Journal of
Engineering and
Applied Sciences

SINCE 2000
MEDWELL PUBLICATIONS

A Random Forest Classifier for Digital Newspaper Readers

[1]Enrique De La Hoz-Domínguez, [2]Adel Mendoza-Mendoza and [2]Daniel Mendoza-Casseres
[1]*Department of Industrial Engineering, Universidad Tecnológica de Bolívar, Cartagena, Colombia*
[2]*Department of Industrial Engineering, Universidad del Atlántico, Puerto Colombia, Colombia*

Key words: Random forest, classification, newspapers, supervised learning, recommender systems

Corresponding Author:
Adel Mendoza-Mendoza
Department of Industrial Engineering, Universidad del Atlántico, Puerto Colombia, Colombia

Page No.: 3668-3673
Volume: 15, Issue 22, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences
Copy Right: Medwell Publications

Abstract: In this research, the potential of machine learning methods based on Decision Trees (DT) and Random Forest (RF) models is developed in the context of classifying readers of a digital newspaper. For this purpose, the number of visits of users to each section of the newspaper in a 6-month interval has been taken into account. The models of DT and RF developed in this study, classify the profiles of readers who access the journal with an accuracy of 98.07% and AUC value of 99.27%, thus, demonstrating that it serves as a valid tool for making strategic and operational decisions when creating, manage and present content in the user-website interaction.

INTRODUCTION

Machine Learning (ML) is a branch of Artificial Intelligence (AI) related to the ability to create data-based models to identify hidden patterns and non-contextual information about a phenomenon without the need to establish the intrinsic relationships that characterize the problem. Thus, ML algorithms are able to progressively improve their performance by counting a greater amount of data to ’learn’ as well as being able to discover hidden patterns in complex, heterogeneous and high dimensional data sets[1-3]. Consequently, the ML has become the key technology for the development of many real applications in different fields: from the prediction of complex diseases, the bankruptcy forecast for companies, the internet search engines, educational data mining and the computer vision[4-6].

In the field of client management in virtual environments, there are machine learning algorithms, such as Logistic Regression (LR), Gradient Boosting Machines, Support Vector Machines (SVMs), Decision Trees (DTs) and Random Forests (Rfs) which are able to relate variables of non-linear and heterogeneous inputs to a pattern and response, even when relationships between model variables can not be determined due to their complexity, high variability or lack of businesssense[7-9]. Machine learning models are widely used by large Internet-based companies which have the capabilities and resources to develop data collection and modeling. However, ML models are rarely seen in the context of small businesses, the focus of study of this research is a regional digital newspaper, categorized as a Small and medium-sized enterprises (SME’S) where the volume of readers is not very large and don not have a large number of subscribers which could mean not having a sufficient number of training samples which in turn could compromise the learning process of the algorithm. Among several classification algorithms, the DT have characteristics that are particularly suited to the process of classification of website users. The DT can be understood intuitively, even without having a statistical or mathematical training in addition to being able to generate a visualization of the results, thus, reinforcing the understanding of it. On the other hand, DTs are able to deal with missing NA values and can combine categorical and numeric data in the same model while developing a selection of main characteristics in parallel to the modeling.

The transformations in the way people keep themselves in- formed, associated with the revolution of social networks and the excessive supply of information, forces the digital media to understand the behavior of its readers in order to be competitive[10,11]. The dynamics of business in the digital world requires digital newspapers to understand the behavior of their readers. So, through the present research the following research questions are answered. How to identify the key variables in the consumption flow of the digital newspaper reader? How to define a machine learning model to classify digital newspaper readers? How to graphically represent the profiles of readers of digital newspapers to have a comprehensive perspective? In correspondence with what has been previously proposed, a method of classifying readers of digital newspapers is presented, identifying the significant newspaper sections and the representative classification of the readers according to the use of the website.

MATERIALS AND METHODS

Data and methods: The 689 readers who have interacted with the newspaper in the period from January to June 2019 are included in the study. The number of visits to each section of the newspaper (Table 1) was analyzed by each reader to identify the intensity of use of the website. Thus, the average activity in each section has been used to define standardized vectors of behavior which are independent of the number of visits of the newspaper. In order to guarantee legitimate results, only the information corresponding to users who at least visited the newspaper six times in the last three months has been used.

