https://doi.org/10.37815/rte.v34n3.945
Artículos originales

Movilidad humana hacia Ecuador, una visión desde el análisis de datos

Human mobility to Ecuador, a vision from data analysis

Paúl Córdova[1] https://orcid.org/0000-0002-5686-7370,
Andrés Tobar[2] https://orcid.org/0000-0003-4822-0515

[1]*Banco Pichincha*, Quito, Ecuador
paul_cordova94@hotmail.com

[2]*Banco Solidario*, Quito, Ecuador
andrestorres492@gmail.com

Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0.

Enviado: 2022/06/18
Aceptado: 2022/09/16
Publicado: 2022/11/30

Resumen

La movilidad humana hacia Ecuador ha sido una temática en creciente difusión en la coyuntura actual, por las condiciones post pandemia que atraviesa el país. En tal motivo, este artículo examina la información concerniente a los ciudadanos que han recibido su visado por parte del Ministerio de Relaciones Exteriores y Movilidad Humana. Con base en los datos disponibles, el análisis se ha enfocado en los solicitantes que se encuentran en territorio nacional provenientes de Colombia y Venezuela. Se han empleado dos modelos de aprendizaje automático con la finalidad de destacar y validar una posible relación entre las características socioeconómicas de los ciudadanos con respeto a su categoría migratoria; identificando si variables como la edad, género, estado civil y nacionalidad pueden influir en la solicitud de visado respecto al tipo de residencia. Según los hallazgos encontrados, ciudadanos colombianos solteros independiente de su edad optan por una residencia temporal; mientras que, para el caso de los divorciados, por una residencia permanente. Para los ciudadanos venezolanos solteros con una edad menor a 23 años la mayoría posee una residencia permanente; no obstante, aquellos con una edad mayor acceden a una residencia temporal.

Palabras clave: Árbol de decisión, movilidad humana, análisis de datos, regresión logística.

Sumario: Introducción, Materiales y Métodos, Resultados y Discusión.

Como citar: Córdova, P. & Tobar, A. (2022). Movilidad humana hacia Ecuador, una visión desde el análisis de datos. *Revista Tecnológica - Espol, 34(3), 31-45.*
http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/945

Abstract

Human mobility to Ecuador has been a topic of increasing diffusion in the current situation, due to the post-pandemic conditions that the country is going through. For this reason, this article examines the information concerning the citizens who have received a visa from the Ministerio de Relaciones Exteriores y Movilidad Humana. Based on the available data, the analysis has focused on applicants from Colombia and Venezuela. Two machine learning models have been used in order to highlight and validate a possible relationship between the socioeconomic characteristics of citizens with respect to their migratory category; identifying whether variables such as age, gender, marital status, and nationality can influence the visa application with respect to the type of residence. According to the findings, single Colombian citizens, regardless of their age, opt for temporary residency, while divorced Colombians opt for permanent residency. For single Venezuelan citizens under 23 years of age, the majority have permanent residency; however, those older than 23 years of age opt for temporary residency.

Keywords: Decision tree, human mobility, data analysis, logistic regression.

Introducción

La movilidad humana es una temática recurrente en las agendas de estado de los gobiernos a lo largo del mundo por concepto de fenómenos biológicos, políticos y sociales. Una necesidad imperante de los países es poder trabajar en conjunto para garantizar una movilidad digna para los ciudadanos y crear ciudades inclusivas que pueda acoger a los migrantes del exterior.

En la actualidad, la Covid-19 ha demostrado la fragilidad del sistema migratorio dado que las restricciones fronterizas y la reducción del acceso mediante la imposición normativa, han generado afectaciones a los derechos de los ciudadanos inmersos en los flujos migratorios.

En una investigación realizada por Nueva Sociedad (Liberona Concha, 2020), se identificaron que las políticas migratorias en los países de América Latina se han encaminado a la restricción de la movilidad humana, provocando desregularización y mayores afectaciones para los migrantes. Estos aspectos han ocasionado la vulneración de derechos y afectaciones sociales para aquellos que huyen de la violencia, estados fallidos o crisis económicas.

En Ecuador se han normado algunos requisitos para la obtención de la visa de residencia temporal, entre los principales se encuentran: la documentación oficial, pasaporte válido y vigente, certificado de antecedentes penales del país de origen, no ser considerado una amenaza para la seguridad interna, acreditar los medios de vida lícitos para la subsistencia, pago de la tarifa y presentación de la solicitud.

Históricamente en Ecuador ha existido una tendencia a que ciudadanos del país vecino, Colombia, ingresen con intención de establecer sus vidas. A partir de una instigación nacional se afirma que, basado en los datos del censo de 2010, la mayoría de los inmigrantes dentro del país son colombianos, comportamiento que se ha conservado con el tiempo (Loor Valeriano, 2012).

