Sistema para la identificación del país de destino en procesos de migración colombianos

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

**Resumen— Desde la antigüedad el ser humano ha estado en constante tránsito. Las causas son múltiples: búsqueda de trabajo o de nuevas oportunidades económicas, para reunirse con sus familiares, para estudiar, escapar de conflictos, persecuciones, del terrorismo o de violaciones o abusos de los derechos humanos. Con este trabajo se analizan diferentes modelos de machine learning para predecir el país destino de los colombianos que emigran a partir de información obtenida en el portal de datos abiertos Colombia.**

I. INTRODUCCIÓN

La migración hace referencia al desplazamiento de una persona o un grupo de personas desde el lugar que habitan (su residencia) hasta otro: es decir, a grandes rasgos, un cambio de residencia. En el caso de la salida de las personas de una región o país, se utiliza el término *emigración* y, desde el país de destino, este fenómeno se entiende como *inmigración*.

Las causas de la migración incluyen, entre otras: búsqueda de trabajo o de nuevas oportunidades económicas, para reunirse con sus familiares, para estudiar, escapar de conflictos, persecuciones, del terrorismo o de violaciones o abusos de los derechos humanos. Algunos lo hacen debido a los efectos adversos del cambio climático, desastres naturales u otros factores ambientales.

“En la actualidad, el número de personas que vive en un país distinto de su país natal es mayor que nunca. Según el [Informe sobre las migraciones en el mundo 2020 de la OIM](https://publications.iom.int/system/files/pdf/wmr_2020_es.pdf), a junio de 2019 se estimaba que el número de migrantes internacionales era de casi 272 millones en todo el mundo, 51 millones más que en 2010. Casi dos tercios eran migrantes laborales. Los migrantes internacionales constituían el 3,5% de la población mundial en 2019, en comparación con el 2,8% en 2000 y el 2,3% en 1980”. [1]

“La migración humana se clasifica mediante seis grandes grupos o subcategorías: según su escala geográfica, las características del lugar de origen y destino, su temporalidad, su grado de libertad, su causa y según la edad de los migrantes”. [2]

En este trabajo buscaremos a partir de la Información brindada en datos abiertos Colombia de la población colombiana residente y registrada en las diferentes misiones consulares en el exterior que incluye el género, la edad, el nivel de estudios, la ocupación, lugar de residencia, el consulado de la circunscripción que lo atiende, entre otros, definir un modelo que permita identificar el país destino del migrante.

II. DESCRIPCIÓN DEL FENÓMENO, PROCESO MODELADO Y PROBLEMÁTICA A ABORDAR

En este proyecto, se abordará el fenómeno de la emigración en Colombia a partir de los datos de colombianos en el exterior de 25 - 40 años.

La pregunta para el problema se define a continuación:

*¿Es posible identificar a qué país se desplaza la población entre 25 y 40 años, a partir de su área de conocimiento, edad, género, estado civil, nivel académico?*

Este problema es importante por su relevancia a nivel internacional ante situaciones como la actual pandemia para análisis de propagación del virus por procesos de migración, o en el caso de personas desplazadas por el conflicto, para apoyar organizaciones humanitarias que puedan brindar asistencia a esta población.

A nivel nacional hemos orientado el problema en la identificación de tendencias de oportunidades de desplazamientos a otros países por profesión, edad, género, estado civil y/o nivel académico.

*A. Proceso Modelado*

El proceso modelado se presenta en el siguiente diagrama esquemático.

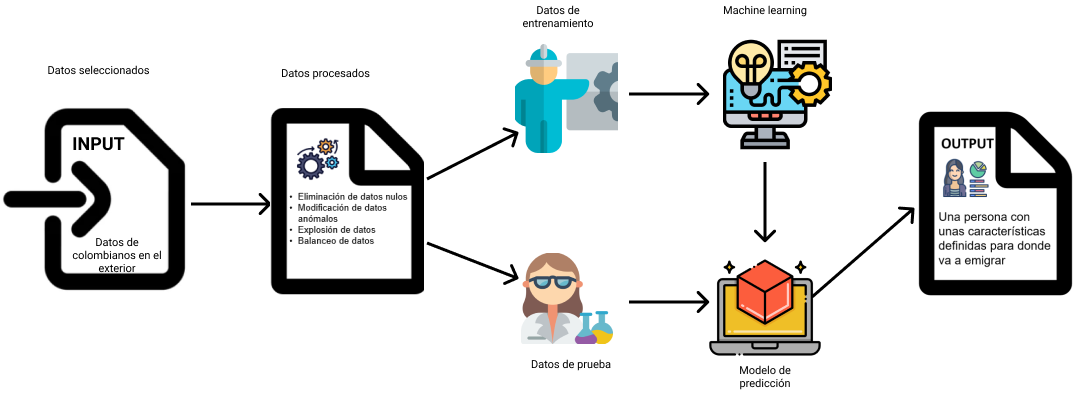
****

Fig. 1 Diagrama esquemático del proceso de solución

III. REVISIÓN DE LITERATURA

A continuación, se listan los artículos revisados con soluciones similares.

