A Recommender System for Digital Newspaper Readers Based on Random Forest

Enrique Delahoz-Dominguez[1], Rohemi Zuluaga-Ortiz[2](✉), Adel Mendoza-Mendoza[3], Jey Escorcia[4], Francisco Moreira-Villegas[5], and Pedro Oliveros-Eusse[6]

[1] Engineering Faculty, Universidad Tecnológica de Bolívar, Cartagena, Colombia
edelahoz@utb.edu.co
[2] Engineering Faculty, Universidad del Sinú, Cartagena, Colombia
rohemi.zuluaga@unisinu.edu.co
[3] Engineering Faculty, Universidad del Atlántico, Barranquilla, Colombia
[4] Faculty of Business Sciences, Universidad de la Costa, Barranquilla, Colombia
[5] Faculty of Natural Sciences and Mathematics, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
[6] Faculty of Social Sciences and Humanities, Universidad de la Costa, Barranquilla, Colombia

Abstract. In this research, the potential of machine learning methods based on decision trees (DT) and Random Forest (RF) models 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 3-month interval has been taken into account. The models of DT and RF developed in this paper 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.

Keywords: Random Forest · Classification · Newspapers · Supervised learning · Recommender systems

1 Introduction

Machine learning (ML) is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment [1, 2]. Also, it's a branch of artificial intelligence (AI) that creates data-based models to identify hidden patterns and non-contextual information about a phenomenon without establishing the intrinsic relationships that characterize the problem [3]. Thus, machine learning algorithms can gradually improve their performance by counting a more significant amount of data and uncovering hidden patterns in complicated, heterogeneous, and high-dimensional data sets [4], these types of algorithms automatically alter or adapt their architecture through repetition, so they become better at achieving the desired task [1]. Machine Learning is gradually evolving, is, and will be a potential game-changer

K. Saeed and J. Dvorský (Eds.): CISIM 2022, LNCS 13293, pp. 191–201, 2022.
https://doi.org/10.1007/978-3-031-10539-5_14

in the history of computing, logical algorithm patterns, and the design of complex data structures [5]. Consequently, the ML has become the key technology for developing many real applications in different fields: from predicting complex diseases [6], the bankruptcy forecast for companies [7], the internet search engines [8] educational data mining [9–11], speech recognition and computer vision [12].

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 can relate variables of non-linear and heterogeneous inputs to a pattern and response, even when relationships between model variables cannot be determined due to their complexity, high variability or lack of business sense [13–15].

Nowadays, sizeable Internet-based companies use Machine learning models, which have the capabilities and resources to develop data collection and modeling. However, it isn't common to find ML models in the context of small businesses. Our research is designed in a regional digital newspaper, categorized as Small and medium-sized enterprises (SME'S), and characterized by a small volume of readers and a low proportion of subscribers. The above features could compromise the algorithm learning process and make unsuitable recommendations for a machine learning model. Among several classification algorithms. As it is cited in the work of Akinsola [16] Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a value that the node can assume. Also, instances are classified starting at the root node and sorted based on their feature values. Decision tree learning, that are used in Data Mining and Machine Learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Moreover, the DT has characteristics that are particularly suited to classifying website users. The DT can be understood intuitively, even without statistical or mathematical training, and can visualize the results, thus reinforcing its understanding. On the other hand, DTs can deal with missing NA values and combine categorical and numeric data in the same model while developing a selection of main characteristics parallel to the modeling.

The transformations in the way people keep themselves informed, associated with the revolution of social networks and the excessive supply of information, force the digital media to understand the behavior of its readers to be competitive [17].