Observations corresponding to 70% (482) of the total data used in the study were randomly selected for the training phase of the model and the remaining 30% (207) of readers were subsequently used as evaluation elements. The new datasets developed for the cross-validation process were used for the training of DT decision tree models and random forest models.

Table 1: Summary of predictor variables

Section	Min	Max	Mean	SD
Front page	0	55.5	16.70	12.1
Politics	0	44.4	24.60	11.1
Economy	0	20.0	4.20	5.4
Sports	0	25.0	5.62	6.5
Culture	0	25.0	6.13	6.8
Interview	0	33.3	13.10	6.5
Opinion	0	44.4	17.40	9.5
International	0	44.4	19.20	8.9
Video	0	10.0	65.00	31.6

Output variables: The output variables of the classification process represents reader's profile: Visual, informed and NetNEE as described by Dominguez *et al*.[12].

Visual profile: Readers with a high utilization rate of videos and little relation with the contents of reading. This profile resembles those known in the literature as digital natives, those who prefer the graphics to the texts; they use external short-cuts to access the web and frequently share information with their friends on social networks[13].

Informed profile: They are characterized by the widespread use of the sections of them newspaper, show a global interest and be at day of what happens in their environment; this group represents the 50.5% of the total sample of users used for this study. This profile can be likened to that of Immigrants digital who prefer sequential processes, in social terms it is as if they learn a new language, culture and communication approach.

NetNee profile: Responds to a behavior of little interest in the contents of the newspaper in their visits to the web. It hardly interacts with the other sections, its Entrance to the web is done through external platforms, such as social networks or forums, which shows that, at the first moment He was attracted by the information but when he enters the Web, he leaves immediately. According to Hernandez *et al*.[14], this profile could be similar to the sniffer visitor, a silent participant with a passive activity being there, reading, watching the messages in the forums, stalking but in no way contributes nor comment on the generated discussion.

Data exploratory analysis: We used Principal Component Analysis (PCA) to deliver a visual representation of the data. Prior to PCA, the data was arranged by the frequency of visiting and use of the newspaper's sections. Only the two first Principal Components (PCs) were selected as they represent the 80.4% of the information. Figure 1 shows a visual representation of the reader behavior in a new orthogonal space composed of the two PCs above mentioned. On one hand, points on the plot are the readers and their colors are related to their profile. On the other hand, information related to the level of contribution of each variable is highlighted with colors being green the maximum.

The larger the value of the contribution, the more the variable contributes to the selected principal components. Thus, the variable which provides less information is Interview. The informed group (Green points) are located over the horizontal axis in the growing direction of the sections Main, Sports, Culture, Economy, International,

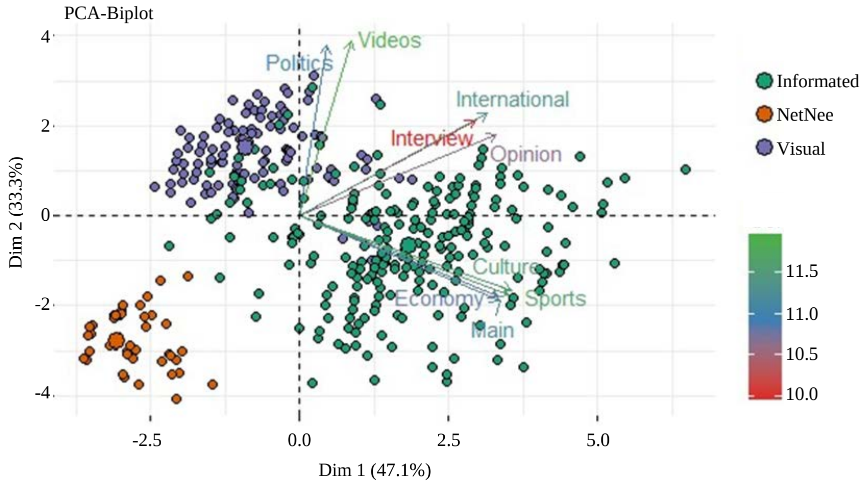

Fig. 1: Exploratory analysis based on principal components for the user's behavior

Opinion and Inter-view, this location maximizes the relations of these users with the newspaper, representing high visiting frequency and inter-action. In the other hand, the visual readers (purple points) al-located on the top of the vertical axis interact with the videos and Politics sections but in the lower position of the growing vectors. In the opposite, the NetNee users (orange points) are located in the third quadrant of the plane, this location minimizes the majority of the sections since they are in the opposite direction of growth of them, representing low interest to interact with the newspapers.