* A Machine Learning Approach to Modeling Human Migration.
* Using an interpretable Machine Learning approach to study the drivers of International Migration.
* International migration management in the age of artificial intelligence.
* Modelización y predicción de la propagación de la enfermedad del coronavirus de 2019 en China incorporando datos sobre migración humana.
* 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.
* Vulnerabilidad climática y migración humana en perspectiva global.
* Modeling Movement: A machine-learning approach to track migration routes after displacement.
* General Concept of the Storage and Analytics System for Human Migration Data.

Los artículos [3] y [4] presentan alternativas con modelos de aprendizaje automático a los modelos tradicionales de movilidad humana. El primer modelo tradicional denominado “gravity model” define que la probabilidad de un viaje entre dos ubicaciones se obtiene directamente a partir de la distancia entre ellas. El segundo modelo tradicional llamado “radiation model” postula que

la probabilidad de un viaje no depende tanto de la distancia sino de la cantidad de oportunidades intermedias (este modelo se amplía en el artículo [7]). Estos modelos se basan en relaciones lineales entre variables independientes.

Las alternativas propuestas en los artículos [3] y [4] se basan en árboles (modelos XGBoost) y redes neuronales que permiten combinar de forma no lineal otras características provistas por el Banco Mundial o el Censo de Estados Unidos.

El artículo [3] desarrolla la primera propuesta de aprendizaje automático del problema de predicción de la migración humana e incluye procedimientos para trabajar con el desequilibrio del conjunto de datos, el ajuste de hiper parámetros y la evaluación del desempeño. Así mismo, para la red neuronal desarrollan una función de pérdida a la medida o personalizada. Finalmente, a partir de dos datasets (el USA Migration dataset y el Global Migration dataset), logran encontrar un mejor desempeño de los modelos de aprendizaje automático comparado con los modelos tradicionales. A continuación, se presenta la figura 1, que contiene el Top10, por cada dataset, de las características más relevantes utilizadas en el modelo XGBoost.

El artículo [4] utiliza una red neuronal de 3 capas con las siguientes características mostradas en la figura 2.

En los dos casos utilizan métricas como el RMSE (Root mean squared error) y r2 (Coefficient of determination), entre otras.

Los otros artículos complementan el contexto de análisis, presentando, por ejemplo, en el artículo 3 como los sistemas de inteligencia artificial están apoyando los procesos de migración y toma de decisiones de asilo en países como Canadá y Alemania. También se encuentran artículos que estudian causas de migración como los cambios climáticos (artículo [10]) y que amplían aplicaciones a partir del análisis de movilidad humana como en la era del COVID (artículo [6]).

El artículo [8] presenta modelos para predecir la migración en Alemania para dos grupos demográficos: gente joven entre 18 y 24 años y población familia con edades de 30 a 40 años y de 0 a 17 años. Los modelos utilizados fueron: (1) Linear regression; (2) random forest; (3) extreme gradient boosted tree; y (4) deep neural network.

Los resultados obtenidos para cada modelo con los dos grupos, se ven en la figura 3, destacándose para la predicción de gente joven el XGBoost y Random Forest.

El artículo [11] presenta métodos de machine learning para predecir rutas de migración internas (en un país - caso de estudio Yemen) para personas desplazadas por el conflicto, con el objetivo de apoyar organizaciones humanitarias que puedan brindar asistencia a esta población vulnerable. Entre los métodos utilizados se encuentran KNN (K-Nearest Neighbors), Random Forest y Radial SVM (Support Vector Machines).

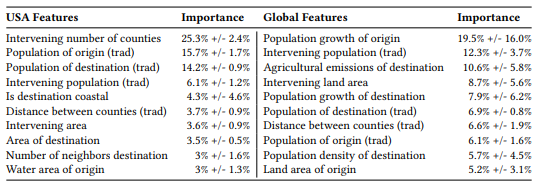


Fig. 2 Características más relevantes utilizadas en el modelo XGBoost.

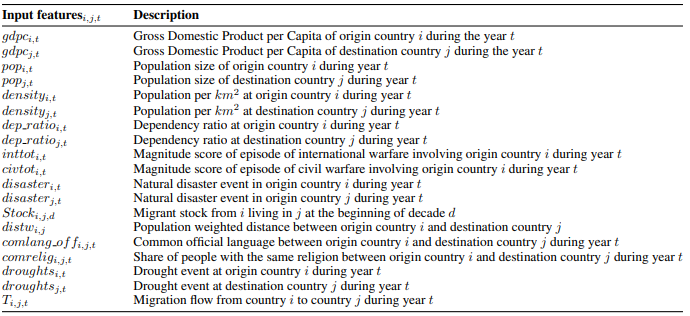


Fig. 3 Características de una red neuronal de 3 capas.

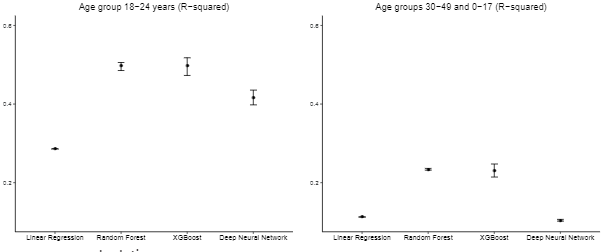


Fig. 4 Resultados obtenidos para cada modelo.