It's not a secret that the newspaper industry has been in a steady decline triggered by a loss in readership and ad revenue which have been migrating to other media, most notably digital [18]. Additionally, according to [19] Digital Innovation has been the engine to drive change in the industry of news and media. Even though the transition from print to digital started more than two decades ago, the changes to news and media companies during this pandemic time have been dramatic. However, to take advantage of this evolution, news and media firms are leveraging the strength of their digital platforms to generate previously unattainable insights into reader behavior. Companies are enhancing both reader engagement and online revenue success by employing this information with greater sophistication. Knowing this, it's important for this type of companies to keep improving their services and intern processes, in order to keep them in force. The Pandemic has accelerated the Digital Transformation for everything and everyone, and only the ones that take this threat as an opportunity will succeed in the present and in the near future. So, to keep with the dynamics of the business in the digital world require digital newspapers to understand their readers' behavior. 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 the 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.

2 Materials and Methods

2.1 Data

The database comprises six hundred eighty-nine readers who have interacted with the newspaper from January to March 2019. We assessed the number of visits made by each user to each newspaper section (see Table 1) to identify the intensity of use of the website. Thus, the average activity in each section allows defining standardized behavior vectors, independent of the number of visits to the newspaper. For reproducibility purposes, only the information corresponding to users who visited the newspaper six times in the last three months has been used. The Observations corresponding to 70% (482) of the total data used in the study correspond to the model's training phase, and the remaining 30% (207) of readers will be evaluation elements. The new datasets developed in the cross-validation process will train the Decision tree and Random Forest, models.

Table 1. Summary of the predictor variables.

Section	Min	Max	Mean	SD
Main page	0	55.5	16.7	12.1
Politics	0	44.4	24.6	11.1
Economy	0	20	4.2	5.4
Sports	0	25	5.62	6.5
Culture	0	25	6.13	6.8
Interview	0	33.3	13.1	6.5
Opinion	0	44.4	17.4	9.5
International	0	44.4	19.2	8.9
Video	0	100	65	31.6

2.2 Target Variables

The output variables of the classification process represent the reader's profile: Visual, Informed, and Net-NEE, as described in [20].

Readers with a visual profile show a high utilization rate of videos and little relation with reading contents. This profile resembles those known in the literature as digital natives, those who prefer the graphics to the texts; they use external shortcuts to access the Web and frequently share information with their friends on social networks [21].

Consequently, readers in the Informed profile highlight the widespread use of the newspaper sections and show global interest in what happens in their environment; this group represents 50.5% of the total sample of users used for this study. This profile is like that of digital immigrants, who prefer sequential processes. It is as if they learn a new language, culture, and communication approach in social terms.

The NetNee profile responds to the behavior of little interest in the newspaper's contents. It hardly interacts with the other sections; its entrance to the Web is through external platforms, such as social networks or forums, which shows that, at the first moment, he engages with the information. Still, when he enters the Web, he leaves immediately. According to Hernández et al. [22], 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.

2.3 Decision Trees

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. Also, Decision tree classifiers usually employ post-pruning techniques that evaluate the performance of decision trees, as they are pruned by using a validation set [16].

The decision tree implemented is based on the architecture of the "CART" algorithm developed using the rpart package [23] of the software RSTUDIO. The data set was divided repeatedly through the cross-validation technique during the training process, creating ten new data sets, thus generating a trial and error process to characterize the parameters. The Gini Diversity Index (GDI) was considered as an optimization criterion for the models for evaluating the models. The GDI measures the level of impurity of each node; therefore, a node will be pure when all the observations belong to the same category.

The design process of the decision tree has been developed in the following way: The minimum number required for the creation of a participation node is equal to ten and at least one observation for a response node. The tree creation procedure was repeated five hundred 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

Pruning methods aim to simplify decisions trees that overfitted data [24] which consists in examining the nodes that have a more negligible effect in the general classification. In other words, pruning means changing the model by deleting the child nodes of a branch node. The pruned node is regarded as a leaf node. Leaf nodes cannot be pruned [25]. In this research, the pruning process was applied to penalize the decision tree's complexity, ensuring significant partitions were involved in the model.