Decision tree: Decision Trees (DT) are machine learning models that resemble the shape of a tree where the leaves represent the output categories and the branches represent the partitions of the predictive variables that determine the results of classification or regression.

Design of the decision tree in this research: The decision tree implemented is based on the architecture of the "CART" algorithm developed using the R part package of RSTUDIO[15]. During the training process, through the cross-validation technique, the data set was divided repeatedly, creating 10 new data sets, thus generatinga trial and error process for the characterization of the parameters. For the evaluation of the models, the Gini Diversity Index (GDI) was taken into account as an optimization criterion for the models. The GDI measures the level of impurity of each node, therefore, a node will be considered pure when all the observations belong to the same category; The GDI of a pure node is equal to[16].

The design process of the decision tree has beendeveloped in the following way: The minimum number required for the creation of a participation node is equal to 10 and at least 1 observation for a response node. The tree creation procedure was repeated 500 times and each time a different subset of data was tested. Beforehand, it was expected to observe a high variability between the performance of these 500 DT.

Pruning: An alternative to avoid overfitting in decidingtrees is pruning, which consists in examining the nodes that have less effect in the general classification[17]. In this research, pruning process was applied to penalize the complexity of the decision tree, ensuring significant partitions were involved in the model.

Random forest: Random forest models are methods articulated between machine learning algorithms, which consist in the recurrent and growing construction of multiple decision trees through a process of aggregation by bootstrapping[18]. In other words, multiple decision trees are created with different compositions of variables in such a way that each tree yields an independent result, to then carry out a process of democracy where the category most voted by the trees is determined as the final output (Fig. 2).

This characteristic, of generating separate answers for each decision tree and then joining them in a general prediction, produces robust models, less prone to extreme values than asimple decision tree, thus, improving the prediction and classification capacity of the model. The RF model presents a variable selection technique, so, it can handle data sets with a large number of variables if it is necessary to use previous processes to reduce dimensions. In addition, the model allows to identify the importance of each variable for the correct classification of observations, through a permutations test. The Random Forest model used consists of 500 trees created under the

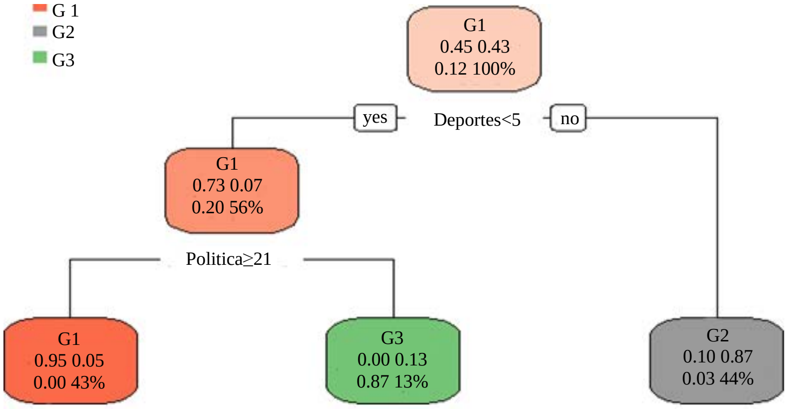

Fig. 2: Schema of DT with 3 branch nodes and 3 leaf nodes

following guidelines. We considered 3 as the minimum number of observations that give rise to aresponse node. The number of variables used to create the trees was evaluated from 4-8.

Evaluation performance: The success resulting from the classification process is given by the difference between the predicted value and the real value. This relationship is de- scribed by the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) metrics. The metrics used to evaluate the performance will be the correct Classification rate (C), Kappa (k) and the area under the ROC curve or AUC[19]. The area under the curve represents the rate of Tp and FP at various discrimination thresholds. A model with a perfect classification will have an AUC = 1. On the other hand, a completely random model will have a value of AUC = 0.5.