Finalmente, el artículo [12] presenta una arquitectura desde el punto de vista de almacenamiento para sistemas de análisis de datos de migración humana con características como la escalabilidad horizontal y vertical y propone un modelo conceptual para los datos. Esta propuesta se puede complementar con la presentada en el artículo 7.

IV. PROCESO DE OBTENCIÓN Y GENERACIÓN DEL DATASET

Los datos se obtienen del portal de datos abiertos Colombia [13] y corresponden a la población colombiana residente y registrada en las diferentes misiones consulares en el exterior que incluye el género, la edad, el nivel de estudios, la ocupación, lugar de residencia, el consulado de la circunscripción que lo atiende, entre otros.

Los datos corresponden a 166.912 registros y 9 características. Se identifican 152 clases (países) en los datos sin procesar y 99 con los datos procesados.

A continuación, se presenta una tabla resumen y los gráficos de los datos:

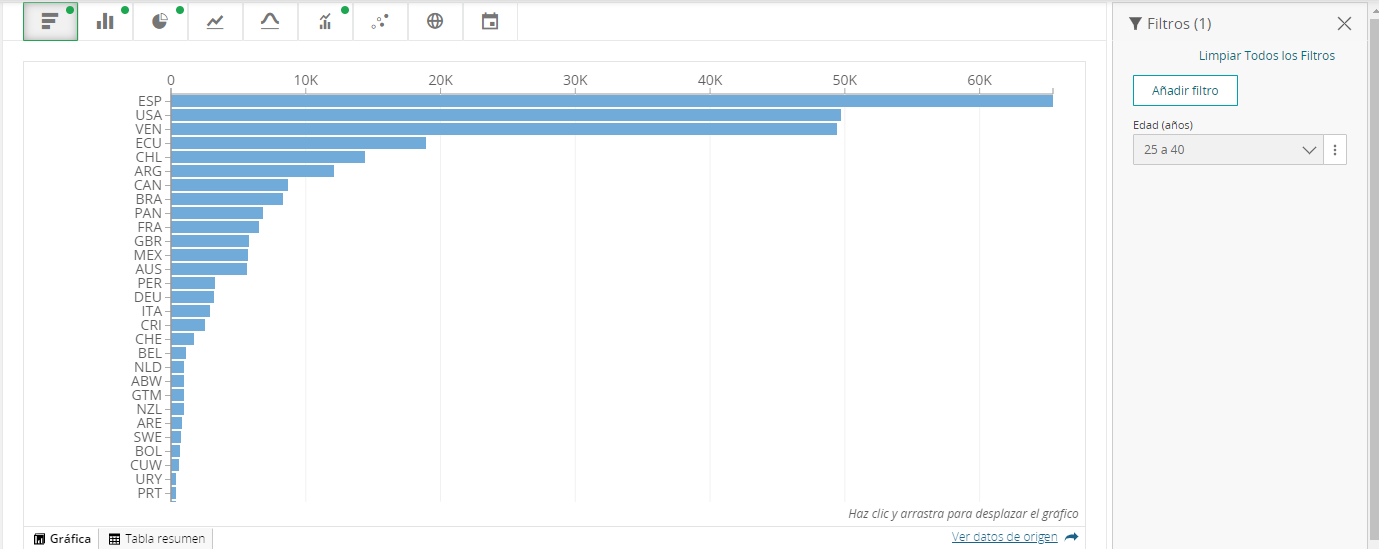


Fig. 5 Número de personas que han migrado desde Colombia a otros países. Tomada de [13].

TABLA I.

DESCRIPCIÓN DE LOS DATOS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dato** | **Característica o Salida** | **Tipo de datos** | **Si aplica o no aplica** | **Descripción** |
| País | Salida | String | Si aplica | Indica el país de residencia de la persona. |
| Código ISO país | Salida | String | N/A | ISO del país de residencia. |
| Oficina de registro | Salida | String | N/A | Consulado donde la persona realizó su registro consular. |
| Grupo de edad | Característica | String | Si aplica | El grupo se basa en la edad de acuerdo con lo siguiente: Entre 25 y 28 años Adulto joven Entre 29 y 40 años Adulto |
| Edad | Característica | String | Si aplica | Edad de la persona. Cuando no se tienen datos registrados se presenta el valor -1. |
| Área | Característica | String | Si aplica | Área de ocupación. Cuando no se tiene registrada esta información se presenta el texto (NO INDICA) |
| Sub área de conocimiento | Característica | String | Si aplica | Sub área de ocupación.Cuando no se tiene registrada esta información se presenta el texto (NO INDICA) |
| Nivel académico | Característica | String | Si aplica | Nivel de estudio. Cuando no se tiene registrada esta información se presenta el texto (NO INDICA) |
| Estado civil | Característica | String | Si aplica | Cuando no se tiene registrada esta información se presenta el texto (DESCONOCIDO) |
| Género | Característica | String | Si aplica | Cuando no se tiene registrada esta información se presenta el texto (DESCONOCIDO) |
| Etnia | Característica | String | Si aplica | Cuando no se tiene registrada esta información se presenta el texto (SIN ETNIA REGISTRADA) |
| Estatura | Característica | Entero | Si aplica | Estatura dada en centímetros, el valor -1 indica que no se encuentra una estatura registrada |
| Cantidad de personas | Por definir | Entero | Si aplica | Cantidad de personas residentes en el exterior que cumplen con las condiciones demográficas |