2.4 Random Forest

Random Forest is a well-known and powerful supervised classification method. Due to its high accuracy and robustness, and some ability to offer insights by the ranking of its features, RF has effectively been applied to various Machine Learning applications. RF consists of a set of decision trees, each of which is generated by the bagging algorithm with no pruning, forming a "Forest" of classifiers voting for a particular class [26]. Random Forest models are methods articulated between machine learning algorithms; it entails the repeated and growing building of many decision trees using an aggregation method called bootstrapping [27]. In Breiman's approach, each tree in the collection is formed by first selecting at random, at each node, a small group of input coordinates to split on, and secondly, by calculating the best split based on these features in the training set [28]. Thus, generating several decision trees with varied variable compositions such that each tree provides an independent outcome, followed by a democratic approach in which the category with the most votes is chosen as the final output. The ability to generate different responses for each decision tree and then combine them into general forecast results in robust models that are less susceptible to extreme values than a basic decision tree, boosting the model's prediction and classification capabilities.

Also, to train an RF, it's needed two parameters, the number of tress (ntree) in the forest and the number of randomly selected features/variables used to evaluate at each tree node (mtry), must be supplied, as well as a training database with ground-truth class labels [26].

The RF model incorporates a variable selection strategy, which enables it to handle data sets with many variables if preceding processes are used to minimize dimensions. Additionally, the model allows for determining the importance of each variable for correctly classifying observations using a permutation test.

The Random Forest model used consists of 500 trees created under the following guidelines. We considered three as the minimum number of observations that give rise to a response node. The number of variables used to create the trees varied from 4 to 8.

2.5 Performance Metrics

The classification process's success is set by the difference between the anticipated and actual values. The True Positive (VP), True Negative (VN), False Positive (FP), and False Negative (FN) metrics all describe this relationship [2]. The model's evaluation will calculate the Correct classification rate (C), Kappa (K) [29], and the area under the receiver operating characteristic curve (ROC) [30]. The region beneath the curve AUC (area under the ROC curve) shows the rate of TP and FP at various thresholds of discrimination. An AUC value equal to one indicates a model with perfect classification. Meanwhile, an AUC equivalent to 0.5 represents an utterly rando model.

3 Results

3.1 Exploratory Data Analysis

A visual representation of the data was created using Principal Component Analysis (PCA). Initially, we arranged data according to the frequency of viewing and use of newspaper sections. The first two principal components (PCs) account for 80.4% of the data. Figure 1 depicts the reader's activity in two dimensions. As a result, the plot's points represent readers, and the shapes respond to each profile.

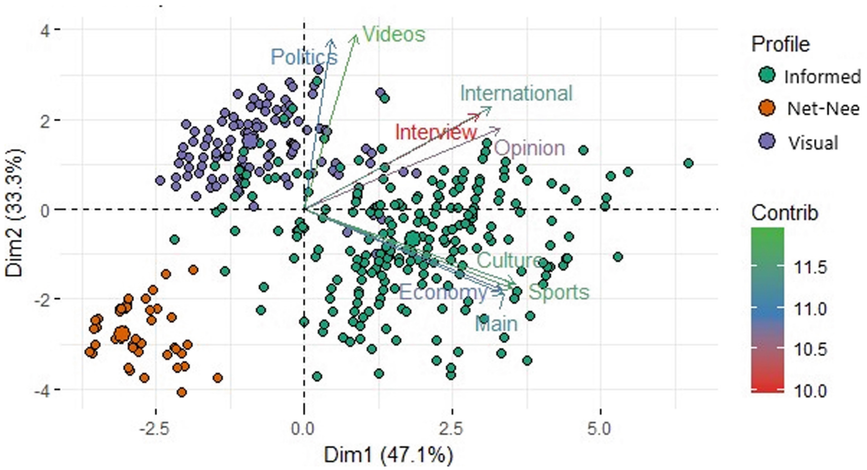

Fig. 1. Two-dimensional representation for users' behavior on the web.