RESULTS AND DISCUSSION

Decision tree: The decision tree created in Fig. 1 was developed after the creation of 500 trees, trained in different subsets of data, exchanging the roles of training and evaluation among them. The decision tree was able to correctly predict the type of reader with 93.1% for training and 88.1% for the test. Analyzing the result of the test, it was found that the decision tree DT is able to identify reader types G1-3 with 81, 100 and 90.4% sensitivity, respectively. Of the 9 sections of the diary used for the classification process, the decision tree identifies the “sports” and “policy” sections as the key discrimination variables, none of the remaining 7 sections has been used by the decision tree model to predict the profile of the reader.

Random forest results: The random forest model, builton 500 trees, showed an accuracy of 98.55% during the training phase and successfully classified 97.10% of the test cases, improving the performance of the decision tree as shown by Breiman[20]. The evaluation of the specificity shows values of 100, 94.68 and 100% for the profiles, Visual, Informed and NetNee, respectively and a global value of AUC_TEST = 99.27%.

Table 2: Performance metrics of DT and RF models

	DT		RF	
Metric	Training	Test	Training	test
Accuracy (%)	0.9376	0.8696	0.9751	0.9807
Kappa	0.8963	0.7899	0.9588	0.9685
Area under the ROC curve, AUC	0.9724	0.0.9519	0.9967	0.9927

The RF model was replicated 10 times to determine the consistency of the model with respect to the decision tree. The results show a reduced variability (sigma = 0.006) and a better overall performance (Fig. 3 and Table 2).

The level of the AUC performance metric of 99.27%, achieved by the model for the test phase demonstrates the relevance of the machine learning model presented in this research to be replicated and reproduced in other web user classification environments. The results obtained are similar to those found by Adeniyi *et al.*[21] in their study on the classification of web users using the KNN algorithm, their object of study was the Really Simple Syndication system (RSS), achieving 70% accuracy in the quality of the recommendations, this means the level of fit of the news recommended to the user according to their immediate requirements.

The Random Forest model yielded a robust result, significantly improving the performance of the DT-based model. The structure of the model makes it possible to identify the visual, informed and NetNee reader profiles with very high precision. The robustness of the model allows its implementation as a system for recommending content to users of a digital newspaper, generating a

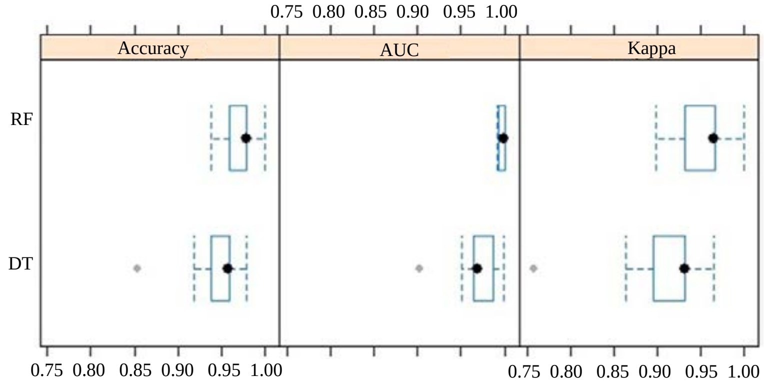

Fig. 3: Boxplot illustrating the results for the three metrics selected

continuous learning process for updating profiles according to the interaction of readers with the website.

It is important to express the particularities to the development of this study which makes it interesting for small digital businesses which do not have the resources or the infrastructure necessary to access advanced recommendationsystems. Therefore, the use of this model by other means of digital communication and the comparison of results will generate a scalable model.

As far as the researcher's knowledge is concerned, the scientific community is presented with a novel model for the classification of readers in a digital newspaper, usingprofiles associated with the frequency and use of newspaper sections as classification variables. The predictions reached by the DT and RF models are generated from the real behavior of readers, resulting in a robust model for making strategic and operational decisions in the administration of a digital newspaper.

CONCLUSION

The process of interaction between users and the website of the newspaper has been modeled effectively, using the 9 sections of the newspaper as predict or variables, using the frequency of visit and use of the sections to make up the dataset of 489 users. The policy and sports sections are the ones with the greatest discrimination capacity to determine user profiles: Visual, informed and NetNee.