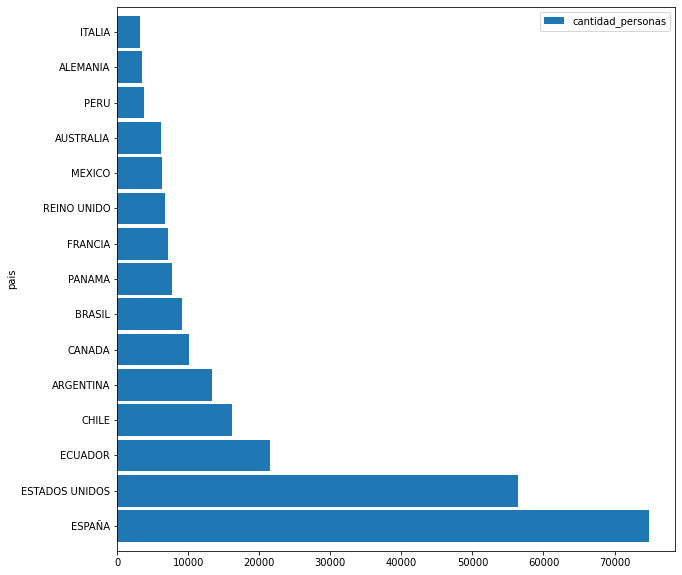


Fig. 6 Número de personas que han migrado por país

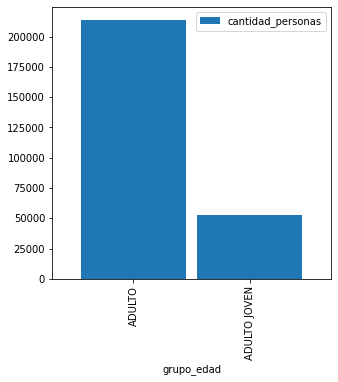
**

Fig. 7 Número de personas por grupo de edad

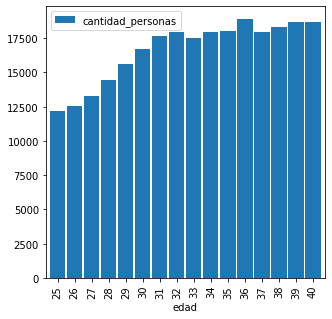


Fig. 8 Número de personas por edad

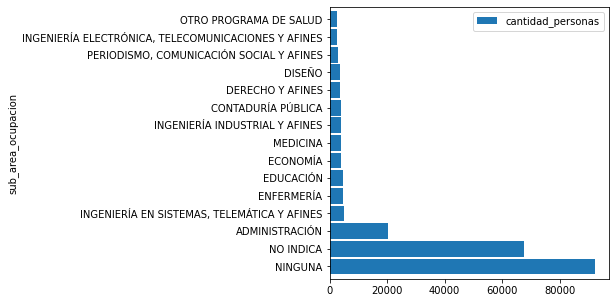


Fig. 9 Número de personas por subárea de ocupación

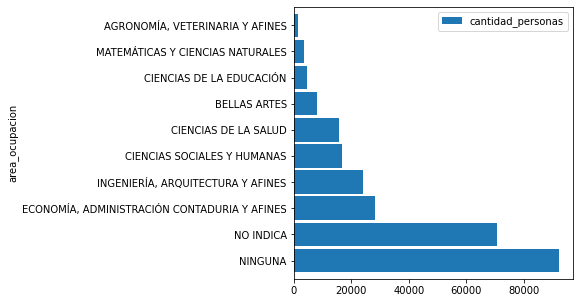


Fig. 10 Número de personas por área de ocupación

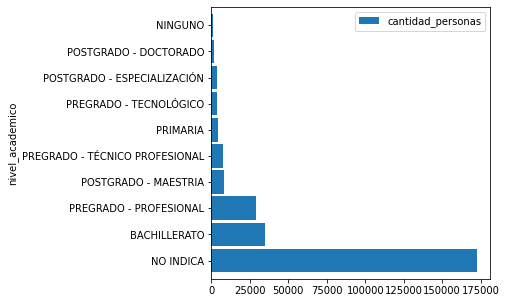


Fig. 11 Número de personas por nivel académico

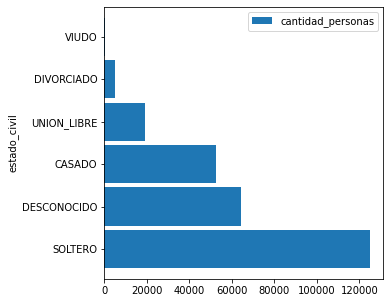


Fig. 12 Número de personas por estado civil

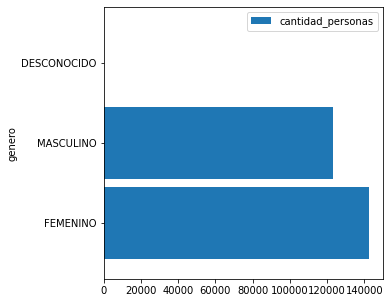


Fig. 13 Número de personas por género

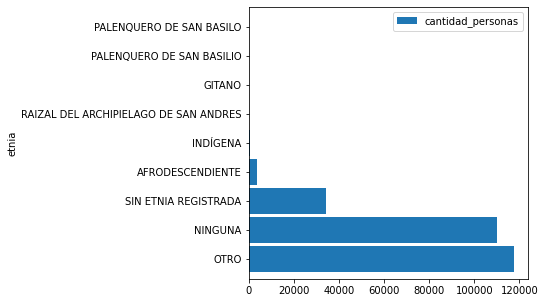


Fig. 14 Número de personas por etnia

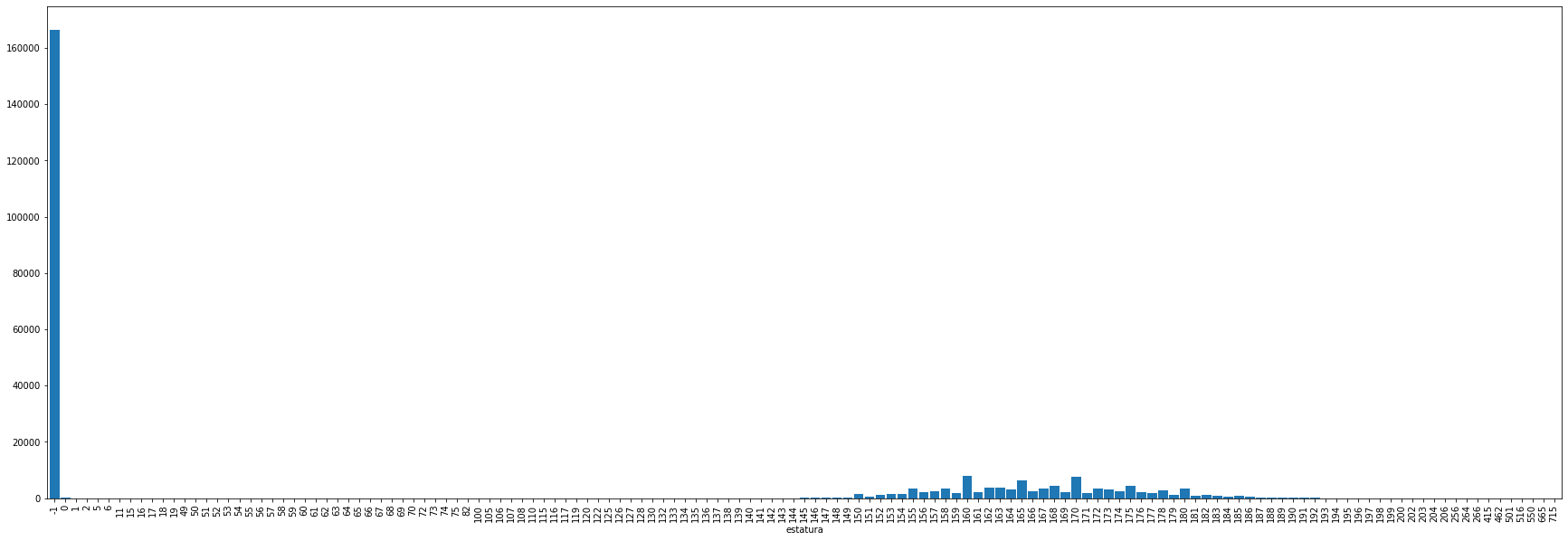


Fig. 15 Número de personas por estatura

V. PREPROCESAMIENTO DE DATOS

Para el procesamiento de datos, se aplicaron 5 métodos que se explican a continuación.

1. *Eliminación de datos nulos.*

En la revisión del dataset, se encontró que existían valores en donde no se encontraba información registrada y por lo tanto venía categorizada con etiquetas específicas, es por esta razón que se decidió reemplazar estas categorías por NaN y posteriormente se hizo su respectiva eliminación.

En este proceso se pasó de tener 167.842 registros a tener 19.497 registros.

1. *Modificación de datos anómalos.*

Se identificó a través de las gráficas (Fig. 15 Número de personas por estatura), que la estatura tenía outliers, esto se pudo comprobar al realizar un diagrama de bigotes, en donde encontramos que existían valores desde 0 Cm a 550 Cm.

