04-024

GROUNDWATER RECHARGE AREAS MAPPING IN HIGH-ALTITUDE ANDEAN MOUNTAINS THROUGH MACHINE LEARNING ALGORITHMS

Aliaga Medrano, Evelyn (1); Soria Cespedes, Freddy (1); d'Abzac, Paul (1)

(1) Universidad Católica Boliviana San Pablo

The high-altitude wetlands in the Central Andes are unique ecosystems located above 4000 masl in the Bolivian Altiplano. The analysis and classification of spatial information is a crucial step in the identification of wetlands in scarped topography. The objective of this study was to test machine learning algorithms to map Andean wetlands. The first step consisted on applying the machine learning algorithms Least Absolute Shrinkage and Selection Operator LASSO and Receiver Operating Characteristic ROC for the sensitivity analysis. Then, there were compared the Random Forest Regressor RFR, Support Vector Regressor SVR, and Multivariate Adaptative Regression Splines MARS regression supervised machine learning algorithms for the wetlands mapping. Results were validated by Google Earth satellite images and a regression coefficient. The RFR showed good results for areas with slopes of 0 - 32 degrees; the SVR showed good performance for areas with slopes of 44 - 76 degrees, while for areas with slopes of 0 - 12 degrees its performance was inaccurate. The application of the MARS showed trivial results compared to those obtained by the first two algorithms; some results were good for certain areas, areas with slopes of 0 - 12 degrees and 44 - 77 degrees were erroneously flagged.

Keywords: andean wetlands; Random Forest Regressor; Support Vector Regressor; Multivariate Adaptative Regression Splines regression; supervised machine learning algorithms

DESEMPEÑO DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA EL MAPEO DE ÁREAS DE RECARGA SUBTERRÁNEA EN ZONAS ANDINAS DE BOLIVIA

Los bofedales de altura de los Andes Centrales son ecosistemas únicos ubicados por encima de los 4000 msnm en el Altiplano boliviano, que brindan valiosos servicios ecosistémicos. El mapeo de humedales brinda información valiosa para la conservación y el manejo. El objetivo fue probar el desempeño de algoritmos de aprendizaje automático para mapear bofedales andinos. El primer paso fue la aplicación de los algoritmos de regresión LASSO (Least Absolute Shrinkage and Selection Operator) y ROC (Receiver Operating Characteristic) para el análisis de sensibilidad. Luego, se compararon los algoritmos MARS (aprendizaje automático supervisado por regresión), RFR (Regresor de Bosque Aleatorio), SVR (Regresor de Vector de Soporte) y MARS (Regresión adaptativa multivariante). Los resultados se validaron con imágenes de Google Earth y un coeficiente de regresión. El RFR mostró buenos resultados para pendientes de 0 - 32 grados; el RVS mostró un buen desempeño para pendientes de 44 a 76 grados, pero no así para pendientes de 0 a 12 grados. La aplicación del MARS mostró algunos resultados buenos para ciertas áreas, pero presentó errores para áreas con pendientes de 0 a 12 grados y de 44 a 77 grados.

Palabras clave: bofedales andinos; Regresor de Bosque Aleatorio; Regresor de Vector de Soporte; Regresión Adaptativa Multivariante; algoritmos de aprendizaje automático supervisado

Agradecimientos: Este trabajo se desarrolla a través del apoyo del CINAES, Centro de Investigación en Agua, Energía y Sostenibilidad de la Universidad Católica Boliviana San Pablo. La presentación de la comunicación es financiada a través del PROGRAMA DE DESARROLLO INCLUS

1. Introduction

Governments around the world are taking actions towards poverty reduction. The 2030 Sustainable Development Goals are based on that premise, among which Goal 6, Clean water and sanitation and Goal 15, Life of terrestrial ecosystems, make a reference to water. Under that context, water conservation is not just protecting water sources; it also implies caring for the environment along the water cycle.

In the remote Andes, wetland systems are an important component of the water cycle. Andean wetlands are worth studying due to their social and natural relevance. From the human perspective, wetlands provide water during the dry season and forage for the cattle along the entire year (Beck et al., 2000). From a research perspective, wetlands are interesting because their responses result from a mix of hydro-climatic alterations, including the effects of retreating glaciers, and the influence of unplanned anthropogenic interferences. Both aspects increase the sensitivity of wetlands to global climatic changes.

To investigate wetlands responses a common approach is to apply field techniques to measure surface and subsurface water flow trends (Pan & Zhang, 2021). However, in remote areas where topography varies considerably, challenges arise from the difficulties to interpret and predict spatial patterns from sparse and limited field data. A potential tool to advance on that research is machine learning, which can continuously update information based on previous knowledge for investigating the spatial distribution of wetlands. Thus, the general objective of this research was to generate maps of potential infiltration by the application of Artificial Intelligence (AI) algorithms based on mapped data, i.e., field data, of High Andean wetland areas in La Paz as an approach to identify surface water movement patterns.

2. Methodology and data

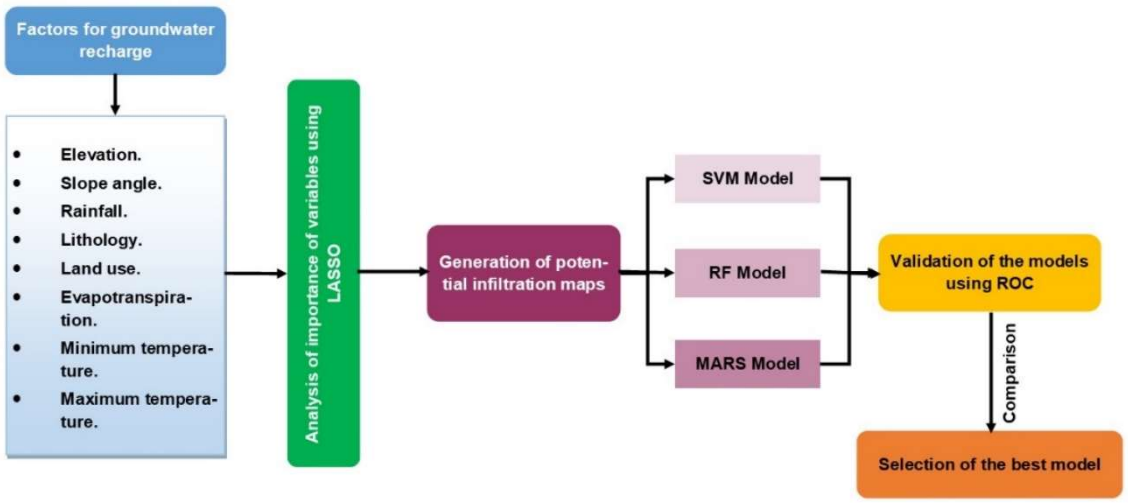
This study follows the methodology proposed by Pourghasemi et al. (2020), adapted to high Andean areas. The hypothesis is that unlike traditional statistical methods, the application of AI in the modeling process can handle noisy data, adequately handling data inaccuracies.