The model identifies users with a percentage of use of the sports section <5% and a use of the policy section >21% as a reader of the "Visual" profile with a 95% probability. Those readers with a percentage use of the sports section >5% are classified in the "Visual" profile with an 87% probability. Users with a consumptionof the policy section >21% and sports <5% will be classified in the NetNee profile.

The Random Forest model consistently presented better results than the Decision Tree for the three-performance metrics used. The model created from 500 trees yielded a performance of 98.07, 96.85% and 0.9927 for accuracy, Kappa and AUC, respectively, representing a robust, replicable and reproducible model.

REFERENCES

01. Obermeyer, Z. and E.J. Emanuel, 2016. Predicting the future-big data, machine learning and clinical medicine. New Engl. J. Med., 375: 1216-1219.
02. Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM. SIGMETRICS. Perform. Eval. Rev., 41: 70-73.
03. Yu, Q., Y. Miche, E. Severin and A. Lendasse, 2014. Bankruptcy prediction using extreme learning machine and financial expertise. Neurocomputing, 128: 296-302.
04. Kourou, K., T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis and D.I. Fotiadis, 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J., 13: 8-17.
05. Mahdavinejad, M.S., M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi and A.P. Sheth, 2018. Machine learning for internet of Things data analysis: A survey. Digital Commun. Networks, 4: 161-175.
06. Beaulac, C. and J.S. Rosenthal, 2019. Predicting university students' academic success and major using random forests. Res. Higher Educ., 60: 1048-1064.
07. Stalidis, G., D. Karapistolis and A. Vafeiadis, 2015. Marketing decision support using artificial intelligence and knowledge modeling: Application to tourist destination management. Procedia-Social Behav. Sci., 175: 106-113.

08. Sundsoy, P., J. Bjelland, A.M. Iqbal and Y.A. de Montjoye, 2014. Big data-driven marketing: How machine learning outperforms marketers gut-feeling. Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, April 1-4, 2014, Springer, Washington, USA., pp: 367-374.
09. Erevelles, S., N. Fukawa and L. Swayne, 2016. Big data consumer analytics and the transformation of marketing. J. Bus. Res., 69: 897-904.
10. Allcott, H. and M. Gentzkow, 2017. Social media and fake news in the 2016 election. J. Econ. Perspect., 31: 211-236.
11. Kumpel, A.S., V. Karnowski and T. Keyling, 2015. News sharing in social media: A review of current research on news sharing users, content and networks. Social Media Soc., Vol. 1, 10.1177/2056305115610141
12. Dominguez, E.D.L.H., M.A. Mendoza and O.D.L.H. Hoz, 2017. Classification of readers profiles of a digital journal. Revista Udca Actualidad Divulgacion Cientifica, 20: 469-478.
13. Ahn, J. and Y. Jung, 2016. The common sense of dependence on smartphone: A comparison between digital natives and digital immigrants. New Media Soc., 18: 1236-1256.
14. Hernandez, D.H., R.A. Martinell and D. Cassany, 2014. [Categorizing users of digital systems (In Spanish)]. Pixel-Bit. Media Educ. Mag., 44: 113-126.
15. Therneau, T. and E. Atkinson, 2018. An introduction to recursive partitioning using the RPART routines. Mayo Foundation for Medical Education and Research, Arizona.
16. Coppersmith, D., S.J. Hong and J.R. Hosking, 1999. Partitioning nominal attributes in decision trees. Data Min. Knowl. Discovery, 3: 197-217.
17. Rokach, L., 2016. Decision forest: Twenty years of research. Inf. Fusion, 27: 111-125.
18. Breiman, L., 2001. Random forests. Mach. Learn., 45: 5-32.
19. Handelman, G.S., H.K. Kok, R.V. Chandra, A.H. Razavi and S. Huang et al., 2019. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. Am. J. Roentgenology, 212: 38-43.
20. Breiman, L., 2017. Classification and Regression Trees. CRC Press, Boca Raton, Florida, USA.,.
21. Adeniyi, D.A., Z. Wei and Y. Yongquan, 2016. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Comput. Inf., 12: 90-108.