A continuación, se presenta el gráfico de bigotes en donde se evidencian los outliers:

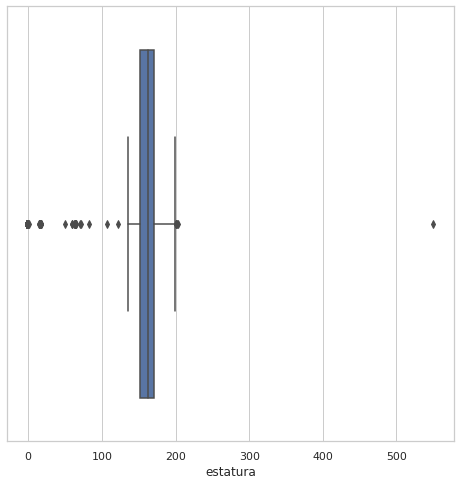


Fig. 16 Diagrama de bigotes con outliers

Es por lo anterior, que se utilizó la técnica de los cuartiles, en donde en primera instancia se calcularon los cuartiles, el rango Inter cuartil y los respectivos bigote superior e inferior.

Una vez tenidos estos cálculos, se crearon un dataframe (BS) con la información de los registros cuya estatura fueran mayores al bigote superior y otro dataframe (BI) con la información de los registros menores al bigote inferior.

Posteriormente, se recorrió el dataframe (BI) y se modificaron los valores del dataframe principal que cumplieran con que su estatura fuese igual al valor de la estatura de BI como nan, dando como resultado la identificación de los outliers del nivel inferior.

Una vez identificados los outliers, se procedió a cambiar estos valores identificados por la mediana. El anterior procedimiento se hizo también para los BS, aunque como diferencia estos outliers fueron cambiados por la moda.

A continuación, se presenta el gráfico de bigotes en donde se evidencia el proceso de identificación y posterior reemplazo de outliers:

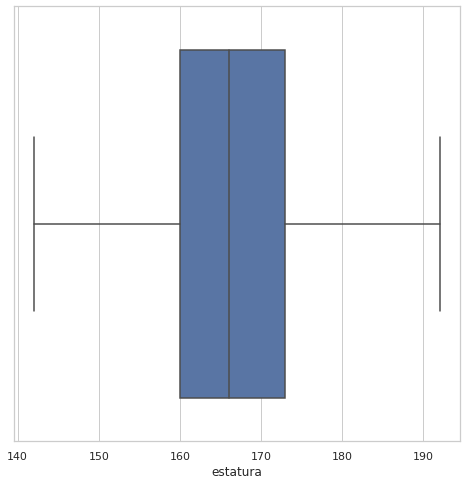


Fig. 17 Diagrama de bigotes sin outliers

Una vez realizado este proceso, se realiza la visualización nuevamente de la estatura y como se verá en la Fig 18., la información se ve mucho más adecuada.

Sumado a lo anterior, también se identificó que para el campo de etnia se debía modificar los registros de 'PALENQUERO DE SAN BASILO' por 'PALENQUERO DE SAN BASILIO' ya que hacían referencia a la misma etnia, solo que el dataset tenía problemas de escritura.

1. *Explosión de datos*

Debido al campo cantidad de personas de nuestro dataset, el cual hace referencia a la cantidad de personas que cumple con dichas coincidencias de características, por ejemplo: Si en existen dos personas que tengan exactamente el mismo país, edad, estatura, etnia, etc. En el dataset estarán representados por un solo registro, sin embargo, en cantidad de personas estará un dos haciendo alusión al número de personas que cumplen con esos datos.

Dado lo anterior, se tuvo la necesidad de explosionar los datos, es decir, se insertaron nuevos registros multiplicando la información por la cantidad de personas de cada registro. Es importante aclarar que, una vez hecho esta parte, el dato cantidad de personas no se necesitó utilizar más en el proceso.