Por otra parte, las circunstancias que ha atravesado la República Bolivariana de Venezuela han generado que su población migre en busca de asilo y con intenciones de mejorar su calidad de vida (Gandini, 2019). No obstante, según datos obtenidos por parte de Colectivo Geografía Crítica de Ecuador respecto de la situación de inmigrantes en pandemia, se observa

que el 86,7% de los migrantes venezolanos no cuentan con afiliación a la seguridad social y el 89,9% no tiene seguro de salud privado (Proyecto Inmovilidad en las Américas, s.f.).

Durante esta investigación se busca describir las características de los ciudadanos que han recibido un visado para su permanencia en Ecuador, profundizando por medio del análisis de las condiciones socioeconómicas de los migrantes y su categoría migratoria. A partir de los datos provistos por el Registro de Movilidad Humana se pretende aplicar un enfoque de aprendizaje supervisado (Yaser S. et al 2012), empleando a la categoría migratoria como variable dependiente, por medio del uso del algoritmo de árbol de decisión (J. R. Quinlan, 1996).

La correcta comprensión de los posibles patrones en las solicitudes de visado permitirá conceptualizar formas para una adecuada inserción de los flujos migratorios provenientes principalmente de Colombia y Venezuela. Logrando, de esta manera, generar ideas para promover una inclusión social y económica, prevenir y contrarrestar la discriminación, y gestionar una política de gobernanza inclusiva.

Materiales y Métodos

La sección metodológica de este estudio se estructuró en las siguientes secciones de acuerdo con el flujo de un proyecto de ciencia de datos. Se comenzó por el análisis exploratorio para identificar las características de las variables dentro del conjunto de conjunto de datos. Posterior se realizó la preparación de los datos seleccionados mediante proceso de ingeniería de variable con la finalidad de generar el conjunto de datos de entrada que sería empleado en el modelamiento. Finalmente, se ejecutó un algoritmo supervisado para identificar las reglas de decisión que influenciaron el entrenamiento del modelo estadístico.

Se optó por empelar el lenguaje de programación Python por su versatilidad para ejecutar procesos de extracción, transformación y carga de datos. Igualmente, por su amplia disponibilidad de librería para desarrollar ciencia de datos e implementar modelos de aprendizaje automático.

Conjunto de datos

Del portal de datos abiertos de las instituciones públicas de Ecuador, se recuperó el registro de movilidad humana (visado) provisto por el Ministerio de Relaciones Exteriores y Movilidad Humana. Este archivo contiene información de las visas que han sido otorgadas a los ciudadanos de diferentes nacionalidades en el ámbito nacional y en el exterior desde el mes de enero de 2021 hasta marzo de 2022. A continuación, se muestra el diccionario de datos con las descripciones de los campos que componen la información de la movilidad humana:

Tabla 1

Diccionario de datos

NOMBRE VARIABLE	DESCRIPCIÓN DE LA VARIABLE
Fecha Trámite	Contiene la información de la fecha de inicio del servicio requerido por el ciudadano.
Ubicación	Especifica la localidad donde el ciudadano solicitó el servicio de Visas, el cual es brindado por el MREMH a nivel nacional (Ecuador) y en el exterior.
Categoría Migratoria	Especifica los diferentes tipos de visas de permanencia temporal o permanente que el Estado otorga a los extranjeros en Ecuador de conformidad al hecho que motiva su presencia en el país.
Género	Este término se utiliza para anonimizar los datos de los ciudadanos que solicitan el servicio de movilidad humana, siendo masculino y femenino. En caso de que no se haya registrado el género del ciudadano, se colocará "No se especifica en el sistema".