The analysis has three stages. At first, data was extracted and classified into 16 effective factors, studied by the LASSO (Least Absolute Shrinkage and Selection Operator) algorithm to identify statistical correlation problems. Then, the aim was to compare Machine Learning (ML) algorithms to find the one(s) with less overfitting problems to obtain maps; the Artificial Intelligence (AI) algorithms tested were SVM (Support Vector Machines), MARS (Multivariate Adaptative Regression Splines), and RF (Random Forest). The last stage was focused on testing the sensitivity of AI algorithms. Results were validated by visual adjustment with Google Earth imagery to provide measures to test the predictive capacity of each AI algorithm. The methodology is summarized in Fig. 1.

For the data cleaning process before applying the AI algorithms, an analysis of importance was carried for 16 effective factors. That analysis ranks factors in order of importance by the LASSO algorithm, regarding their predictive potential of infiltration patterns. The application of the algorithm also indicates the presence of statistical correlation problems.

The models to predict potential infiltration patterns development by AI are based on regression analysis. For this work, there were applied the SVM, RF and MARS algorithms. Data was randomly divided in two sets, to train and test the algorithms. Finally, the predictive accuracy was also evaluated by the ROC AI algorithm; ROC uses classification analysis for model performance evaluation. Based on the elements obtained, is identified the best algorithm for generating maps of potential infiltration patterns for high Andean areas.

Figure 1: Diagram of the research methodology



Note: Elaborated after Pourghasemi et al. (2020)

Data required for the preliminary analysis was: terrain elevation, land slope angle, lithology, land use, rainfall rates, evapotranspiration, minimum and maximum air temperature. Climatic data was obtained from the Bolivian Surface Water Balance 1980 – 2016 (Bolivia: Ministry of Environment and Water, 2016); topographic data was obtained from the NASA´s repository (https://search.earthdata.nasa.gov).

The sixteen effective factors were chosen based on available databases and then they were elaborated by Geographic Information Systems (GIS) tools. DEM (Digital Elevation Model) raster is obtained from the NASA repository (https://search.earthdata.nasa.gov); from it there were built the Slope Angle factor, Topographic Wetness Index (TWI) factor, Multi-resolution Index of Valley Bottom Flatness (MRVBF) factor, and Aspect. Regarding hydrological information, annual precipitation, evapotranspiration, annual maximum temperatures, and annual minimum temperatures are obtained from the Bolivian Surface Water Balance 1980 – 2016 (Bolivia: Ministry of Environment and Water, 2016). Annual 24-hour maximum rainfall factor was calculated from series provided by the National Meteorology and Hydrology Service (SENAMHI) of Bolivia.

The assumption to estimate infiltration is that the higher the permeability, the greater the infiltration of surface water into the ground. Infiltration was estimated from mapped values under the SCS (US Soil Conservation Service) method for abstractions (Chow et al., 1988). The method empirically relates direct runoff to precipitation depth. The dimensionless curve number (CN) varies from 0 to 100; for impervious surfaces and water surfaces CN = 100; for natural surfaces CN < 100. The calculation of the CN value is based on three humidity conditions: normal (CN II), dry (CN I) or humid (CN III); CN calculations also consider four hydrological groups of soil types: A, B, C and D and 20 land use types. Final infiltration values were obtained combining raster data pf geology, land use and a DEM into new CN empiric values.

2.1. Variables for the algorithms

Variables were introduced in the algorithms as: X = effective factors, Y = CN on behalf of permeability information. The algorithm assumes that X is standardized before the calculation. This procedure is performed before splitting data into training and testing sets.

CN values were calculated by the algorithms using a best fitting parameters approach. The models were trained applying each of the algorithms, using the best fitting parameters and adjusted to the X and Y variables of the training set for each hydrologic unit (HU) of the area.

2.2. On the interpretation of the effective factors

The following considerations apply: Fault Density and Distance from Faults factors are considered as effective factors because geologic faults serve to transport and storage groundwater. Geology and Land Use factors contribute with information about soil characteristics. Drainage Density and Distance from Rivers factors are considered as effective factors because drainage networks provide implicit evidence of excessive permeability and surface porosity (Pourghasemi et al., 2020).

2.3. On the LASSO, SVM, RF, MARS, and ROC algorithms

The LASSO algorithm applies a linear regression analysis method. It was implemented for sensitivity analysis, that is, to identify the variable(s) of importance within the effective factors for mapping potential infiltration patterns. The LASSO algorithm is not aimed specifically at mapping. This stage is necessary to understand the behavior of the variables that will be considered for the AI algorithms. A requirement for the LASSO application is that the features within a data set do not need to have the same measuring units.