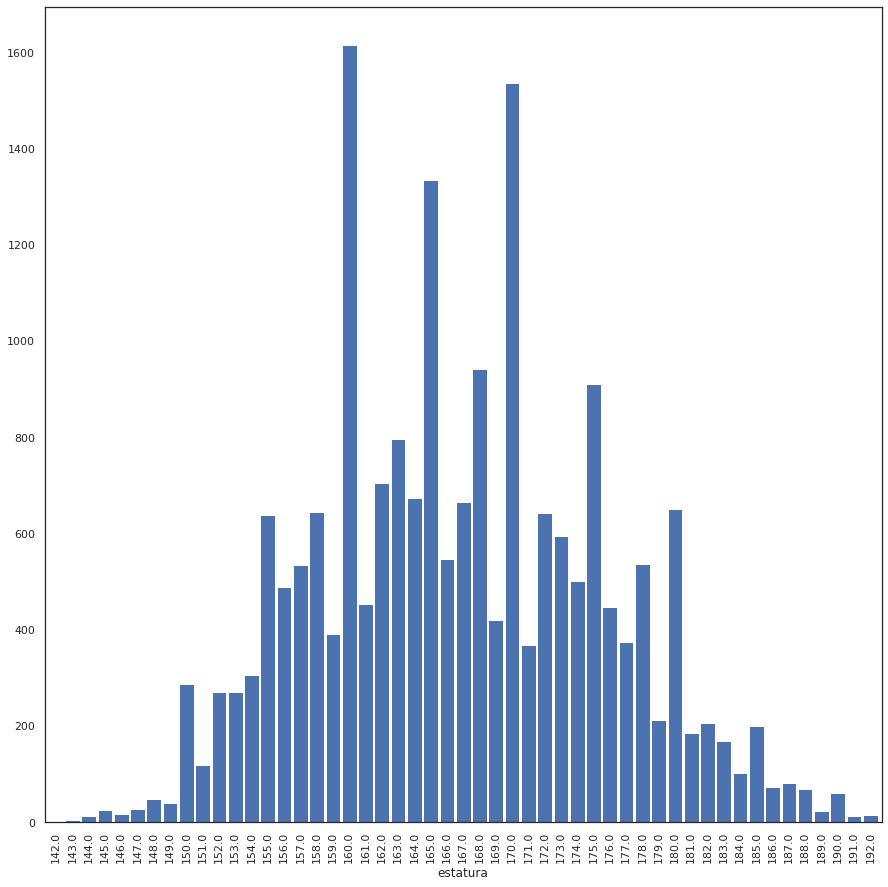


Fig. 18 Número de personas por estatura sin outliers

En este proceso se crearon en primera instancia dos dataframe con la información de los datos cuyo campo cantidad\_personas es diferente de 1 (data\_diferente\_1) y otro con los registros de cantidad\_personas igual a 1 (data\_igual\_1). Se hizo una sumatoria del campo cantidad\_personas en el dataframe data\_diferente\_1 para identificar la cantidad de registros nuevos (1290) que se deberá insertar.

Posteriormente, se recorrió el dataframe data\_diferente\_1 para multiplicar la información, es decir, insertar la cantidad de registros iguales a su información según indique su campo cantidad\_personas, creando un nuevo dataframe llamado data\_tem el cual contendrá estos nuevos registros.

1. *Categorización de datos*

Dentro del procesamiento de los datos, más específicamente para el entrenamiento del dataset, dado que algunos de los algoritmos solo permitian datos numéricos y por lo tanto fue necesario la realización de una categorización de las características cualitativas, para la realización de este proceso convertimos a category cada una de las características cualitativas como género, etnia, estado\_civil, etc.

1. *Balanceo de los datos.*

Se identificó desbalance en los datos, con una gran concentración en los países de España y Estados Unidos, como se ve en la Fig 6. esto hace que afecte el algoritmo, en los procesos de generalización con las clases minoritarias.

Para solucionar este problema, se probaron 3 modelos de balanceo de datos, los cuales “crean” datos sintéticos a partir de los datos que tiene la clase a balancear. Cabe resaltar que se realizaron pruebas con la clase imblearn utilizando variaciones de su clase y diferentes parámetros. Se encontraron los siguientes algoritmos:

* SMOTE: Por sus siglas en inglés “Synthetic Minority Oversampling Technique”, es un algoritmo que balancea las clases minoritarias, para dejarlas con la misma cantidad de datos de la clase mayoritaria. Se le define un parámetro de tipo string (‘monority’,’not minority’, ‘not mayority’, ‘all’), que indica a que clases se le quiere hacer el remuestreo. Se aplica para la clase SMOTE y todas sus variaciones. [19]
* SMOTETomek: Es un tipo de balanceo combinado, es decir, balancea las clases minoritarias y las clases mayoritarias, hasta tener un número parejo de datos. Se agrega un algoritmo de ayuda al SMOTE, que es el Tomek link, que sirve para encontrar un vecino contrario más cercano y el algoritmo decide si eliminar el par o solo uno.[20]
* SMOTE-NC: Es un tipo de balanceo para características, tanto continuas como categóricas, usando igualmente la lógica del SMOTE.[22]

Se escogió el modelo de SMOTETomek, con el parámetro ‘all’ (remuestrea todas las clases) ya que es el que mejor resultado que da, cuando se balancean los datos. Sin embargo, se tuvo que aplicar una eliminación de datos para que los algoritmos pudieran funcionar, en este caso, se eliminaron todos los países que tuvieran 6 o menos datos, esto es porque todos los algoritmos que se probaron, exigen al menos 6 datos para poder hacer relaciones y crear los datos sintéticos.

VI. ENTRENAMIENTO Y EVALUACIÓN

A partir de las soluciones encontradas en los artículos y de los modelos vistos en clase se utilizaron los siguientes modelos de aprendizaje:

1. *XGBoost*

Es un algoritmo de aprendizaje automático de conjunto basado en árboles de decisiones que utiliza un marco de [impulso de gradiente](https://en.wikipedia.org/wiki/Gradient_boosting) . En los problemas de predicción que involucran datos no estructurados (imágenes, texto, etc.), las redes neuronales artificiales tienden a superar a todos los demás algoritmos o marcos. Sin embargo, cuando se trata de datos tabulares / estructurados de tamaño pequeño a mediano, los algoritmos basados ​​en árboles de decisión se consideran los mejores en su clase en este momento. [14]

1. *Redes Neuronales Artificiales*

Las RNA son sistemas de procesamiento de la información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas (Hilera y Martínez, 1995). Cada neurona está conectada con otras neuronas mediante enlaces de comunicación, cada uno de los cuales tiene asociado un peso. Los pesos representan la información que será usada por la red neuronal para resolver un problema determinado. [15]

1. *KNN (K-Nearest Neighbors)*

El vecino más cercano de K es uno de los algoritmos de aprendizaje automático más simples basado en la técnica de aprendizaje supervisado.El algoritmo K-NN asume la similitud entre el nuevo caso / datos y los casos disponibles y coloca el nuevo caso en la categoría que es más similar a las categorías disponibles. [16]

1. *Random Forest*

Como su nombre lo indica, consiste en una gran cantidad de árboles de decisión individuales que operan como un [conjunto](https://en.wikipedia.org/wiki/Ensemble_learning) . Cada árbol individual en el bosque aleatorio lanza una predicción de clase y la clase con más votos se convierte en la predicción de nuestro modelo. [17]

1. *Radial SVM (Support Vector Machines)*

Son una técnica de clasificación famosa y muy sólida que no utiliza ningún tipo de modelo probabilístico como cualquier otro clasificador, sino que simplemente genera hiperplanos o simplemente coloca líneas, para separar y clasificar los datos en algún espacio de características en diferentes regiones. En este caso en particular el Kernel o función matemática utilizada es de base radial. [18]

VII. ANÁLISIS Y EXPLICACIÓN DE LAS MÉTRICAS DE DESEMPEÑO

Para analizar los diferentes modelos se utilizaron las siguientes métricas:

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

Los modelos se aplicaron para los datos sin ningún tipo de procesamiento, luego con el procesamiento sin el balanceo y finalmente con todas las técnicas de preprocesamiento completas definidas en la sección V. Se consignan los resultados para cada caso.

1. *Datos sin procesamiento*

TABLA II.

MÉTRICAS DE DESEMPEÑO PARA LOS MODELOS UTILIZANDO DATOS SIN PROCESAMIENTO

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | ***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 |

Se puede observar que con los datos sin procesamiento los resultados de las métricas son muy pobres para todos los modelos. Los modelos RNA y XGBoost obtienen mejores resultados para el *Test set score*, son los que menor *mean\_squared\_error* presentan y se destacan para el *accuracy\_score***,** el *precision\_score*, el *recall\_score* y el *f1\_score*. Los modelos KNN y Random Forest presentan un mejor resultado para el *Train set score*.

1. *Datos procesados sin balanceo*

TABLA III.

MÉTRICAS DE DESEMPEÑO PARA LOS MODELOS UTILIZANDO DATOS PROCESADO SIN BALANCEO

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | ***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 |

Se puede observar que para los datos sin procesamiento los resultados de las métricas mejoran respecto a los obtenidos en los modelos con datos sin procesamiento. Los modelos RNA y XGBoost obtienen mejores resultados para el *accuracy\_score,*el *precision\_score*, el *recall\_score*, el *f1\_score* y el *Test set score*, mientras que para el *Train set score* los mejores resultados se obtienen para los modelos KNN y Random Forest. SVM y KNN son los que menor *mean\_squared\_error* presentan, aunque la diferencia es mínima comparado con los otros modelos

1. *Datos completamente procesados*

TABLA IV

MÉTRICAS DE DESEMPEÑO PARA LOS MODELOS UTILIZANDO DATOS COMPLETAMENTE PROCESADOS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | ***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 |

Para el caso de los modelos utilizados con los datos completamente procesados los mejores resultados obtenidos para todas las métricas corresponden a KNN.

VIII. CONCLUSIONES

A partir de las diferentes pruebas realizadas se puede identificar la importancia del preprocesamiento de los datos. Los resultados de los diferentes modelos mejoran en el momento en que se empiezan a aplicar las diferentes técnicas de preprocesamiento como la eliminación de datos nulos, modificación de datos anómalos y balanceo de datos.

Para este caso en particular fue fundamental realizar el balanceo de los datos a partir de las técnicas investigadas, lo anterior por cuanto el número de registros para cada una de las clases difería de forma importante. El balanceo seleccionado SMOTETomek corresponde a un método combinado, el cual realiza remuestreo sobre todas las características.

Los mejores resultados se obtuvieron desde el punto de vista de tiempo de ejecución y de las métricas para KNN en el caso de los datos completamente procesados. Los modelos RNA y XGBoost presentaron los mejores resultados para los datos sin procesamiento y esa es precisamente una de las principales ventajas de este último modelo: requieren mucha menos limpieza y pre procesado de los datos en comparación a otros métodos de aprendizaje, además no se ven muy influenciados por outliers.

Una variable importante a la hora de obtener un modelo de machine learning es el tiempo de procesamiento y los rendimientos exigidos para el respectivo hardware. Para los modelos analizados se presentaron los mayores tiempos de entrenamiento para los Support Vector Machines y las Redes neuronales artificiales.

REFERENCIAS

[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].