System to identify the processes and countries Colombian migrate to

Mayra A Caicedo, Nini J Velez, Sebastian Sanchez

*Universidad Autónoma de Occidente*

*Cali, Colombia*

mayra\_ale.caicedo@uao.edu.co

nini.velez@uao.edu.co

gomez.sebastian@uao.edu.co

**Abstract— Since ancient times, the human being has been in constant transit. The causes are multiple, including but not limited to the search for better economic opportunities, reunite with family members, studying, escaping armed conflict, persecution, human rights violations, among other factors. Using publicly available data, this project analyses different machine learning models to predict the various destinations Colombians migrates to.**

I. INTRODUCTION

While "migration" refers to the displacement of people from their home country to another, the term "emigration" is a term that identifies those leaving their home country; in contrast, the term "immigration" identifies those arriving at a foreign destination.

The causes of migration include, among others: The search for new economic opportunities, reuniting with family members, studying, escaping armed conflict, persecution, terrorism, human rights violations, including the adverse effects caused by climate change, environmental factors, and natural disaster.  
 "Today, the number of people living in a country other than their home country is greater than ever. According to the IOM World Migration Report 2020, as of June 2019, the number of international migrants was estimated to be almost 272 million worldwide, 51 million more than in 2010. Almost two-thirds were labor migrants. International migrants constituted 3.5% of the world's population in 2019, compared to 2.8% in 2000 and 2.3% in 1980". [1]

“The human migration is classified into six large groups or subcategories: according to their geographical scale, the characteristics of the place of origin and destination, their temporality, their degree of freedom, their cause and according to the age of the migrants”. [2]

The information contained in this project was gathered from publicly available data, including different consular which classifies information by gender, age, education, occupation, place of residence, consulate of the constituency to identify the country of destination of the migrants accurately.

II. DESCRIPTION OF THE PHENOMENON, MODELED PROCESS AND PROBLEM TO ADDRESS

This project will address the emigration phenomenon of Colombians, between the 25 to 40 years of age traveling to other countries.

The question for the problem is defined below:

*¿It is possible to identify to which country the population between 25 and 40 years of age travels, based on their area of ​​knowledge, age, gender, marital status, academic level?*

This question is of international relevance, especially when the world is facing a pandemic, and how this undoubtedly affects the different processes previously implemented by nations facing unprecedented immigration challenges.

We have tackled the issue at the national level by identifying trends in travellers to other countries by profession, age, gender, marital status, and/or academic education level.

*A. Modelling Process*

The model below is presented in a schematic diagram.

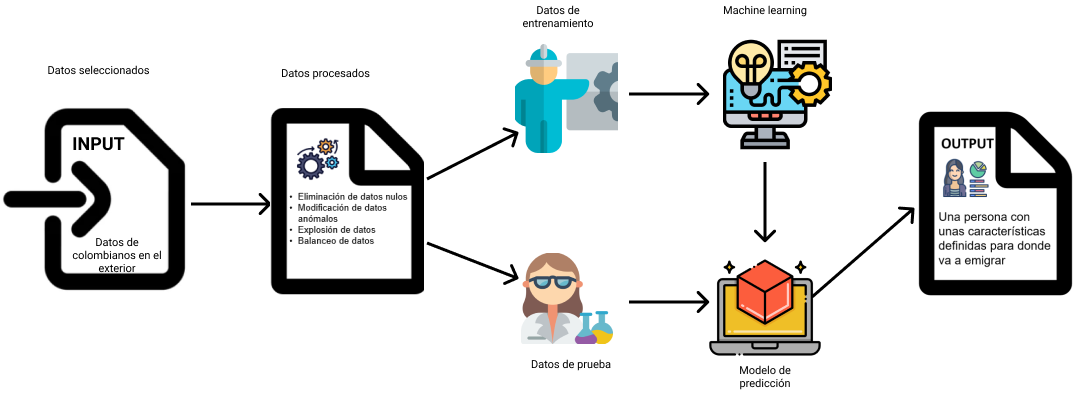
****

Fig. 1 Schematic diagram of the solution process

III.LITERATURE REVIEW

The reviewed articles with similar solutions are listed below:

* A Machine Learning Approach to Modelling Human Migration.
* Using an interpretable Machine Learning approach to study the drivers of International Migration.
* International migration management in the age of artificial intelligence.
* Modelling and predicting the spread of the 2019 coronavirus disease in China incorporating data on human migration.
* Modelo de radiación generalizada para la migración humana.
* How Well Can the Migration Component of Regional Population Change be Predicted? A Machine Learning Approach Applied to German Municipalities.
* A Classification and Data Visualization Tool Applied to Human Migration Analysis.
* Climate vulnerability and human migration in a global perspective.
* Modelling Movement: A machine-learning approach to track migration routes after displacement.
* General Concept of the Storage and Analytics System for Human Migration Data.

Articles 1 and 2 present alternatives with machine learning models to traditional models of human mobility. The first conventional model, called the "gravity model," defines that the probability of a trip between two locations as obtained directly from a distance between them. The second traditional model, called the "radiation model," postulates that the probability of a trip does not depend so much on the distance but on the number of intermediate opportunities (this model is expanded in article 5). These models are based on linear relationships between independent variables.

The alternatives proposed in articles 1 and 2 are based on trees (XGBoost models) and neural networks that allow the non-linear combination of other characteristics provided by the World Bank or the United States Census.

Article 1 develops the first machine learning proposal of the human migration prediction problem and includes procedures to work with data set imbalance, hyperparameter adjustment, and performance evaluation. Likewise, for the neural network, they develop a custom or personalized loss function. Finally, from two datasets (the USA Migration dataset and the Global Migration dataset), they manage to better understand the performance of machine learning models compared to traditional models. Figure 1 below, containing the Top10 for each dataset of the most relevant characteristics used in the XGBoost model.

Article 2 uses a 3-layer neural network with the following characteristics shown in figure 2.

In both cases, they use metrics such as RMSE (Root mean squared error) and r2 (Coefficient of determination), among others.

The other articles complement the analysis context. For example, article 3 how artificial intelligence systems support migration and asylum decision-making processes in countries such as Canada and Germany. In addition, some articles study the causes of migration, such as climate changes (article 8), and expand applications from the analysis of human mobility as in the era of COVID (article 4).

Article 6 presents models for predicting migration in Germany for two demographic groups: young people between 18 and 24 years old and family population between 30 and 40 years old and between 0 and 17 years old. The models used were: (1) Linear regression; (2) random forest; (3) extreme gradient boosted tree; and (4) deep neural network.

The results obtained for each model with the two groups are seen in figure 3, highlighting the XGBoost and Random Forest for the prediction of young people.

Article 9 presents machine learning methods to predict internal migration routes (in one country - Yemen case study) for people displaced by the conflict to support humanitarian organizations that can assist this vulnerable population. The methods used are KNN (K-Nearest Neighbors), Random Forest, and Radial SVM (Support Vector Machines).

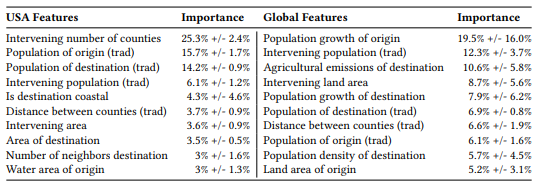


Fig. 2 Most relevant characteristics used in the XGBoost model.

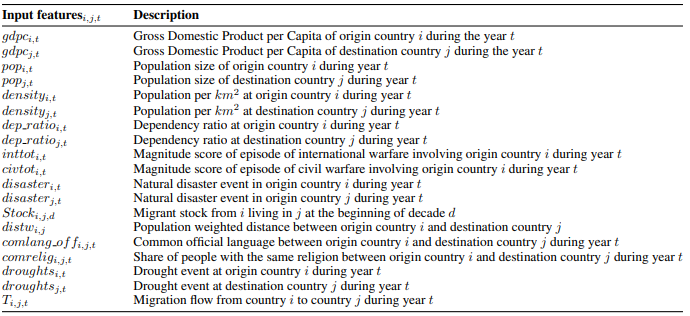


Fig. 3 Characteristics of a 3-layer neural network.

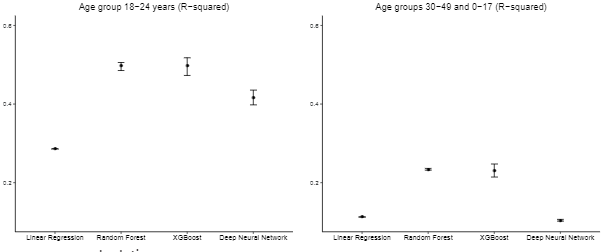


Fig. 4 Results obtained for each model.

In closing, article 10 presents architecture from the storage point of view for human migration data analysis systems with characteristics such as horizontal and vertical scalability and proposes a conceptual model for the data. This proposal can be complemented with the one presented in article 7.

IV. PROCESS OF OBTAINING AND GENERATING THE DATASET

The data are obtained from the Colombia open data portal [13] and correspond to the Colombian population resident and registered in the different consular missions abroad, which includes gender, age, level of studies, occupation, place of residence, the consulate of the constituency that serves it, among others.

The data correspond to 166,912 records and 9 characteristics. 152 classes (countries) are identified in the raw data and 99 in the processed data.

Below is a summary table and graphs of the data:

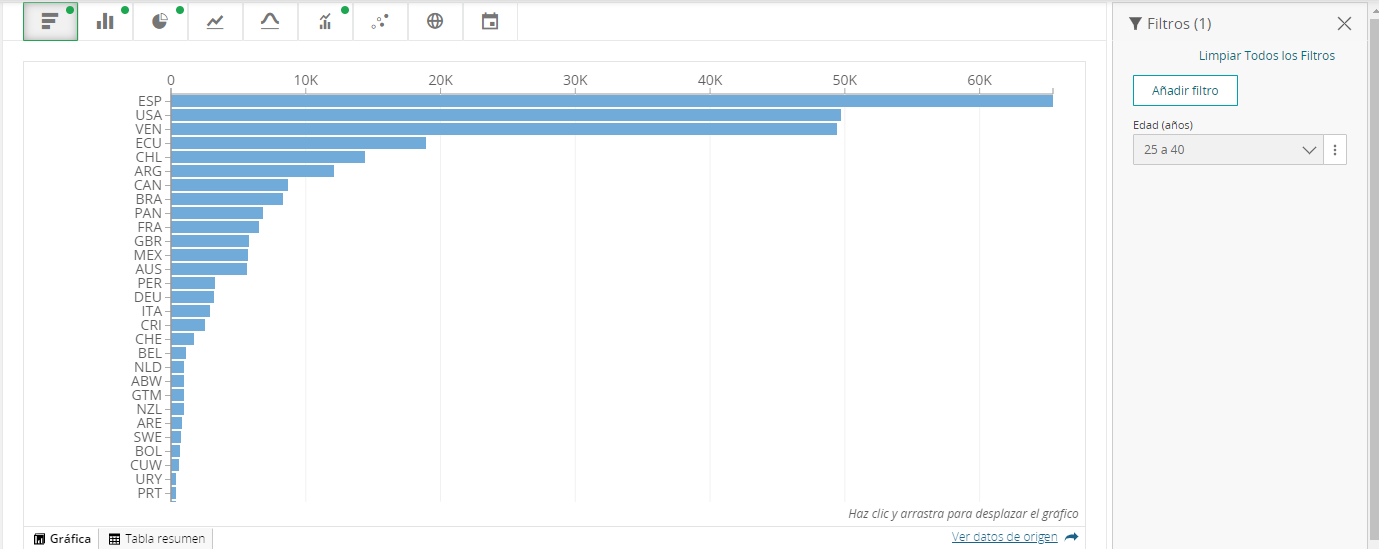


Fig. 5 Number of people who have migrated from Colombia to other countries. Taken from [13].

TABLE I.

DATA DESCRIPTION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fact** | **Feature or output** | **Type of data** | **Whether it applies or not** | **Description** |
| Country | Output | String | Applies | Indicates the country of residence of the person. |
| ISO country code | Output | String | N/A | ISO of the country of residence. |
| Registration office | Output | String | N/A | Consulate where the person made his consular registration. |
| Age group | Feature | String | Applies | The group is based on age according to the following: Between 25 and 28 years old Young adult Between 29 and 40 years old Adult |
| Age | Feature | String | Applies | Age of the person. When there is no registered data, the value -1 is presented. |
| Area | Feature | String | Applies | Occupancy area. When this information is not registered, the text is presented (DOES NOT INDICATE) |
| Sub area of ​​knowledge | Feature | String | Applies | Sub area of ​​occupation When this information is not registered, the text is presented (NOT INDICATED) |
| Academic level | Feature | String | Applies | Level of study. When this information is not registered, the text is presented (DOES NOT INDICATE) |
| Civil status | Feature | String | Applies | When this information is not registered, the text (UNKNOWN) is displayed |
| Gender | Feature | String | Applies | When this information is not registered, the text (UNKNOWN) is displayed |
| Ethnicity | Feature | String | Applies | When this information is not registered, the text (WITHOUT REGISTERED ETHNIC) is presented |
| Height | Feature | Integer | Applies | Height given in centimeters, the value -1 indicates that a registered height is not found |
| Amount of people | N/A | Integer | Applies | Number of people residing abroad who meet the demographic conditions |

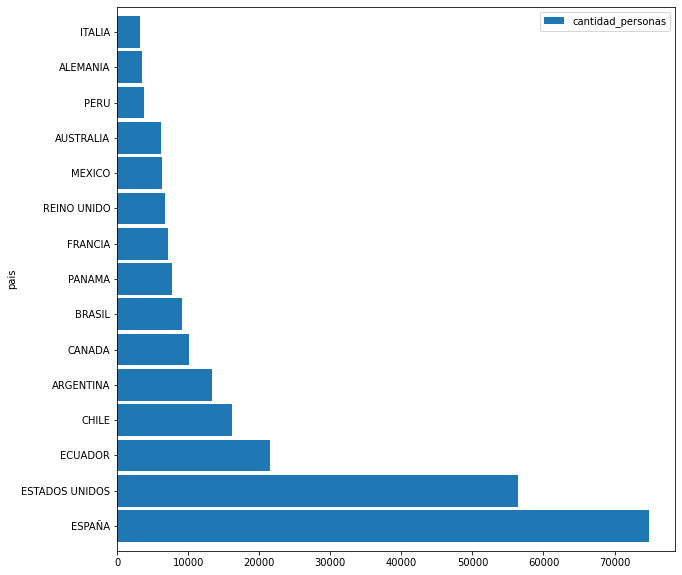


Fig. 6 Number of people who have migrated by country

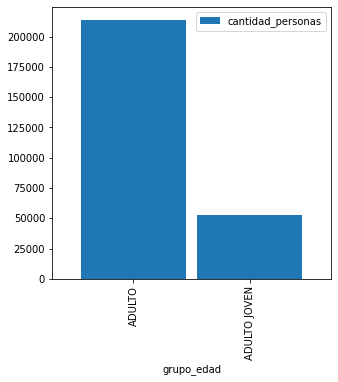
**

Fig. 7 Number of people by age group

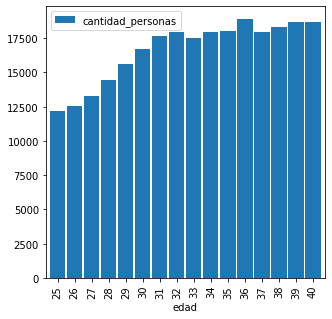


Fig. 8 Number of people by age

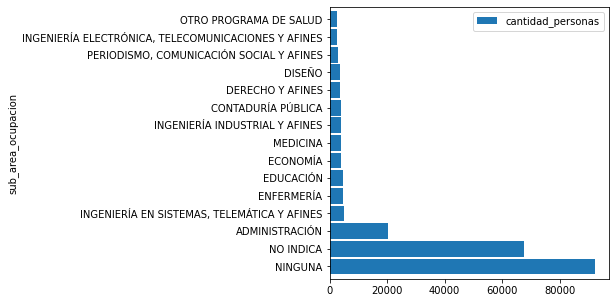


Fig. 9 Number of people by occupation subarea

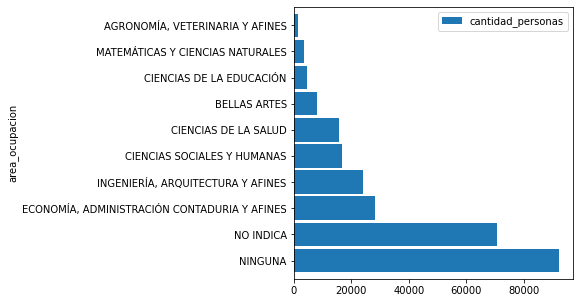


Fig. 10 Number of people per occupation area

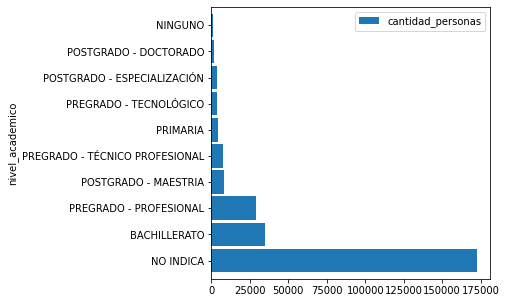


Fig. 11 Number of people per occupation area

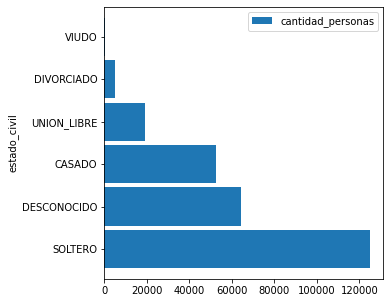


Fig. 12 Number of people by marital status

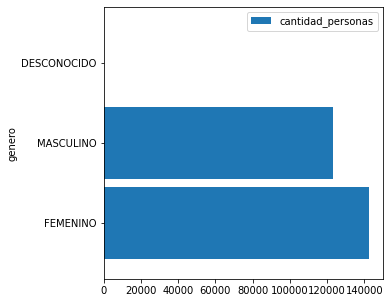


Fig. 13 Number of people by gender

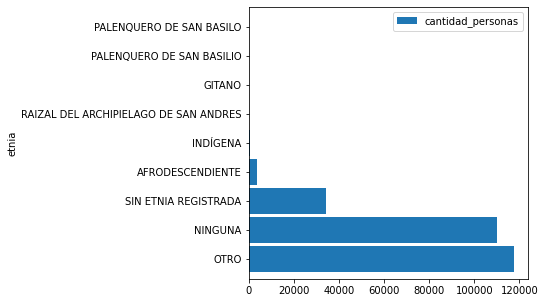


Fig. 14 Number of people by ethnicity

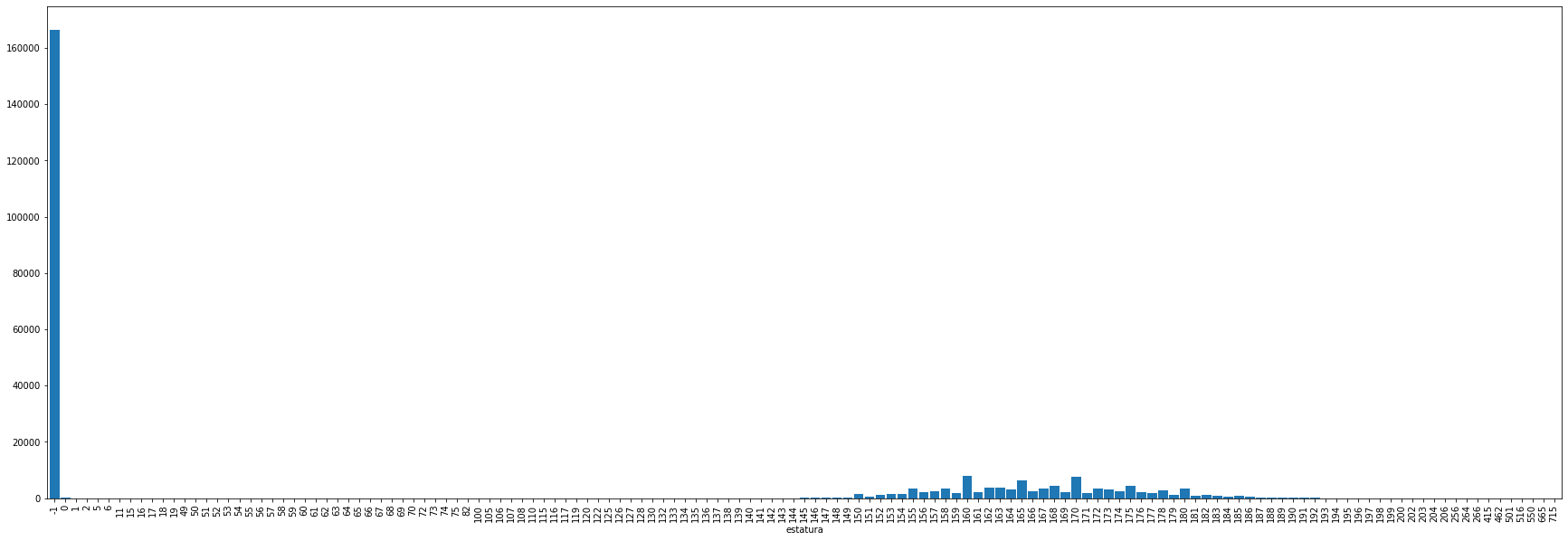


Fig. 15 Number of people per height

V. DATA PRE-PROCESSING

For data processing, 5 methods were applied which are explained below.

1. *Null data removal.*

the review of the dataset, it was found that there were values ​​where there was no registered information, and therefore it was categorized with specific labels; it is for this reason that it was decided to replace these categories with NaN and, subsequently, their respective elimination was made.

In this process, it went from having 167,842 records to having 19,497 records.

1. *Modification of anomalous data.*

It was identified through the graphs (Fig. 15 Number of people per height), that the height had outliers, this could be verified by making a whisker diagram, where we found that there were values ​​from 0 Cm to 550 Cm.

Below is the whisker graph where the outliers are evident:

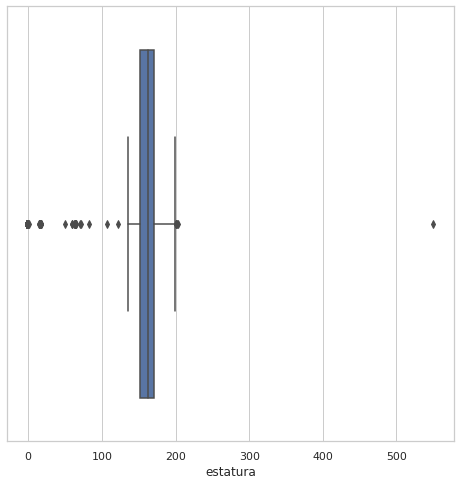


Fig. 16 Mustache diagram with outliers

This is why the quartiles technique was used, wherein the first instance, the quartiles, the interquartile range, and the respective upper and lower whiskers were calculated.

Once these calculations were made, a dataframe (BS) was created with the information of the records whose height was greater than the upper whisker and another dataframe (BI) with the information of the records smaller than the lower whisker.

Subsequently, the dataframe (BI) was traversed, and the values ​​of the main dataframe were modified to ensure that its height was equal to the value of the BI's height as nan, resulting in the identification of the lower-level outliers.

Once the outliers were identified, these values ​​identified by the median were changed. The previous procedure was also done for the BS, although as a difference, these outliers were changed by fashion.

Below is the whisker chart showing the process of identification and subsequent replacement of outliers:

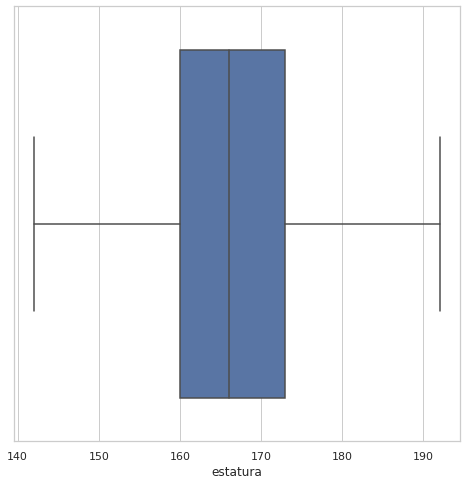


Fig. 17 Mustache diagram diagram without outliers

Once this process has been carried out, the height is displayed again and as will be seen below, the information looks much more appropriate.

In addition to the above, it was also identified that for the ethnicity field, the records of 'PALENQUERO DE SAN BASILO' should be modified by 'PALENQUERO DE SAN BASILIO' since they made reference to the same ethnic group, only that the dataset had problems of writing.

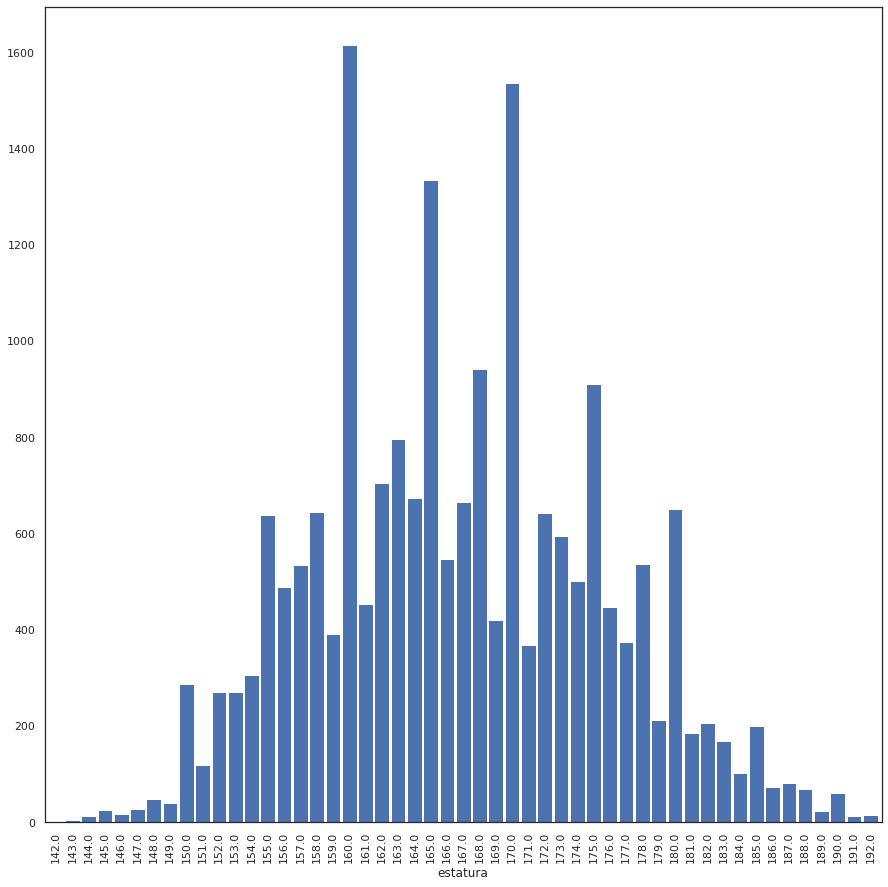


Fig. 18 Number of people per height without outliers

1. *Data explosion*

The number of people field in our dataset refers to the number of people who meet these characteristics matches, for example: If two people have the same country, age, height, ethnicity, etc. In the dataset, they will be represented by a single record; however, in a number of people, it will be a two referring to the number of people who comply with that data.

Given the above, it was necessary to explode the data; that is, new records were inserted by multiplying the information by the number of people in each record. It is important to clarify that once this part was done, the number of people data did not need to be used anymore in the process.

In this process, two dataframes were created in the first instance with the information of the data whose field quantity\_people are different from 1 (data\_different\_1) and another with the records of quantity\_people equal to 1 (data\_equal\_1). A summation of the number\_people field was made in the dataframe data\_different\_1 to identify the number of new records (1290) that should be inserted.

Subsequently, the dataframe data\_different\_1 was traversed to multiply the information, that is, insert the number of records equal to your information as indicated by its number\_people field, creating a new dataframe called data\_tem which will contain these new records.

1. *Data categorization*

Within the data processing, more specifically for the dataset training, since some of the algorithms only allowed numerical data and therefore it was necessary to carry out categorization of the qualitative characteristics, to carry out this process, we converted to category each of the qualitative factors such as gender, ethnicity, civil\_status, etc.

1. *Balancing the data.*

An imbalance was identified in the data, with a high concentration in Spain and the United States, as seen in Fig 6. This affects the algorithm in generalization processes with minority classes.

To solve this problem, 3 data balancing models were tested, which "create" synthetic data from the data that the class has to balance. It should be noted that tests were carried out with the imblearn class using variations of its class and different parameters. The following algorithms were found:

● SMOTE: For its acronym in English, "Synthetic Minority Oversampling Technique", it is an algorithm that balances the minority classes to leave them with the same amount of data as the majority class. A string type parameter is defined ('monority',' not minority', 'not majority', 'all'), indicating which classes you want to resample. It applies to the SMOTE class and all its variations. [19]

● SMOTETomek: It is a type of combined balancing; it balances the minority classes and the majority classes until having an even number of data. A helper algorithm is added to the SMOTE, the Tomek link, which is used to find a nearest opposite neighbor, and the algorithm decides whether to eliminate the pair or just one.

● SMOTE-NC: It is a type of balancing for both continuous and categorical characteristics, using the SMOTE logic.

The SMOTETomek model was chosen, with the parameter 'all' (it resamples all the classes) since it is the one that gives the best result when the data is balanced. First, however, a data elimination had to be applied so that the algorithms could work; in this case, all the countries that had 6 or fewer data were eliminated, this is because all the algorithms that were tested require at least 6 data to be able to make relationships and create synthetic data.

VI. ENTRENAMIENTO Y EVALUACIÓN

Based on the solutions found in the articles and the models seen in class, the following learning models were used:

XGBoost

It is a decision tree-based ensemble machine learning algorithm that uses a gradient impulse framework. In prediction problems involving unstructured data (images, text, etc.), artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small to medium-sized tabular/structured data, decision tree-based algorithms are considered best in class at this time. [14]

Artificial neural networks

ANNs are information processing systems whose structure and function are inspired by biological neural networks (Hilera and Martínez, 1995). Each neuron is connected to other neurons through communication links, each of which has an associated weight. The weights represent the information that will be used by the neural network to solve a given problem. [fifteen]

KNN (K-Nearest Neighbors)

K's closest neighbor is one of the simplest machine learning algorithms based on supervised learning technique. K-NN algorithm assumes similarity between new case / data and available cases and places new case in category which is more similar to the categories available. [16]

Random Forest

As the name implies, it consists of a large number of individual decision trees that operate as a set. Each individual tree in the random forest casts a class prediction, and the class with the most votes becomes our model prediction. [17]

Radial SVM (Support Vector Machines)

They are a famous and very robust classification technique that does not use any kind of probabilistic model like any other classifier, but simply generates hyperplanes or simply places lines, to separate and classify the data in some feature space in different regions. In this particular case, the Kernel or mathematical function used is radial based. [18]

VII. ANALYSIS AND EXPLANATION OF PERFORMANCE METRICS

Metrics implemented to analyze the different models:

*Test set score, Train set score, mean\_squared\_error, accuracy\_score, precision\_score, recall\_score, f1\_score*

The models were applied to the data without any type of processing, then with the processing without balancing and finally with all the complete preprocessing techniques defined in section V. The results are recorded for each case.

1. *Data without processing.*

TABLE II.

PERFORMANCE METRICS FOR MODELS USING DATA WITHOUT PROCESSING

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | ***Test set score*** | ***Train set score*** | ***mean\_squared\_error*** | ***accuracy\_score*** | ***precision\_score*** | ***recall\_score*** | ***f1\_score*** |
| XGBoost | 0,29 | 0,296 | 936.953 | 0.291 | 0.197 | 0.291 | 0.196 |
| RNA | 0.29 | 0.30 | 948.708 | 0.289 | 0.195 | 0.289 | 0.198 |
| KNN | 0.13 | 0.51 | 1857.679 | 0.135 | 0.132 | 0.135 | 0.134 |
| Random Forest | 0.175 | 0.528 | 1494.836 | 0.175 | 0.139 | 0.175 | 0.152 |
| SVM | 0.25 | 0.48 | 980.379 | 0.246 | 0.150 | 0.246 | 0.164 |

It can be seen that with the data without processing, the results of the metrics are very poor for all the models. The RNA and XGBoost models obtain the best results for the Test set score; they are the ones with the lowest mean\_squared\_error and stand out for the accuracy\_score, the precision\_score, the recall\_score, and the f1\_score. The KNN and Random Forest models present a better result for the Train set score.

1. *Data processed without balancing*

TABLE III.

PERFORMANCE METRICS FOR MODELS USING NON-BALANCED PROCESSED DATA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | ***Test set score*** | ***Train set score*** | ***mean\_squared\_error*** | ***accuracy\_score*** | ***precision\_score*** | ***recall\_score*** | ***f1\_score*** |
| XGBoost | 0.348 | 0.534 | -4.838 | 0.348 | 0.257 | 0.348 | 0.271 |
| RNA | 0.36 | 0.37 | -4.338 | 0.357 | 0.255 | 0.357 | 0.260 |
| KNN | 0.21 | 0.86 | -5.857 | 0.206 | 0.199 | 0.206 | 0.202 |
| Random Forest | 0.244 | 0.809 | -4.405 | 0.244 | 0.220 | 0.244 | 0.230 |
| SVM | 0.31 | 0.78 | -6.560 | 0.309 | 0.199 | 0.309 | 0.197 |

It can be observed that for the data without processing the results of the metrics improve compared to those obtained in the models with data without processing. The RNA and XGBoost models obtain better results for the accuracy\_score, the precision\_score, the recall\_score, the f1\_score and the Test set score, while for the Train set score the best results are obtained for the KNN and Random Forest models. SVM and KNN have the lowest mean\_squared\_error, although the difference is minimal compared to the other models.

1. *Fully processed data*

TABLE IV

PERFORMANCE METRICS FOR MODELS USING FULLY PROCESSED DATA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | ***Test set score*** | ***Train set score*** | ***mean\_squared\_error*** | ***accuracy\_score*** | ***precision\_score*** | ***recall\_score*** | ***f1\_score*** |
| XGBoost | 0.871 | 0.949 | 1.671 | 0.871 | 0.867 | 0.871 | 0.867 |
| RNA | 0.40 | 0.40 | 4.414 | 3.96 | 0.355 | 0.396 | 0.343 |
| KNN | 0.92 | 0.99 | 1.092 | 0.921 | 0.916 | 0.921 | 0.918 |
| Random Forest | 0.875 | 0.982 | 1.584 | 0.875 | 0.871 | 0.875 | 0.872 |
| SVM | 0.90 | 0.96 | 1.267 | 0.901 | 0.904 | 0.901 | 0.900 |

In the models used with the fully processed data, the best results obtained for all the metrics correspond to KNN.

VIII. CONCLUSIONS

From the different tests carried out, the importance of data preprocessing can be identified. The results of the different models improve when the different preprocessing techniques such as null data elimination, anomalous data modification, and data balancing begin to be applied.

For this particular case, it was essential to balance the data based on the investigated techniques, as the number of records for each of the classes differed significantly. Therefore, the selected SMOTETomek balancing corresponds to a combined method, which performs resampling on all characteristics.

The best results were obtained from a runtime and metric point of view for KNN for fully processed data. However, the RNA and XGBoost models presented the best results for the data without processing. That is precisely one of the main advantages of the latter model: they require much less cleaning and preprocessing of the data than other learning methods, and they are not visible, heavily influenced by outliers.

An important variable when obtaining a machine learning model is the processing time and performance required for the respective hardware. For the analyzed models, the longest training times for Artificial Neural Networks and Support Vector Machines were presented.

REFERENCES

[1] U. Nations, "Migración | Naciones Unidas", United Nations. [Online]. Available: https://www.un.org/es/global-issues/migration. [Accessed: 11- Jun- 2021].

[2]"▷Tipos de migración humana en el mundo y sus causas", Ayuda en Acción, 2018. [Online]. Available: https://ayudaenaccion.org/ong/blog/ayuda-humanitaria/tipos-de-migracion-humana/. [Accessed: 11- Jun- 2021].

[3] C. Robinson and B. Dilkina, "A Machine Learning Approach to Modeling Human Migration", Dl.acm.org, 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3209811.3209868. [Accessed: 11- Jun- 2021].

[4] H. Silvere Kiossou, Y. Schenk, F. Docquier, V. Ratheil, S. Nijssen and P. Schaus, "Using an interpretable Machine Learning approach to study the drivers of International Migration", Arxiv.org, 2020. [Online]. Available: https://arxiv.org/pdf/2006.03560.pdf. [Accessed: 11- Jun- 2021].

[5] A. Beduschi, "International migration management in the age of artificial intelligence", Watermark.silverchair.com, 2020. [Online]. Available: https://watermark.silverchair.com/mnaa003.pdf?token=AQECAHi208BE49Ooan9kkhW\_Ercy7Dm3ZL\_9Cf3qfKAc485ysgAAAsgwggLEBgkqhkiG9w0BBwagggK1MIICsQIBADCCAqoGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMMa7-uCaLrPZoWf-WAgEQgIICe\_zq-XjqRM425UpVPhUDxFISI6kzu3Xcw-3a7YNyHz8qnK9VIXfLWi1gwX5SzsTGFVpRf-FcePzjmqGdaH1t2X2EkfdLn7h8tOigkuRDN-sYuoGlVIBFZ1YiGxxasliZdhxYfGzZcapdr8Jbp-6ts8ASyprdPHtl-wjK1JldPJbAmnmvwUwWKARTiP7M4R\_bPYo8Tic\_GOdpMPgoNgp4c0cCwfWZiFcefsUnTYHBccx94ZvPyrGEUDlno6PiuzncCRPH0kcNNczfJ9-FOftDBNhvrDigtoTgSCuIaksjHxPqhsHTMzehQWn0lkl\_Nt1KXpX\_Y4DMnEKEz2kJTZJPSlH\_RzYFuDmYMb3Ob27zgYuPG9inVUoGMmqcvrjq45rkB5s0t\_101A0UAFpu0a5PkhyL-iSvO6dOgsYSEJ5X9WCHrJjorbyAfK3AtsRO6ham6YUh7GRFgQtX2ubY25tpXx2jevVMvYDeg--9RpLAa0q0YMRca8zybJdSOmKOiXXUE83EOw7LOUQM\_UEE-Z5TYGumu-RpEK565hRlUqdRsmNnDQaapoPD5yB2z8gYaChq9aj-5Cl0UE6Lw4Jl1ecMEl\_I8CBPB1BiXaDe9hwyhy0lHJurXfxK\_cp-Eb90NIOCR3B8Ai4m\_U5PicVKEFGAiCPZnIPPIO\_vqbE2bS-AhafESPnVHX3bDBiFpVUhjYMVHf9AAa9mXzgYHlypzK6kdmsrHTVH2hTqn8\_CcaPEU0V0gPQ8a3W9WlQCkcZKqA1OEh3-m\_w92Vw-H\_IgXvQVlFULSHRux1vAJK5tN1iHT6eVVkx6KYRWcKWrqSUqVHe893F8hEUP\_KiUXNAR. [Accessed: 11- Jun- 2021].

[6] C. Zhan, C. Tse, Y. Fu, Z. Lai and H. Zhang, "Modeling and prediction of the 2019 coronavirus disease spreading in China incorporating human migration data", https://journals.plos.org, 2019. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241171. [Accessed: 11- Jun- 2021].

[7] L. Singh et al., "Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration", Dl.acm.org, 2019. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3292500.3330774. [Accessed: 11- Jun- 2021].

[8] H. Weber, "View of How Well Can the Migration Component of Regional Population Change be Predicted? A Machine Learning Approach Applied to German Municipalities", Comparativepopulationstudies.de, 2020. [Online]. Available: https://comparativepopulationstudies.de/index.php/CPoS/article/view/328/308. [Accessed: 11- Jun- 2021].

[9] D. Dominguez, P. Soria, M. Gonzalez, F. Rodríguez Ortiz and A. Sanchez, "A Classification and Data Visualization Tool Applied to Human Migration Analysis", https://www.researchgate.net/publication/333765430\_A\_Classification\_and\_Data\_Visualization\_Tool\_Applied\_to\_Human\_Migration\_Analysis, 2019. [Online]. Available: https://www.researchgate.net/publication/333765430\_A\_Classification\_and\_Data\_Visualization\_Tool\_Applied\_to\_Human\_Migration\_Analysis. [Accessed: 11- Jun- 2021].

[10] M. Greceque, J. DeWaard, J. Hellmann and G. Abel, "Climate Vulnerability and Human Migration in Global Perspective", mdpi, 2017. [Online]. Available: https://www.mdpi.com/2071-1050/9/5/720. [Accessed: 11- Jun- 2021].

[11] E. Harrison, "Modeling Movement: A machine-learning approach to track migration routes after displacement", W&M ScholarWorks, 2020. [Online]. Available: https://scholarworks.wm.edu/honorstheses/1531/. [Accessed: 11- Jun- 2021].

[12] L. Rudikowa, V. Denilchik, I. Savenkov, A. Nenko and S. Sobolevsky, "General Concept of the Storage and Analytics System for Human Migration Data", Ezproxy.uao.edu.co, 2018. [Online]. Available: https://ezproxy.uao.edu.co:2150/content/pdf/10.1007%2F978-3-030-02843-5\_21.pdf. [Accessed: 11- Jun- 2021].

[13]"Colombianos en el exterior de 25 a 40 años", Datos.gov.co, 2021. [Online]. Available: https://www.datos.gov.co/Estad-sticas-Nacionales/Colombianos-en-el-exterior-de-25-a-40-a-os/3k8j-4d42. [Accessed: 11- Jun- 2021].

[14] V. Morde, "XGBoost Algorithm: Long May She Reign!", Medium, 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. [Accessed: 11- Jun- 2021].

[15] P. A. and M. J.J., "¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones", Disi.unal.edu.co, 1999. [Online]. Available: https://disi.unal.edu.co/~lctorress/RedNeu/LiRna001.pdf. [Accessed: 11- Jun- 2021].

[16][2]"K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint", www.javatpoint.com. [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning. [Accessed: 11- Jun- 2021].

[17]T. Yiu, "Understanding Random Forest", Medium, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed: 11- Jun- 2021].

[18]"1.4. Support Vector Machines — scikit-learn 0.24.2 documentation", Scikit-learn.org, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html#kernel-functions. [Accessed: 11- Jun- 2021].

[19]"SMOTE — Version 0.8.0", Imbalanced-learn.org, 2021. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over\_sampling.SMOTE.html. [Accessed: 11- Jun- 2021].

[20]"SMOTETomek — Version 0.9.0.dev0", Imbalanced-learn.org, 2021. [Online]. Available: https://imbalanced-learn.org/dev/references/generated/imblearn.combine.SMOTETomek.html. [Accessed: 11- Jun- 2021].

[21]"Tackling class imbalance with SVM-SMOTE", Medium, 2016. [Online]. Available: https://medium.com/vclab/tackling-class-imbalance-with-svm-smote-efa41ec3de5f. [Accessed: 11- Jun- 2021].

[22]"SMOTENC — Version 0.9.0.dev0", Imbalanced-learn.org, 2021. [Online]. Available: https://imbalanced-learn.org/dev/references/generated/imblearn.over\_sampling.SMOTENC.html. [Accessed: 11- Jun- 2021].