The greater the contribution value, the more significantly the variable contributes to the principal components. As a result, the variable that has the tiniest information is Interview. The informed group (circles) is distributed throughout the horizontal axis, in the growing direction of the Main, Sports, Culture, Economic, International, Opinion, and Interview sections. This location maximizes the relations of these users with the newspaper, representing high visiting frequency and interaction. On the other hand, the Visual readers (triangles) allocated on the top of the vertical axis interact with the videos and Politics sections but in the lower position of the growing vectors. On the opposite, the NetNee users (squares), are located in the third quadrant of the plane; this allocation reduces the bulk of sections since they are growing in the opposite direction of their growth, indicating a lack of interest in interacting with the newspaper.

3.2 Decision Tree Results

The decision tree shown in Fig. 2 was developed after creating 500 trees, trained in different subsets of data, exchanging the roles of training and evaluation among them. The decision tree correctly predicted the type of reader with 93.1% for training and

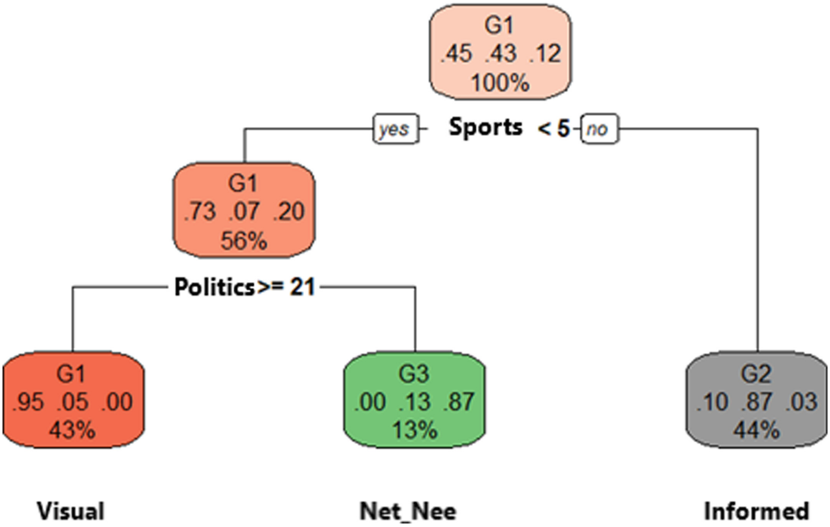

Fig. 2. Schema of DT with 3 branch nodes and 3 leaf nodes.

88.1% for the test. So, the decision tree DT can identify reader types G1, G2, and G3 with 81%, 100%, and 90.4% sensitivity, respectively.

In Fig. 3, we can see the ROC results with an AUC-TRAIN = 0.97 and AUC-TEST = 0.95 for the decision tree predictions. From the nine sections of the diary used for the classification process, the decision tree identifies the "sports" and "politics" sections as the critical discrimination variables. None of the remaining seven sections has been used by the decision tree model to predict the reader's profile.

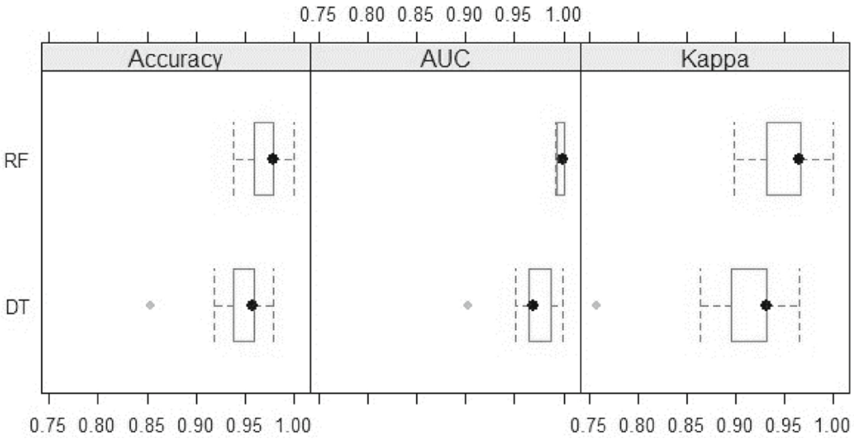

Fig. 3. Boxplot illustrating the results for the three metrics selected

3.3 Random Forest Results

The Random Forest model, built on 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. The specificity evaluation 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%.

The RF model was replicated ten times to test the model's compliance with the decision tree. The results indicate a decrease in variability (sigma = 0.006) and an overall improvement in performance (see Table 2).

Table 2. Performance metrics of DT and RF models.

Metric	DT		RF	
	Training	Test	Training	Test
Accuracy (%)	0.9376	0.8696	0.9751	0.9807
Kappa	0.8963	0.7899	0.9588	0.9685
The area under the ROC curve (AUC)	0.9724	0.9519	0.9967	0.9927

4 Discussion

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. [31] in their study on the classification of web users using the KNN algorithm; the object of study was the Simple Syndication system (RSS), achieving 70% accuracy in the recommendations' quality, which 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 continuous learning process for updating profiles according to readers' interaction with the website.

It is essential 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 recommendation systems.

As far as the researchers' knowledge is concerned, the scientific community is presented with a novel model for classifying readers in a digital newspaper, using profiles 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. Therefore, using this model by other means of digital communication and comparing results will generate a scalable model.

5 Conclusion

The interaction between users and the newspaper's website has been modeled effectively, using the nine sections of the newspaper as predictor variables, using the frequency of visit, and use of the sections to make up the dataset of 489 users. The policy and sports sections have the most significant discrimination capacity to determine user profiles: Visual, Informed and NetNee.

The model identifies users using the sports section less than 5% and the policy section greater than 21% as a reader of the "Visual" profile with a 95% probability. Those readers with 5% or more use of the sports section will classify for the "Informed" profile with an 87% probability. Users who consume the politics section greater than 21% and sports less than 5% will be in the NetNee profile.

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

This research could help this newspaper to fully take advantage of this new insights to improve the User Interface and Experience, so they can build a strong engagement with their readers and keep new readers into the newspaper by offering a tailor-made service for each group of readers that the newspaper has. This kind of study allow the newspaper to take decisions based on data, meaning that all the decisions taken would be fully substantiated and efficient. Taking decisions based on the results of this research would allow resources saving for the newspaper at the time to make or implement any strategies for gaining or retaining their clients well known as readers. This kind of approach is a combination of innovation, technology, statistical methods and Data Science. A solution up to the mark in this digital transformation era.

References

1. El Naqa, I., Murphy, M.J.: What is machine learning? In: El Naqa, I., Li, R., Murphy, M.J. (eds.) Machine Learning in Radiation Oncology, pp. 3–11. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18305-3_1
2. De La Hoz, E., Zuluaga, R., Mendoza, A.: Assessing and classification of academic efficiency in engineering teaching programs. J. Effi. Responsib. Educ. Sci. **14**, 41–52 (2021)
3. Escorcia Guzman, J.H., Zuluaga-Ortiz, R.A., Barrios-Miranda, D.A., Delahoz-Dominguez, E.J.: Information and Communication Technologies (ICT) in the processes of distribution and use of knowledge in Higher Education Institutions (HEIs). Procedia Comput. Sci. **198**, 644–649 (2022)
4. Suthaharan, S.: Big data classification: problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Perform. Eval. Rev. **41**, 70–73 (2014)
5. Nayak, A., Dutta, K.: Impacts of machine learning and artificial intelligence on mankind. In: 2017 International Conference on Intelligent Computing and Control (I2C2) pp. 1–3. IEEE, Coimbatore (2017)