NOMBRE VARIABLE	DESCRIPCIÓN DE LA VARIABLE
Fecha Nacimiento	Contiene la información de la fecha de nacimiento del ciudadano que requiere el servicio de visas.
Grupo Etario	Especifica la identificación de los distintos grupos humanos divididos por edad: 1. Niños y niñas (0 - 11 años) | 2. Adolescentes (12 - 17 años) | 3. Jóvenes (18 - 30 años) 4. Adultos (31 - 64 años) | 5. Adultos mayores (65 y más años).
Edad	Define la edad del ciudadano solicitante del servicio de visas.
Nacionalidad	Define la nacionalidad del ciudadano solicitante del servicio de visas.
País Nacionalidad	Define el país del ciudadano solicitante del servicio de visas.
Ciudad Nacimiento	Define la ciudad de nacimiento del ciudadano solicitante del servicio de visas.
Estado Civil	Define el estado civil del ciudadano solicitante del servicio de visas: Casado | Divorciado | Soltero | Unión de Hecho | Viudo En caso de que no se haya registrado el estado civil del ciudadano, se colocará "No se especifica en el sistema".
Tipo Acto Consular	Define el tipo de acto consular que requiere el ciudadano solicitante.
Tipo de Grupo de Visa	Define el tipo de visa que requiere el ciudadano solicitante: 1 VTSPI - Visitante Temporal (Solicitante de protección internacional) 2 VT - Visitante temporal 3 RTPI - Residente Temporal (Protección internacional) 4 RT - Residente temporal 5 RP - Residente permanente
Detalle de Visa	Define el detalle de la visa que requiere el ciudadano solicitante.
Discapacidad	Define si el ciudadano solicitante es discapacitado (1) y si no es discapacitado (0).

Análisis exploratorio de datos

Se recomienda realizar una fase preliminar antes de comenzar el modelo estadístico, para familiarizarse con los datos a evaluar. A este tipo de aproximación se le conoce como análisis exploratorio de datos, y se realiza sin ninguna hipótesis previa, utilizando enfoques estadísticos y representaciones gráficas. En esta fase comienzan a surgir los vínculos más evidentes entre las variables que finalmente se investigarán con el rigor adecuado (Rojo, 2006).

Para este proceso se trabajaron con las principales librerías enfocas en Ciencia de Datos por medio del lenguaje Python. Empleando pandas para la lectura y manipulación de datos, pandas-profiling para examinar y perfilar la información de forma global. Y la librería dataprep para la creación de visualizaciones y relacionamiento de variables.

Los datos crudos provenientes de la fuente contaban con 15 variables y 65.502 registros. Con respeto al tipo de datos de cada columna se identificaron: fecha (2 variables), categóricas (12 variables) y numéricas (1 variable).

Entre las principales estadísticas descriptivas respecto a la movilidad humana se destacan las siguientes:

- El 84% de las personas que realizaron la solicitud de visado, se encuentran ubicados en Ecuador.
- Las dos principales categorías migratorias son: residente temporal (56%) y residente permanente (23%).
- Existe una proporción cercana entre hombres y mujeres.
- La edad presenta una distribución sesgada a la derecha, con una mediana de 34 años.
- El top cinco de nacionalidades fueron: colombiana, venezolana, estadounidense, china y peruana.

- Cerca del 80% de las personas que realizaron la solicitud de visado, tiene estado civil soltero.

Posterior al examinar las variables de interés se ejecutaron los siguientes procesos de filtrado y modificación del conjunto de datos original:

1. Seleccionar de los solicitantes que se encuentran en Ecuador, excluyendo aquellos ciudadanos que se encuentran en el exterior.
2. Mantener las categorías migratorias temporal y permanente, por lo cual no se consideraron categorías como: visitante temporal y diplomático.
3. Preservar si el país de nacionalidad corresponde a Venezuela o Colombia, porque representan entre ambas la mayor parte de ciudadanos que han recibido visado de residencia.
4. Eliminar las filas duplicadas, teniendo en consideración si los registros presentaban los mismos valores en todas las filas.
5. Seleccionar las columnas de interés para el modelo: género, estado civil, edad en años, nacionalidad y categoría migratoria.

Preparación de datos

A partir de los hallazgos del análisis exploratorio de datos, se procede con la partición de los datos, empleando la información referente al 2021 y 2022 para modelamiento y validación respectivamente. La cantidad de registros que compone cada conjunto de datos se muestra en la Tabla 2.

Tabla 2

Descripción de la partición de datos

CONJUNTO DE DATOS	CANTIDAD	RESIDENCIA TEMPORAL	RESIDENCIA PERMANENTE
Modelamiento	27068	19324	7744
Validación	4713	2955	1758

A partir de los datos de modelamiento se aplicó la técnica “One Hot Encoding” (M. K. Dahouda & I. Joe, 2021), que consiste en la creación de una variable tipo dummy por cada una de las categorías que posee la variable de origen. Adicionalmente para variables binarias únicamente se consideró una de las dos particiones que se generan al momento de ejecutar dicha técnica, esto es para el caso de las variables: género, país de nacionalidad y categoría migratoria. La Tabla 3 contiene el nombre de las variables creadas a partir de la técnica mencionada.