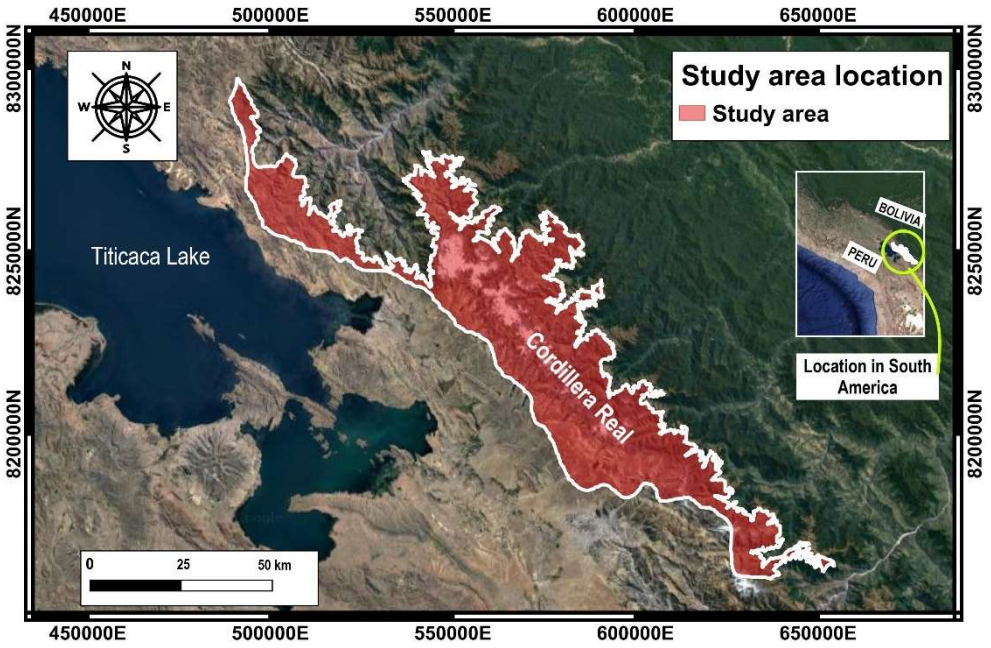
After the sensitivity analysis, the SVM, RF and MARS algorithms were implemented to predict the CN potential infiltration patterns. After modeling, a classification analysis was applied by the ROC algorithm to each model generated by the AI algorithms. The results evidence the predictive capacity of the algorithms, thus allows identify which of them had the best performance in each HU.

3. Study area description

Bolivia has a rich ecosystem diversity. The infiltration inferences are carried in the High Andean areas of La Paz, in the vicinity of the Cordillera Real and the Northern Puna, on an area of about 3647 km^2. The location of the study area referred to the country is presented in Fig. 2.

The area is home to a small rural population, located far from the densely populated cities of La Paz and El Alto. The weather has a dry and a wet season. In the dry season (May – August, winter) temperatures fluctuate from 4.5 °C to -5 °C; in the wet season (December – March, summer) temperatures are around 6.5 °C, with frequent rains, snowfall, and fog in the mountains (Beck et al., 2000). Groundwater reaches its highest levels at the end of the wet season and because of evapotranspiration. During the dry season, wetlands tend to reduce their productivity; however, grazing as the principal economic activity can be carried throughout the year due to the availability of fresh and nutritious forage provided by wetlands.

Figure 2: Study area location

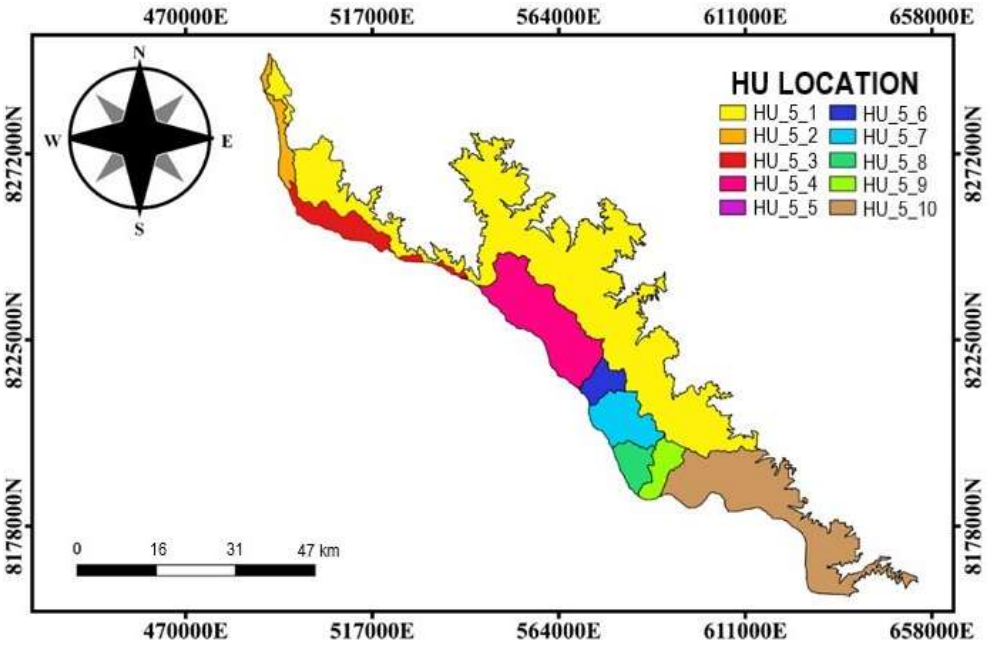


4. Results

4.1. Application of the AI algorithms

The study area was divided into 10 sub-basins named as Hydrographic Units (HU) for mapping purposes (Fig.3). The entire data set is split by a random process into a training and validation data set, following a 70:30 proportion, which means that 70% of the data set is used for model training (calibration stage) and 30% for model validation. A sample size of 2000 was considered for each HU; there was an exception for HU_5_5, where sample size was 100 because the HU size was smaller than the other HUs.

Figure 3: Hydrographic Units (HU) of the study area.



4.2. Application of LASSO algorithm

To interpret the LASSO outputs, the variables distribution and the pairwise relationships constructed is performed for each data set in individual graphs used to explain the behavior of the input variables. Each graph has its own individual scale according to the pairwise relationship it represents. The main diagonal of the schematic graphs demonstrates the distribution of the input variables. That is, it represents the pairwise relationship of the analysis variable with respect to the same variable. The rest of the graphs represent the relationship by pairs of each analysis variable with respect to the other variables.

Through the LASSO analysis it is possible to determine the predominance of hydrological or morphological factors in each HU. For example, for HU_5_10 (see Fig. 4), the analysis determined that the variable of importance is the DEM (DEM_30_p), followed by the Annual 24-hour maximum rainfall (P24H) and the annual precipitation (Prec). LASSO indicates that the minimum annual temperature (Temp_min), topographic aspect (Aspect), drainage density (Dens_dre) and slope (Slope angle) have low importance for the model. As opposed, for HU_5_8 (Fig. 5), the analysis determined that the variable of importance is geology (Geo), followed by land use (Land Use) and minimum annual temperature (Temp_min), whereas have low importance to the model the effective factors maximum annual temperature (Temp_max), Annual 24-hour maximum rainfall (P24H), evapotranspiration (ETP), topographic aspect (Aspect), precipitation (Prec), fault density (Dens_fault), MRVBF, TWI and DEM.

Figure 4: Analysis of Variable of importance HU_5_10 by LASSO algorithm

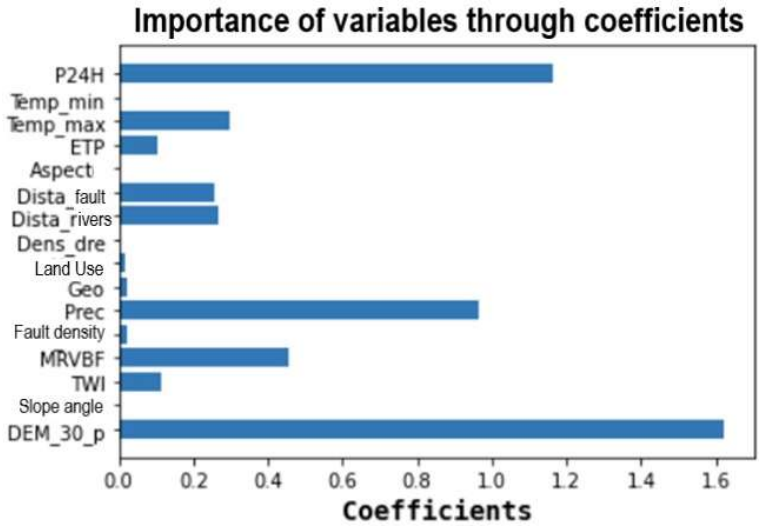
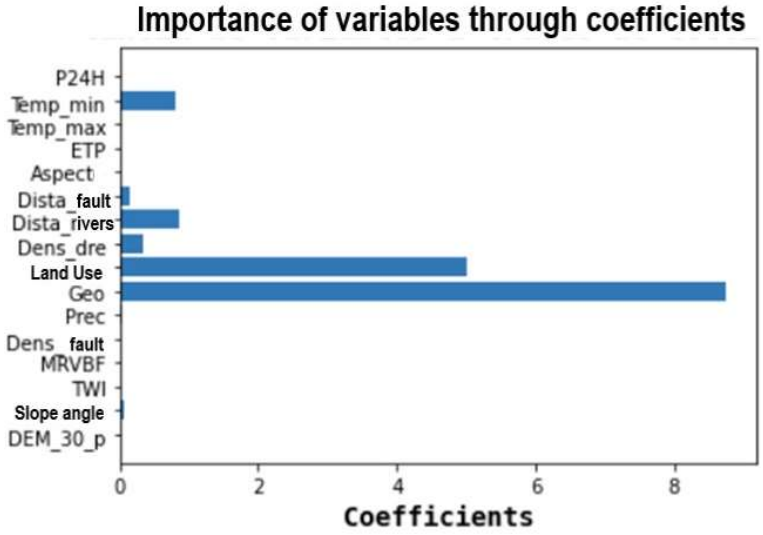


Figure 5: Analysis of Variable of importance HU_5_8 by LASSO algorithm

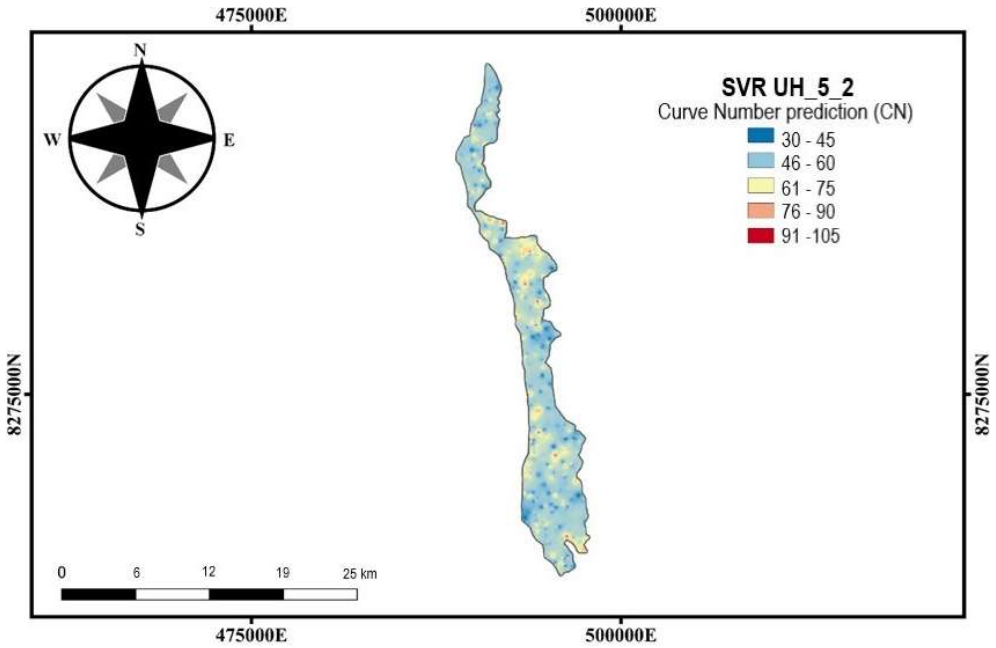


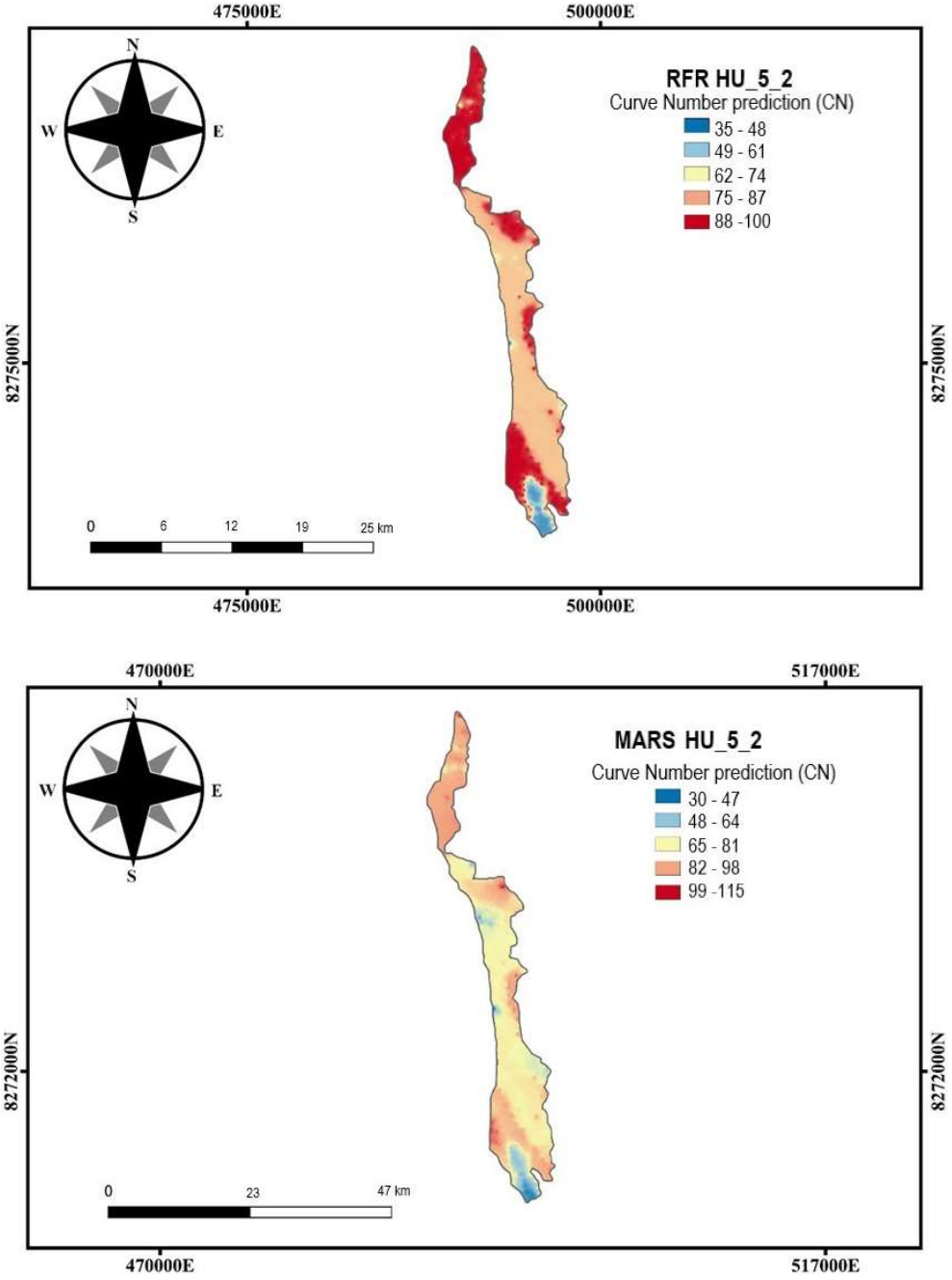
4.3. Application of SVM, RF, and MARS algorithms

The models built by regression analysis of ML, SVM, RF and MARS algorithms, suggest a logical prediction indicating that the least permeable areas are those where there is some water body (e.g., ponding effect) and where land slope is steep. In places where the predicted CN values are lower, topographic slopes are steep.

For HU_5_2, the SVM algorithm predicted a CN in the range 68 to 95 (Fig. 6, upper), where the highest CN are on agricultural zones. The RF algorithm predicted a CN range from 66 to 86 (Fig. 6, middle), where the highest values are on grassland areas. The MARS algorithm calculated CNs in the range 63 to 94 (Fig. 6, bottom), where the highest value is located close to agricultural zones.

Figure 6: Map prediction of HU_5_2 by the application of SVM, RF, and MARS algorithms





4.4. Application of ROC algorithm

The ML algorithm suitability is evaluated by ROC using the Area Under the Curve (AUC) approach. Ideally, the graph should show a false positive rate equal to zero and a true positive rate equal to one, since it seeks to find the largest possible AUC. It is important to consider the slope of the generated curves since the aim is to maximize the rate of true positives and minimize the rate of false positives. Likewise, a calculated random prediction (blue) allows a better understanding of the behavior of the generated curves on graphics.

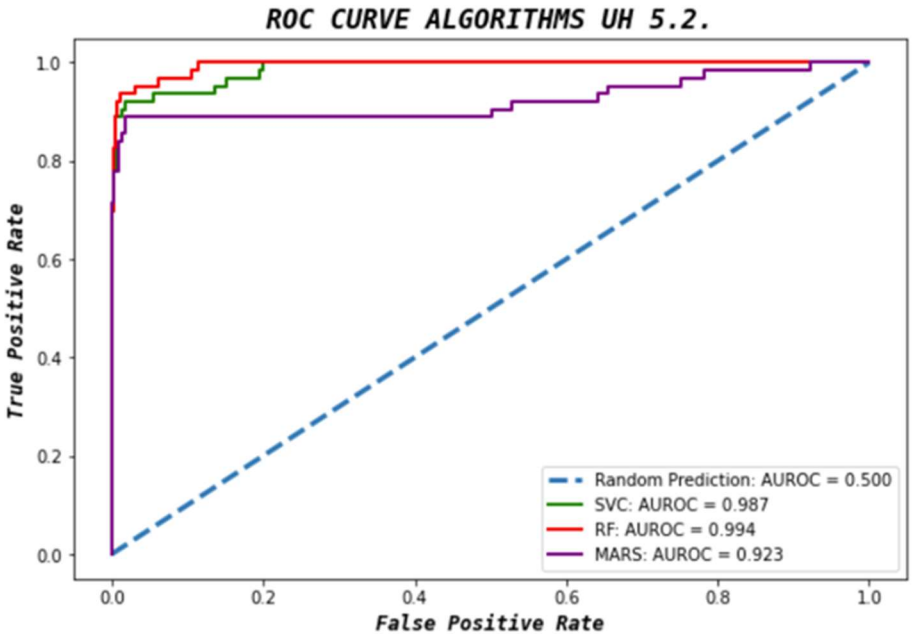
The performance of the algorithms was different for each HU; among all algorithms, the performance of the RF stands out. Citing some examples, in the HU_5_2 (Fig. 7, upper), The RF (red) has an AUC equal to 0.994, being the algorithm with the best performance. The next

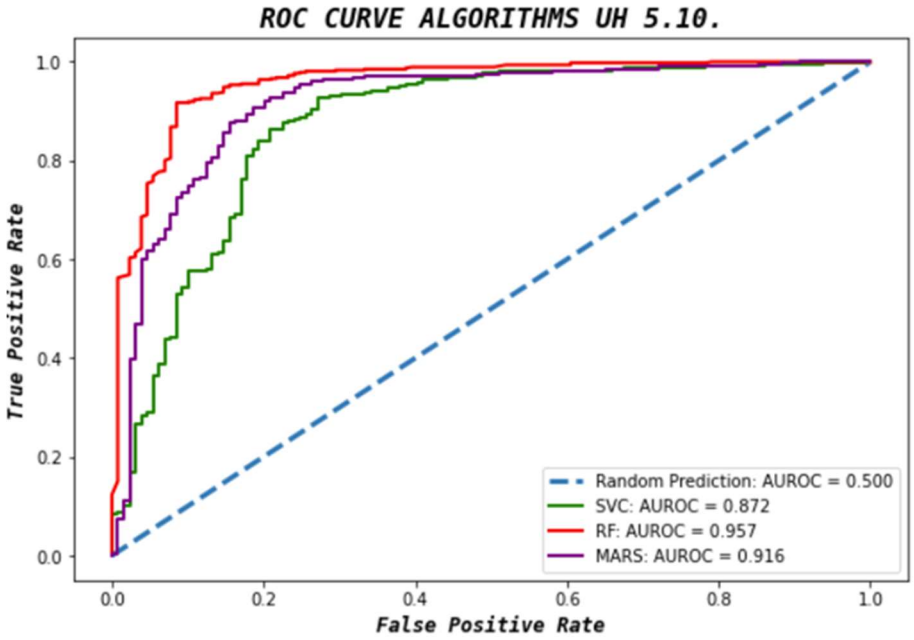
in performance are the SVC (green) with an AUC equal to 0.987 and the MARS (purple) with an AUC equal to 0.923. In the HU_5_10 (Fig. 7, bottom), the RF (red) has an AUC equal to 0.957, i.e., the best performance, followed by the MARS (purple) with an AUC equal to 0.916 and the SVC (green) with an AUC equal to 0.872. In the HU_5_5 (Fig. 8), the RF (red) has an AUC equal to 1.00, being the algorithm with the best performance, followed by the MARS (purple) and SVC (green) with an AUC equal to 0.966. In the latter case, however, the result suggests that all algorithms performed well, which is contradictory because it would mean that any model can be apply regardless the quality of data used; that inconsistency indicates a relationship between the sample size and the model performance, i.e., the algorithms are not adequate when small sample sizes are selected.

4.4.1. Visual comparison

Visual comparison with Google Earth satellite images allowed us to appreciate the predictions made and the spatial patterns that emerge from them. Results from the comparison are presented by potential infiltration pattern maps, which reflect the “real” CN value and the potential attribute predicted by the SVM, RF and MARS supervised ML algorithms for the study area.

Figure 7: Application of ROC algorithm results.

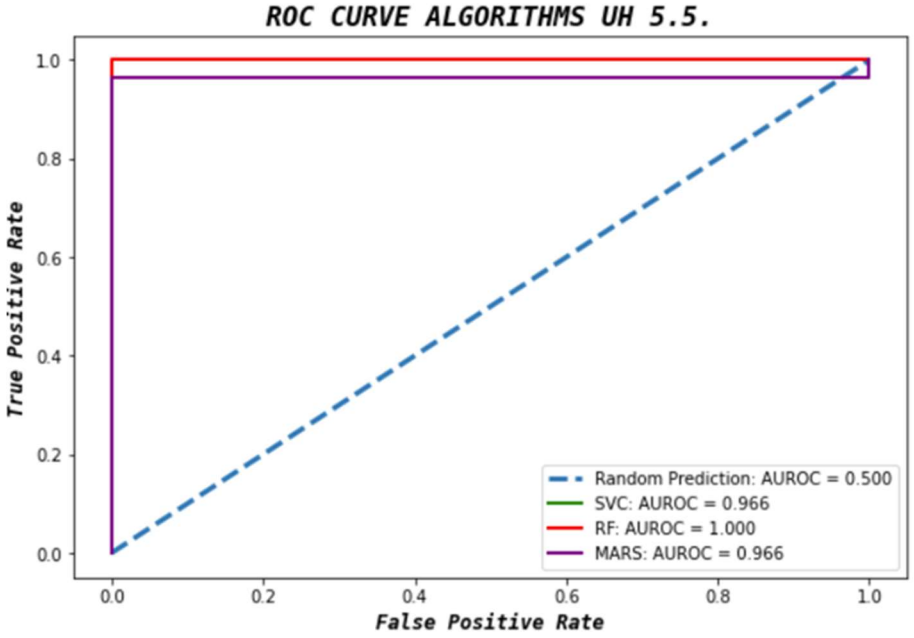




The potential infiltration patterns map generated by the RF algorithm (Fig. 9) shows that the highest predicted CN values are close to 98 (very high runoff potential), spatially located over snow and glacier covered areas, where slope are generally steep. Conversely, the RF allocates the lowest CN values (very low runoff potential) over the lower part of the hills where greater infiltration rates are expected. However, the RF fails to predict some areas where are predicted high CN values also on the lower parts of the hills, which reflects an inconsistency.

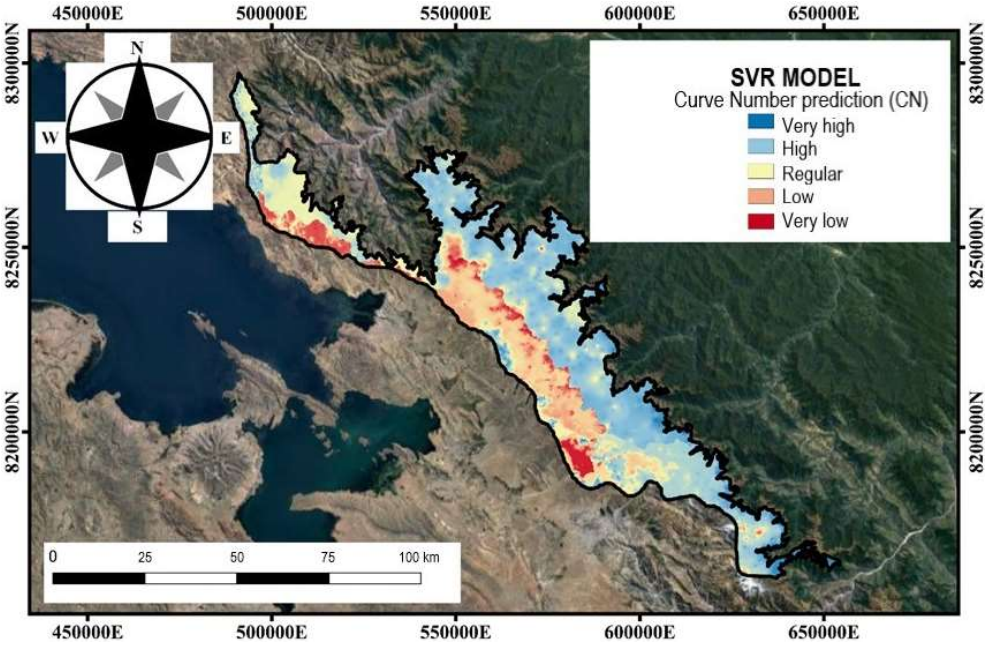
The potential infiltration pattern map generated by the SVM algorithm (Fig. 10) also is accurate when predicting the occurrence of high CN values over steeped areas covers by snow and glaciers, on the peaks of the mountains. However, the algorithm fails when predicting CN values above 100 suggesting the presence of water bodies which was not accurate in all the cases.

Figure 8: ROC curve of HU_5_5 by the application of ROC algorithm.



The potential infiltration pattern map generated by the MARS algorithm (Fig.11) inadequately predicts high CN values in some areas on the lower part of the hills. On the contrary, over steep glacierized peaks, predicted CN values are close to 98 (very high runoff potential).

Figure 9: SVM Model potential infiltration pattern map



5. Discussion

Some operational aspects on the applications of the algorithms are worth mentioning: i) the variables of importance differ at each HU; ii) In some cases, there were removed some effective factors to improve the model performance. That is because there were linear correlations among data set parameters. However, all parameters were used for the mapping.

Figure 10: RFR Model potential infiltration pattern map

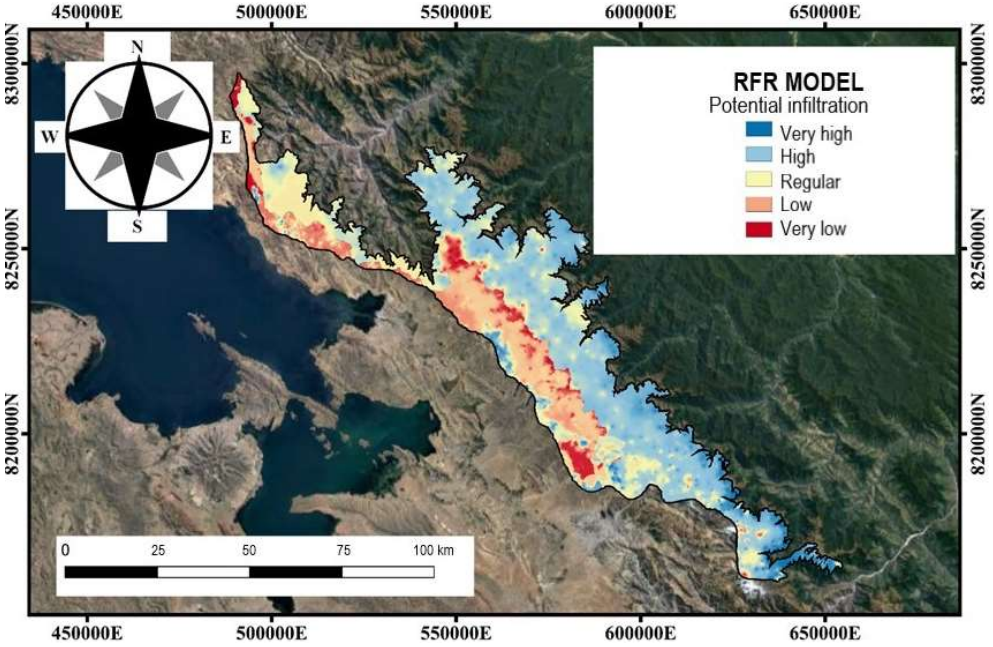
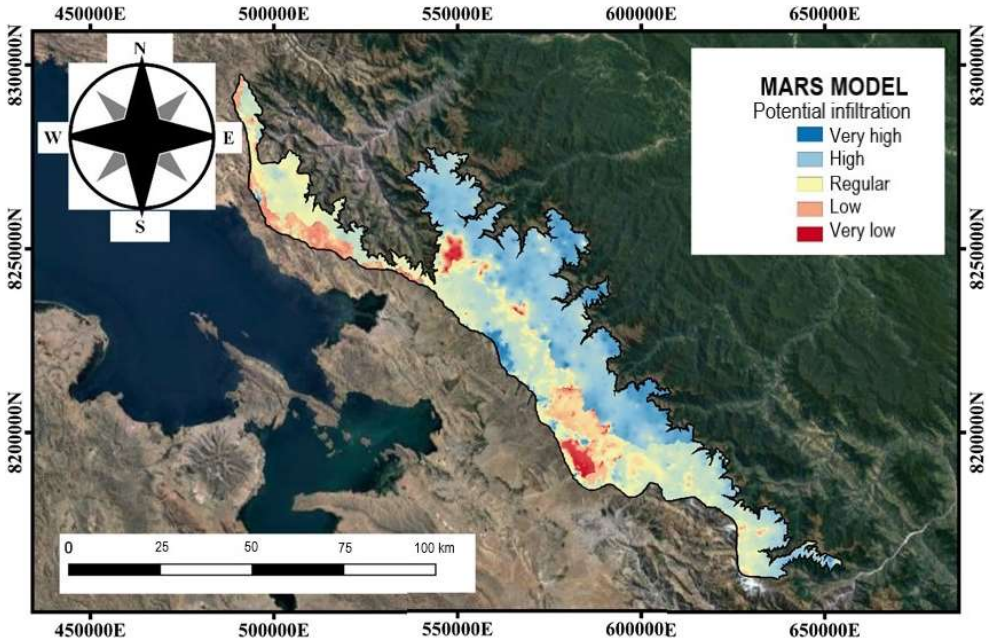


Figure 11: MARS Model potential infiltration pattern map



On the LASSO algorithm application, results suggest a relationship between hydrological factors and the morphology. In contrast, Pourghasemi et al. (2020) indicated that all the factors were important and/or influential for the analysis. Compared to our study, such contrast could be attributed to the high quality of data, data sources and scale homogeneity of the Pourghasemi et al. (2020) study.

On the application of AI SVM, RF, and MARS algorithms, in general there were not many differences among the maps generated by them, highlighting their predictive capacity. However, each algorithm has its own particularities, which can be observed in subtle variations that can be seen when comparing the maps generated by the three algorithms and with respect to those of the study of Pourghasemi et al. (2020). Some of those differences were the non-

contemplation of all the effective factors in the analysis and the differences of the predicted values of the ranges by each algorithm.

Although the ROC algorithm is used to measure the performance of the algorithms, and indeed it fulfills its purpose, it also reveals the importance of the good development of the application procedure of ML algorithms. Having optimal results requires data with good quality, because that data will be used throughout the learning process (Grant, 2019). In our study, we have inherited issues related to the low availability and ease of access, which are common in developing economies and remote areas; that is an aspect that we have to deal with which would not be solved in the near future. However, according to Pan & Zhang (2021), ML is a big step for AI to teach machines to discover hidden patterns in big data and make data-based predictions for future tasks to try to imitate a perfect reality with imperfect data.

To have an adequate application of ML, it is important to choose the learning strategy among one of the three main types of learning: supervised, unsupervised and reinforcement (Sun et al., 2021), since each one has a different form of evaluation. Likewise, the choice of the learning algorithm and the set of training and performance evaluation factors determine the precision of the results and their interpretation. In our study it was demonstrated that machines have the power to do work independently, even without a high-quality data set, but they still require an intervention from humans. In short, ML algorithms result in effective tools for the analysis of noisy data sets in terms of reducing time and computational expense for large volumes of data corroborating that “an additional factor to examine is the advantages of labor exchange in the past; among them, higher amounts of performance, higher quality and fewer errors” (McFrockman, 2020, p. 26).

6. Conclusion

The potential of machine learning as a tool for analyzing data sets is to understand the behavior of a data set. This study shows that the supervised ML, SVM, RF and MARS algorithms can predict potential infiltration patterns for a study area characterized by a data set. For instance, we conclude that the predictive capacity of this tool, whether performing a regression or classification analysis, is adequate for general purposes.

Throughout the results, it was evidenced that the most important part when generating models with ML algorithms is the data set preparation. It was shown that the analysis of variables of importance by the LASSO algorithm does not have two differentiated trends but rather a combined trend of effective hydrological and morphological factors. In addition, by analyzing the distribution of the variables, the models generated by the LASSO algorithm identified a linear correlation between variables.

The maps of potential infiltration patterns obtained by the application of the supervised ML SVM, RF and MARS algorithms describe well the predictive capacity of each one of them since all algorithms were able to predict CN values from the data set. Although results are diverse, the maps generated by the RF algorithm are the closest to reality described by Google Earth imagery. The CN values predicted by that algorithm are in agreement with the characteristics of the study area.

The results of the application of the ROC algorithm indicate that the three ML algorithms have an adequate predictive capacity. However, the RF algorithm slightly outperforms the SVM and MARS algorithms, since its AUC value is above 0.920. This algorithm is identified as the one with the best performance among the three algorithms used in this study. RF is particularly good processing data series with noisy data and linear correlations among its variables.

In this project, it was possible to verify the versatility of the method applied to generate maps of potential infiltration patterns, since it is not conditioned by specific study characteristics. The visual comparison of the maps generated by the three algorithms with Google Earth satellite

images shows that the CN predictions are not far from reality. The infiltration patterns are identifiable through the attributes assigned within the study area and are consistent with its characteristics. Identifying a potential infiltration pattern and mapping it to understand the spatial location of its potential attributes is advantageous for understanding the behavior of environmental factors in the region.

References

Beck, S., Sarmiento, J., Paniagua, N., Miranda, C., & Ribera, M. O. (2000). Humedales de Bolivia, una aproximación a su conocimiento actual. 119–150.

Bolivia. Ministry of Environment and Water. (2016). Bolivian surface water balance: Diffusion document / Ministry of Environment and Water. La Paz: MMAA.

Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). Applied hydrology. McGraw-Hill.

Digital Center for Natural Resources of Bolivia. (2021, April 15). Digital center of natural resources of Bolivia. http://cdrnbolivia.org/geografia-fisica-nacional.htm

Grant, J. (2019). Deep Machine Learning: A Complete Developer's Guide to Deep Learning Algorithms, Concepts, and Techniques for Beginners. https://ler.amazon.com.br/

Mc Frockman, J. (2020). ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING: Deep learning, AI, Python, and cyber security. Technological revolution in finance, medicine, and business with the characters of the time (Spanish Edition). https://ler.amazon.com.br/

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/j.autcon.2020.103517

Pourghasemi, H. R., Sadhasivam, N., Yousefi, S., Tavangar, S., Ghaffari Nazarlou, H., & Santosh, M. (2020). Using machine learning algorithms to map the groundwater recharge potential zones. Journal of Environmental Management, 265, 110525. https://doi.org/10.1016/j.jenvman.2020.110525

Sun, H., Burton, H. V., & Huang, H. (2021). Machine learning applications for building structural design and performance assessment: State-of-the-art review. Journal of Building Engineering, 33, 101816. https://doi.org/10.1016/j.jobe.2020.101816

Comunicación alineada con los Objetivos de Desarrollo Sostenible