6. Obermeyer, Z., Emanuel, E.J.: Predicting the future — big data, machine learning, and clinical medicine. N. Engl. J. Med. **375**, 1216–1219 (2016)
7. Yu, Q., Miche, Y., Séverin, E., Lendasse, A.: Bankruptcy prediction using extreme learning machine and financial expertise. Neurocomputing **128**, 296–302 (2014)
8. Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., Sheth, A.P.: Machine learning for internet of things data analysis: a survey. Digit. Commun. Netw. **4**, 161–175 (2018)
9. De-La-Hoz, E.J., De-La-Hoz, E.J., Fontalvo, T.J., De-La-Hoz, E.J., De-La-Hoz, E.J., Fontalvo, T.J.: Methodology of machine learning for the classification and prediction of users in virtual education environments. Inf. Tecnológica. **30**, 247–254 (2019)
10. Delahoz-Dominguez, E.J., Fontalvo, T., Zuluaga, R.: Evaluation of academic productivity of citizen competencies in the teaching of engineering by using the Malmquist index. Form. Univ. **13**, 27–34 (2020)
11. Delahoz-Dominguez, E., Zuluaga, R., Fontalvo-Herrera, T.: Dataset of academic performance evolution for engineering students. Data Brief **30**, 105537 (2020)
12. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)
13. Erevelles, S., Fukawa, N., Swayne, L.: Big Data consumer analytics and the transformation of marketing. J. Bus. Res. **69**, 897–904 (2016)
14. Stalidis, G., Karapistolis, D., Vafeiadis, A.: Marketing decision support using artificial intelligence and knowledge modeling: application to tourist destination management. Procedia Soc. Behav. Sci. **175**, 106–113 (2015)
15. Sundsøy, P., Bjelland, J., Iqbal, A.M., Pentland, A.“, de Montjoye, Y.-A.: Big data-driven marketing: how machine learning outperforms marketers' gut-feeling. In: Kennedy, W.G., Agarwal, N., Yang, S.J. (eds.) SBP 2014. LNCS, vol. 8393, pp. 367–374. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05579-4_45
16. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., Akinjobi, J.: Supervised machine learning algorithms: classification and comparison. Int. J. Comput. Trends Technol. **48**(3), 128–138 (2017)
17. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**, 211–236 (2017)
18. Adgate, B.: Newspapers Have Been Struggling and then Came the Pandemic. https://www.forbes.com/sites/bradadgate/2021/08/20/newspapers-have-been-struggling-and-then-came-the-pandemic/
19. Deloitte: Digital Transformation Through Data for News and Media Companies. http://www2.deloitte.com/us/en/pages/consulting/articles/digital-transformation-through-data-for-news.html
20. De La Hoz Domínguez, E., Mendoza Mendoza, A., Ojeda De La Hoz, H.: Classification of readers profiles of a digital journal. Rev. UDCA Actual. Amp Divulg. Científica. **20**, 469–478 (2017)
21. Ahn, J., Jung, Y.: The common sense of dependence on smartphone: a comparison between digital natives and digital immigrants. New Media Soc. **18**, 1236–1256 (2016)
22. Hernández, D.H., Ramírez-Martinell, A., Cassany, D.: Categorizando a los usuarios de sistemas digitales. Pixel-Bit Rev. Medios Educ. (2014). https://doi.org/10.12795/pixelbit.2014.i44.08
23. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive partitioning and regression trees (Version R package version 4.1-10). URL HttpsCRAN R-Proj. Orgpackage Rpart (2015)
24. Esposito, F., Malerba, D., Semeraro, G., Kay, J.: A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 476–493 (1997)

25. IBM: Pruning Decision Tress. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/db2/10.5?topic=view-pruning-decision-trees
26. Petkovic, D., Altman, R., Wong, M., Vigil, A.: Improving the explainability of Random Forest classifier – user centered approach. In: Biocomputing, pp. 204–215. WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA (2018)
27. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
28. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. **13**, 1063–1095 (2012)
29. Araújo, F.H.D., Santana, A.M., de Pedro, A., Neto, S.: Using machine learning to support healthcare professionals in making preauthorisation decisions. Int. J. Med. Inf. **94**, 1–7 (2016)
30. Faraggi, D., Reiser, B.: Estimation of the area under the ROC curve. Stat. Med. **21**, 3093–3106 (2002)
31. Adeniyi, D.A., Wei, Z., Yongquan, Y.: Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Appl. Comput. Inform. **12**, 90–108 (2016)