class pay the same tariff, which answers the question of tariff equity. Finally, correcting and refining the mutual insurance company's tariffs would be possible. However, this does not exclude a slight mutualization (in percentage, for example) among these classes.

REFERENCES

[1] O'Neil, C. (2016). Weapons of Math Destruction. new york: Crown.

[2] Charpentier, A., Denuit, M., & Elie, R. (2015). SEGMENTATION ET MUTUALISATION LES DEUX FACES D'UNE MÊME PIÈCE ? Risques n° 103, 19-23.

[3] Feldblum, S., & Schirmacher, E. (2006). Financial Pricing Models for Property-Casualty Insurance Products: Retrospective Analysis. North American Actuarial Journal 10(2), 1-27.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

[5] Genuer, R., & Poggi, J. M. (2017). Arbres CART et Forêts aléatoires Importance et sélection de variables. HAL Id: hal-01387654.

[6] Breiman, L., Friedman, J. H., Olshen, R. A., & J. Stone, C. (1984). Classification And Regression Trees. Boca Raton: Routledge.

[7] Breiman, L. (2001). Random Forests. Machine Learning volume 45, 5–32.

[8] Rodenburg, W., Heidema, A. G., Boer, J. M., Bovee-Oudenhoven, I. M., Feskens, E. J., Mariman, E. C., & Keijer, J. (2008). A framework to identify physiological responses in microarray based gene expression studies: selection and interpretation of biologically relevant genes. the American Physiological Society.

[9] Lariviere, B., & Poel, D. V. (2005). Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques. Expert Systems with Applications 29(2), 472-484.

[10] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. American Chemical Society, 1947–1958.

[11] Shotton, J., Fitzgibbon, A., Cook, M., & Sharp, T. (2011). Real-Time Human Pose Recognition in Parts from Single Depth Images. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 56(1), 1297-1304.

[12] Díaz-Uriarte, R., & Andrés, S. A. (2006). Gene selection and classi microarray data using random forest. BMC Bioinformatics,7, 1-13.

[13] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News.

[14] Liaw, A., & Wiener, M. (n.d.). Random Forest: Breiman and Cutler's Random Forests for Classification and Regression. Retrieved octobre 10, 2021, from https://cran.r-project.org/package=randomForest

[15] Hapfelmeier, A., Hothorn, T., & Ulm, K. (2012). Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Computational Statistics & Data Analysis, 56(6), 1552-1565.

[16] Breiman, L. (2003). Setting up, using, and understanding random forests V4.0. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

[17] Nicodemus, K. K., & Malley, J. D. (2009). Predictor Correlation Impacts Machine Learning Algorithms: Implications for Genomic Studies. Bioinformatics 25(15):1884-90.

[18] Nicodemus, K. K. (2011). Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in bioinformatics ,Volume 12, Issue 4.

[19] Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing.