Classification of mineral zones using machine learning methods. Case study in Colombia.

Emanuel Chaverra Zuleta, Franco Bertaiola Ríos, and Andrés Mauricio Muñoz García
Geophysics and Computer Sciences Group (GGC3), Metropolitan Technological Institute of Medellín (ITM)

Abstract

In the study of deposits and the exploitation of mineral resources, traditionally, the decisions of the geographical positioning of new drilling wells necessary for the generation of geomodels are made subjectively, under a professional criterion that often responds to the experience and assumptions of the geologist. and not to a methodology based on the scientific method. Therefore, there is a need to create a non-subjective decision-making methodology in mineral resource exploration, with greater reproducibility and less uncertainty. Characterizing a deposit begins many times with field work that defines the taking of samples and the measurement of its physical and geochemical parameters, and it must end with the generation of geologically plausible models that favor decision-making on the early stages of development of a mining project. Obtaining these models is a great challenge for Colombian companies and much more because they do not have comprehensive methodologies developed thinking about the specific conditions of the place of study, and supported by computational tools for the treatment and analysis of the data that result from field work. and in the laboratory for collected samples. This work presents the first stage of the development of a methodology for the generation of mineral deposit models using machine learning methods (Machine Learning, ML).

Introduction

Machine Learning (ML) is an artificial intelligence (AI) technique that is responsible for giving computers the ability to learn without prior programming (Samuel, AL 1959). Currently, ML is permeating almost all human activities and is used in a wide range of situations, for example, in social networks to recommend some type of product, service or activities depending on the tastes, choices and needs previously identified of a particular link. In medicine it is useful to detect some type of anomaly in X-ray images, tomography, etc., which allow diagnosing diseases (Sagar, 2019). In banking, they are used to detect fraud, money laundering or simply to classify customers and know what type of products can be recommended according to specific characteristics and behaviors of customers (Badillo, 2019).

The different branches of study of earth sciences are also not unaffected by these computational methods. In the development of some of its activities such as hydrocarbon prospecting, a discipline in which important information is generated through different techniques such as reflection seismic and well logs to classify lithological structures or formations of geological interest, they have been implemented ML algorithms to solve different problems and meet a particular objective, is the case of waveform recognition and first arrivals, analysis of well logs (Huang, Williams, and Katsube, 1996) or seismic tomography (Arraya-Polo et al. ., 2018), and the estimation of rock parameters at a seismic scale using well logs, seismic data and Artificial Neural Networks (ANN),

In the study of mineral resources, these algorithms have also been used for the exploration of deposits. This is the case of the use of Vector Support Machines (SVMs) to obtain a map of prospective zones of a copper deposit located in Iran (Abedi, Norouzi, and Bahroundi, 2012), Artificial Neural Networks. RNAs, Regression Trees (RTs), Random Forests (RF), and SVMs to model epithermal gold (Au) mineral prospectivity of the Rodalquilar district in Spain (Rodriguez-Galiano et al., 2015), and data from gravimetry, magnetometry, electromagnetic methods, geological mapping,

This paper presents the results of the first stage in the development of a methodology based on field data and computational tools such as ML for the generation of mineral deposit models that support decision-making in the early stages of a mining project. . The process and the results obtained in the evaluation and selection of the computational tools that best fit and allow solving the particularities of the study site are described, this particularity represented in the geochemical data of three mining districts in Colombia and ten mines in South Africa.

Method

For the evaluation of some Machine Learning (ML) methods in the identification of the characteristics of mineral samples, three data sets were used. The first, which we will call “Data Colombia (DC)” consists of the chemical concentration of 19 elements reported for 908 samples, the primary data were taken from the geodatabase of the Colombian Geochemical Atlas published by the Geological Service (SGC) in 2016, and consists of the report of the concentration of 56 elements in 13,704 sediment samples (Colombian Geological Service, 2016). The DC samples were taken from three mining districts, Barranco de Loba in the Department of Bolívar, Iquira-Teruel in the Department of Huila and Farallones in the Department of Valle del Cauca (Figure 1 shows its location on the map of Colombia) .

“Data Africa (DA)” is the second set of data (information used by Dixon in his doctoral thesis) from which 10 mines were selected, corresponding to 354 samples and of which 23 chemical elements were reported (Dixon 2014).

Finally, the third set of data is a simplification of the Colombian data, which we will call “Reduced Colombia Data (DCr)”.

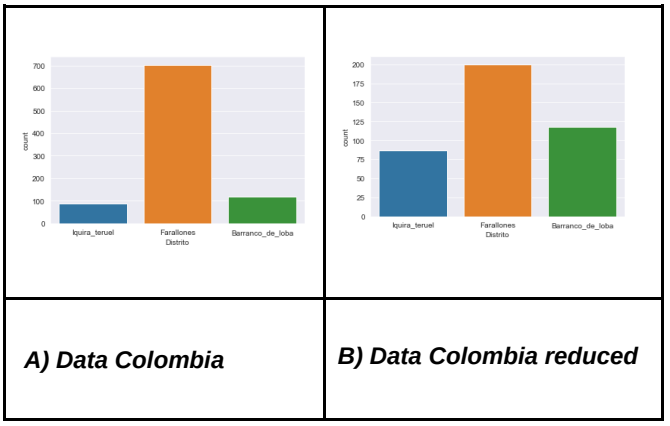
In this research some methods ofexploratory data analysis (Eploration Data Analysis, EDA), dimensional reductionand ML for classification, implemented in Python, of which its performance and precision with respect to the input data could be verified, starting with the Decision Tree method (DTC) which frames a hierarchical data structure. Later, the Random Forest (RF) was used, from which the most popular class was selected from several trees to obtain a robust model. After this, the performance of the Gradient Boosting (GB) method for classification was evaluated and the Hierarchical Classification (CJ) was used for the classification of data according to their spatial relationship. Additionally, the Vector Support Machine (SVM) method was used to classify data from the data's own spatial limits.

Of the methods used in the classification methodology, three are assembly, DTC (Hauska and Swain, 1977), RF (Breiman, 2001) and GB (Friedman, 2002), these methods have an internal structure in the form of a "tree" , this allows making decisions based on specific data conditions. These generalities can be identified even with little input data as in the case of DA where there are approximately 35 data for each class; An approximation of the internal structure of these methods is observed in Figure 11, where the characteristic ramifications are identified when being trained; Additionally, two ML methods based on spatial distance were used, CJ and SVM (Joachims, 1998), due to their nature of operation, these two methods may have drawbacks with the high dimensionality of the inputs. For the experiments carried out here, this dimensional problem would put these two methods at a disadvantage, therefore, to carry out the classification experiment with the total of the 5 methods, their performance was evaluated with different variants of the three original data sets, said data They were transformed by three dimensional reduction algorithms, PCA (Principal Component Analysis) (Roweis and Sam, 1998), t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008). and UMAP (Uniform Manifold Approximation and Projection) (Mclnnes, Healy, and Melville, 2020), this to evaluate when they present the best result and obtain the best combination of data - ML - Dimensional reduction. To perform the classification experiment with the total of the 5 methods, its performance was evaluated with different variants of the three original data sets, said data were transformed by three dimensional reduction algorithms, PCA (Principal Component Analysis) (Roweis and Sam , 1998), t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008). and UMAP (Uniform Manifold Approximation and Projection) (Mclnnes, Healy, and Melville, 2020), this to evaluate when they present the best result and obtain the best combination of data - ML - Dimensional reduction. To perform the classification experiment with the total of the 5 methods, its performance was evaluated with different variants of the three original data sets, said data were transformed by three dimensional reduction algorithms, PCA (Principal Component Analysis) (Roweis and Sam , 1998), t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008). and UMAP (Uniform Manifold Approximation and Projection) (Mclnnes, Healy, and Melville, 2020), this to evaluate when they present the best result and obtain the best combination of data - ML - Dimensional reduction. t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008). and UMAP (Uniform Manifold Approximation and Projection) (Mclnnes, Healy, and Melville, 2020), this to evaluate when they present the best result and obtain the best combination of data - ML - Dimensional reduction.

Figure 1: Location of samples in Colombia

Results

The number of elements in each of the data sets per class was identified by means of Python, in order to determine the correct distribution of these. For DC, a disproportionate amount of data was found in the Farallones district, for this reason a random data reduction was carried out within this class, giving rise to the third set of data called Data Colombia Reduced (DCr). Figure 2 shows figure A) Data Colombia and figure B) Data Colombia reduced.



A) Data Colombia ***B) Data Colombia reduced***

Figure 2: Comparison between original DC vs DCr using Python.

- A description of the pairwise correlation of each of the numerical values of the input information was made,

obtaining its graphical representation allowing to identify linearly dependent data and eliminate them if necessary since this can generate disturbances in some ML models (figure 3).

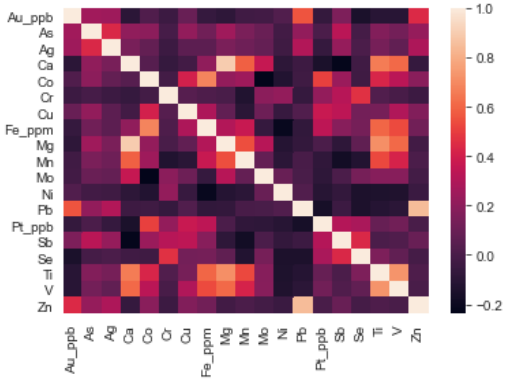
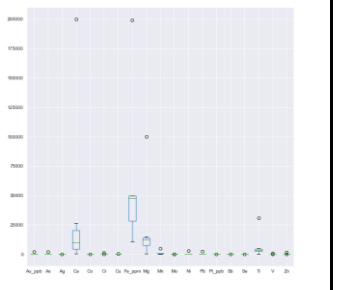
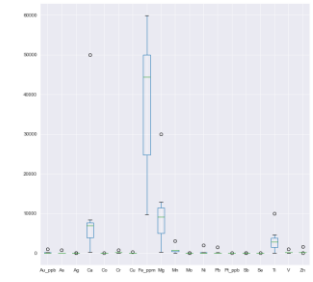


Figure 3: Heat Map Data Africa

- In order to obtain the statistical distribution of the datasets, each of the chemical elements of these were represented graphically, in the box diagrams of Figure 4, of which it can be identified in Figure A) DC with values outliers, which led to the elimination of these outiliers, generating the distribution of Figure B) DC without significant outliers.



A) Data Colombia with outliers



B) Colombia data without significant outliers

Figure 4: ***Outlier boxplot comparison between DC vs DCr using the pandas library.***

- One of the fundamental points to obtain characteristics of an area is to identify which elements stand out in these places, either by their presence or by their absence, as in the case of titanium which is absent in Barranco de Loba and this is returns one of the criteria to perform a classification (Figure 5)

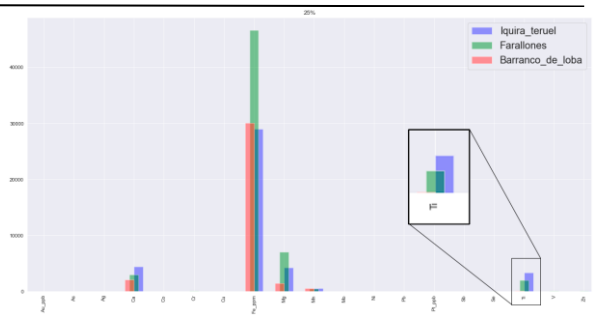


Figure 5: Barplot percentile 25% of DCr of which the absence of Tttanio is observed in barranco de loba

- For the dimensional reduction stage, in which the amount of input data is decreased, three algorithms were used, the first was Principal Component Analysis (PCA) applied to the 3 data sets to obtain their 2D representation and feed the methods of ML, which was also done with the other 2 dimensional reduction algorithms. Figure 6 shows the result of this first algorithm for Data Africa; later, the embedding of stochastic neighbors distributed in t (t-SNE) was used to repeat the aforementioned process (figure 7), although this algorithm has a higher performance than PCA, it has the disadvantage that it does not accept new data to be transformed, therefore it is also I apply the uniform collector approximation and projection (UMAP), From which a very effective dimensional reduction was obtained for all datasets and additionally a model that accepts new data to transform as in PCA. In figure 8 you can see the Colombia data in a 3D space from UMAP, of which the limits are very well defined.

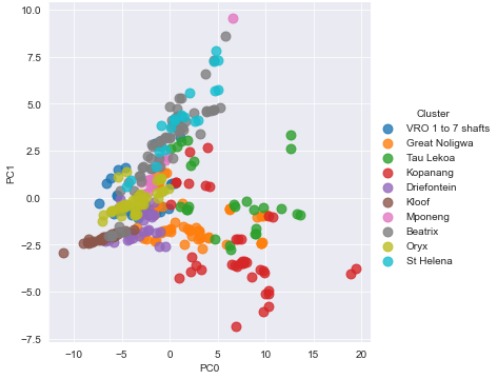


Figure 6: Graphic representation of the PCA dimensional reduction for Data Africa

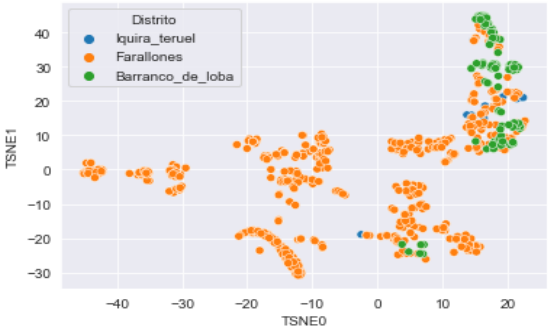


Figure 7: 2D representation of the t-SNE result for DC

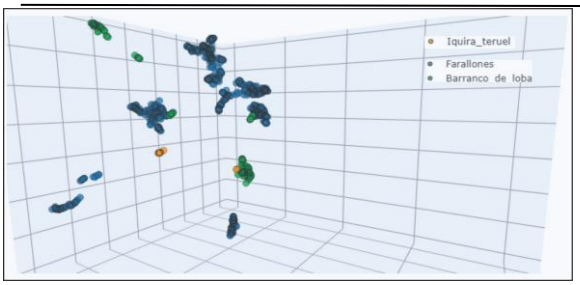


Figure 8: 3D representation of the UMAP result for DC

- In figure 9 you can see the decision tree with its internal structure for the reduced Colombia data, this tool allows a geoscientist to analyze if the selection criteria of the model are in accordance with the known characteristics of the area and in figure 10 it is You can see the importance to the model of each of the input columns. Additionally, figure 11 shows the feature importance from the results of the application of the random forest method, taking into account that the data sets do not reach 1000 rows (since RF is functional with few input data).

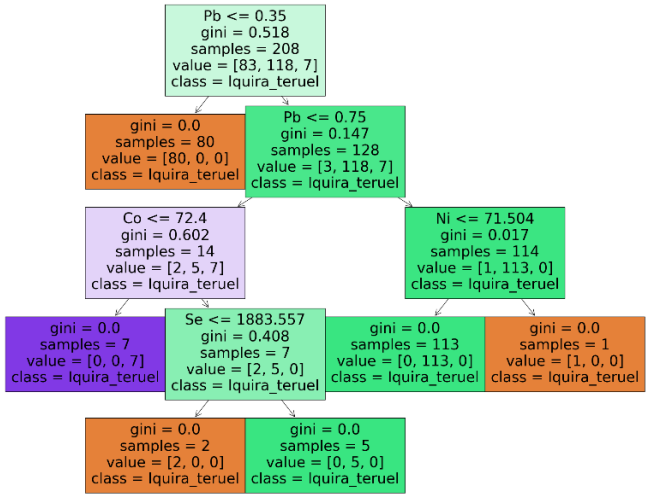


Figure 9: Internal structure of the DTC for DCr

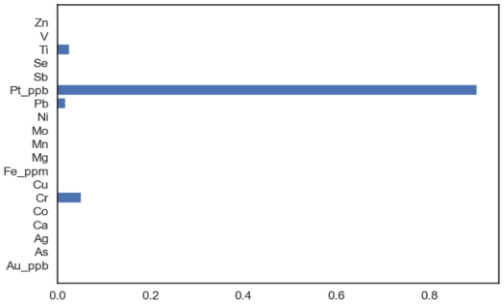


Figure 10: Feature Importance of the DTC for DCr. The relevance of the inputs for the Data can be observed, having Pt as the most outstanding.

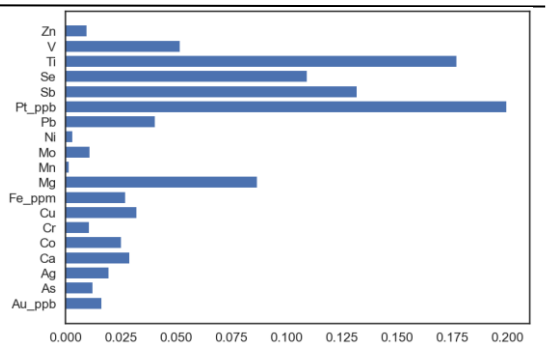


Figure 11: RF Feature Importance for the model obtained from DCr. Pt can be seen as the most outstanding of the data and Mn as the least outstanding.

- In Figure 12 you can see the tuning of the model after one hundred iterations. This graphic representation is relevant because it is taken as an indicator when adjusting the GB hyperparameters.

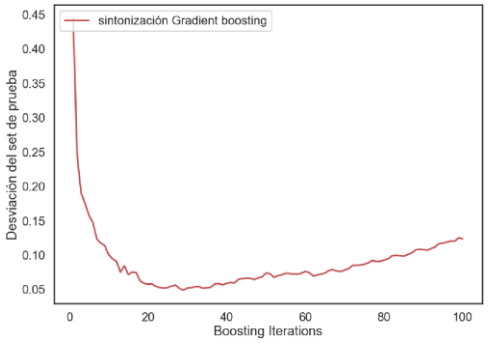


Figure 12: Tuning of the GB for the UMAP result of DC observing in the Y axis the deviation values of the test set and the X axis boosting iterations.

- Figure 13 shows the dendrogram obtained from the hierarchical classification, using the DA. In figure 16 you can see the result of the hierarchical classification with data Africa, which obtained a yield of 11%, this very poor result is due to the fact that, as observed in the dendrogram, the method only manages to identify 2 classes, so when it is requested that the classification be made in the 10 classes of AD the algorithm does not have good selection criteria

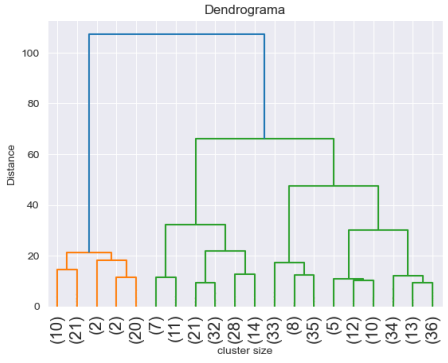


Figure 13: Dendrogram resulting from DA, it is observed that the algorithm only manages to distinguish two of the ten classes (orange and green), this is a bad

indication for its performance.

- The vector support machine algorithm does not present a good performance for Data Africa, this is due to the fact that in this data set there are many classes and relatively few data per class, therefore, as observed in figure 14, there are no manage to create spatial limits clear enough for a good classification, while in data Colombia the algorithm was able to classify the areas thanks to the fact that in this dataset there are only 3 classes in addition to a significant amount of data per class, resulting in better limits defined as shown in figure 15.

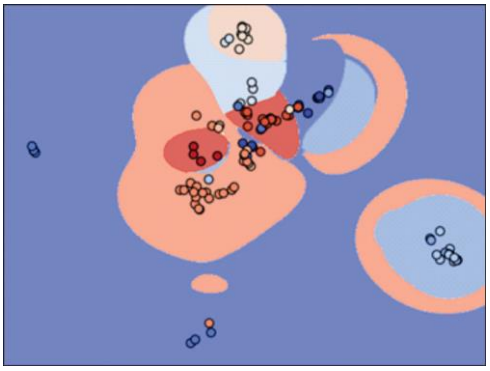


Figure 14: MSV result for DA from UMAP, each color represents a zone (class / mine).

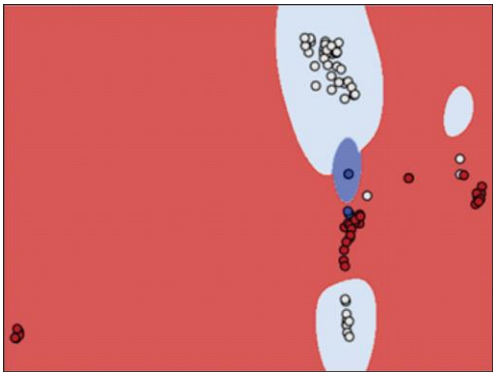


Figure 15: MSV result for the reduced CD from UMAP, each color represents a zone (class / District).

- Very good results were obtained in Data Colombia (see table 1, consolidated results achieved with this data), highlighting RF and DTC. Despite this, and as mentioned previously, the results may be biased by the saturation of data present in Farallones, so they should be compared with the results of Data Colombia reduced.

ML algorithm	Original data (DC)	PCA - DC	TSNE - DC	UMAP- DC
RF	0.999	0.964	0.893	0.960
GB	0.969	0.911	0.884	0.987
DTC	0.996	0.920	0.929	0.995
CJ	0.801	0.801	0.564	0.560
SVM	0.880	0.920	0.898	0.982

Table 1: CD Result. In each case the correlation factor is displayed as the selected metric.

- In the reduced Colombia data, the assembly methods presented a reduced change, even so, they continue to be the best performing, whereas for the spatial methods the results were worse with the exception of the data from UMAP where their predictions improved (see table 2,consolidated results achieved with this data).

ML algorithm	Original data (DCr)	PCA - DCr	TSNE - DCr	UMAP- DCr
RF	1	0.889	0.789	0.955
GB	1	0.922	0.755	0.955
DTC	0.99	0.855	0.800	0.944
CJ	0.147	0.147	0.200	0.727
SVM	0.700	0.767	0.522	0.989

Table 2: Reduced CD Result.In each case, the correlation factor is shown as the selected metric to evaluate the performance of each algorithm.

- In table 3 it can be seen that the RF and DTC methods gave the best results for the Africa data. Both the CJ and SVM algorithms gave mainly CJ unfavorable results. In this dataset the dimensional reduction worsened the predictions, this may be due to the fact that there are a high number of mines as targets and very little input data, therefore the dimensional reduction methods would not have enough information to give better results.

ML algorithm	Original data (DA)	PCA - DA	TSNE - DA	UMAP- DA
RF	0.953	0.887	0.719	0.860
GB	0.187	0.710	0.439	0.037
DTC	0.869	0.719	0.663	0.692
CJ	0.110	0.110	0.0310	0.158
SVM	0.673	0.654	0.178	0.645

Table 3: AD Result.In each case, the correlation factor is shown as the selected metric to evaluate the performance of each algorithm.

Conclusions:

It was possible to determine that both the RF model and the DTC had good performance despite the small number of data in each dataset used in this study, thus verifying that these ML algorithms present an excellent performance even under these conditions, which agrees with the researched literature. For classifications of many classes, it is recommended not to use distance-based methods (SVM and CJ) and preferably to use RF. For classifications of few classes, SVM combined with UMAP is a recommended option.

It was possible to determine the set of dimensional reduction and machine learning algorithms that best fit the specific conditions of the study site, and in this way, classify the data set according to the place where they were taken, now, as part of the study. methodological development involved in this study, it is necessary to scale these results to other areas of Colombia.

Acknowledgments

We thank the Metropolitan Technological Institute of Medellín (ITM), project code P21111, project code 20248 and the Colombian Geological Service for the data provided in the framework of this article

References:

Arraya-Polo, M. et al. (2018). "Deep-learning tomography". In: Leading Edge 37.1, pp. 58–66. url: https: //doi.org/10.1190/tle37010058.1.

Badillo, E. (2019). The Impact of Machine Learning on Economics: What Machine Learning Can (and Cannot) Do for Economic Research. Tech. Rep. Chicago Policy Review. url:https: // chicagopolicyreview. org / 2019/01/21 / the - impact - of - machine - learning - on - economics - what - machine - learning- can- and- cannot- do- for- economicresearch /.

Berger, KJ, OA Hoop, and GC Beroza (2019). "Machine learning for data-driven discovery in solid Earth geoscience". In: Science, pp. 315–323. Breiman, L. (2001). "Random Forest". In: Machine

Bressan, TS, de Souza, MK, Girelli, TJ, & Junior, FC (2020). Evaluation of machine learning methods for lithology classification using geophysical data. Computers & Geosciences, 139, 104475.

Breiman, L. (2001). "Random Forest". In: Machine Learning, pp. 5–32.

Dixon, RD (2014). "Provenance of illicit gold with emphasis on the witwatersrand basin." PhD thesis. University of Pretoria.

Friedman, JH (2002). "Stochastic gradient boosting". In: Computational Statistics Data Analysis, pp. 367–378.

Huang, Z., J. Williams, and J. Katsube (1996). "Permeability prediction with artificial neural network modeling in the Venture gas field, offshore eastern Canada". In: Geophysics 61.2, pp. 422-436. url:https://doi.org/10.1190/1.1443970.

Hauska, Hans and Philip H. Swain (1977). "The decision tree classifier: Design and potential". In: IEEE Transactions on Geoscience Electronics, pp. 142-147.

Iturrarán-Viveros, Ursula (2012). "Smooth regression to estimate effective porosity using seismic attributes." In: Journal of Applied Geophysics 76, pp. 1– 12. url:https://doi.org/10.1016/j.jappgeo. 2011.10.012.

Iturrarán-Viveros, Ursula and Andrés M. Muñoz-García (2018). “Validated artificial neural networks in determining petrophysical properties: A case study from Colombia”. In: Interpretation 6.4, pp. 45–54. url:https://doi.org/10.1190/int-2018-0011.1. - (2021). “Machine Learning as a Seismic Priority Velocity Model Building Method for Full-Waveform Inversion: A Case Study from Colombia”. In: Pure and Applied Geophysics 178.2, pp. 423-448. url: https: //doi.org/10.1007/s00024-021-02655-9.

Iturrarán-Viveros, Ursula and JO Parra (2014). "Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data". In: Journal of Applied Geophysics 107, pp. 45–54. url:https://doi.org/ 10.1016 / j.jappgeo.2014.05.010.

Mitchell, TM (1997). Machine Learning. In M.-H. Education (Ed.), Intelligent Systems Reference Library. McGraw-Hill Series in Computer Science.

Samuel, AL (1959). "Some Studies in Machine Learning Using the Game of Checkers". In: IBM Journal of Research and Development 3.3, pp. 211–229. url:https://doi.org/10.1147/rd.33.0210

Sagar, A. (2019). "Deep Learning for Detecting Pneumonia from X-ray Images". In: Towards Data Science. url: https://towardsdatascience.com/deeplearning-for-detecting- pneumonia- from- xray-images-fc9a3d9fdba8.

Colombian Geological Service (2016), Geochemical Atlas 2016. Retrieved June 5, 2021https://www2.sgc.gov.co/sgc/mapas/Paginas/AtlasGeoquimico.aspx.

Joachims, T. (1998). "Making large scale SVM learning practical". PhD thesis. Technishe Universitat Dourtmund.

Roweis and Sam (1998). "EM algorithms for PCA and SPCA". In: Proceedings of the 1997 conference on Advances in neural information processing systems 10, pp. 626-632.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: Journal of machine learning research, pp. 2579-2605.

Mclnnes, L., J. Healy, and J. Melville (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Tech. Rep. Tutte Institute for Mathematics and Computing.