Tabla 3

Variables One Hot Encoding

NO.	VARIABLE CODIFICADA
1	Género femenino
2	Nacionalidad colombiana
3	Estado civil casado
4	Estado civil divorciado
5	Estado civil no definido
6	Estado civil soltero

NO.	VARIABLE CODIFICADA
7	Estado civil unión de hecho
8	Estado civil viudo
9	Residente temporal

La preparación de los datos o ingeniería de variables es un paso esencial antes de la ejecución de los modelos de aprendizaje automático, con la finalidad de garantizar la correcta lectura de los datos por los algoritmos.

Modelos de aprendizaje supervisado

Es trascendental comprender que: "el aprendizaje automático (ML) se refiere a la capacidad de un sistema para adquirir e integrar conocimiento a través de observaciones a gran escala, y para mejorar y expandirse mediante el aprendizaje de nuevos conocimientos en lugar de ser programado con ese conocimiento". (Park & Woolf, 2009). En este sentido, busca observar y aprender los patrones para luego replicar estos resultados sobre un nuevo conjunto de observaciones con la finalidad de seguir perfeccionándose.

En el ámbito del aprendizaje automático existen tres tipos de aprendizaje supervisado, no supervisado y de reforzamiento (Yaser S. et al, 2012). Enfocando el análisis en el primero, desde el punto de vista supervisado se conoce toda la información referente a la clasificación de las observaciones por lo que se pretende que el algoritmo a emplearse analice los patrones dentro de los datos, con la finalidad de que se adapte a cualquier tipo de procedencia de las observaciones. Se divide a los datos en un conjunto de entrenamiento y otro de testeo empleando al primer conjunto para aprender el comportamiento de las observaciones y el segundo para evaluar el desempeño del modelo o regla que se ha obtenido dentro del primer conjunto. Considerando este enfoque se realiza la división en entrenamiento y testeo particionando los datos de modelamiento en el 70% y 30% respectivamente.

"Un árbol de decisión es una estructura en forma de árbol con ramas que representan grupos de decisiones. Estas decisiones conducen a un conjunto de reglas para categorizar una colección de datos en subgrupos disjuntos y exhaustivos. La ramificación recursiva se realiza hasta que se cumplen los requisitos de parada especificados" (Goicoechea, 2002). La construcción visual se representa en una imagen que se lee de abajo hacia arriba, destacando las reglas de clasificación en cada instancia para la separación de las observaciones en subgrupos.

En aprendizaje automático un árbol de decisión (Quinlan, 1996), es una estructura de fácil interpretación debido a que su comportamiento que es similar a un diagrama de flujo, donde cada una de sus ramas representa una decisión y cada hoja un atributo, una condición. Se encuentra compuesto por:

- **Nodo Raíz:** nodo superior del árbol. El nodo raíz se considera como la decisión que guiará a las ramificaciones.
- **Ramificaciones:** caminos que unen los nodos y muestran la acción que se va a tomar.
- **Nodo de decisión:** Muestra la decisión que se va a tomar.
- **Nodos terminales o hojas:** indica el resultado definitivo.

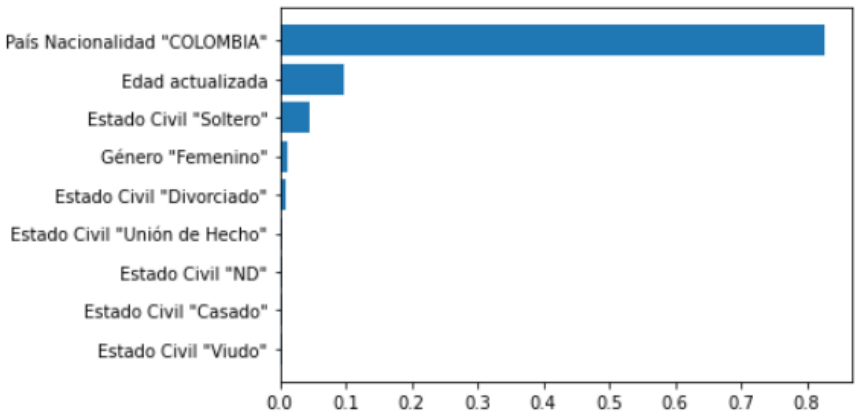
Dado que la variable dependiente es de tipo categórica se utilizará un árbol de clasificación, empleando el Índice Gini para la creación de ramificaciones para las posibles

divisiones. Es importante destacar que este algoritmo se encuentra implementado en la librería skicit-learn (Pedregosa et al, 2011); especializada en Machine Learning en el lenguaje de programación Python.

Un aspecto de interés en el análisis de los resultados corresponde al evaluar la importancia de cada uno de los predictores que conforman el árbol de decisión respecto a la variable dependiente (M. R. A. Iqbal et al, 2012). La Figura 1, destacó que el poder predictivo de la nacionalidad fue determinante al discriminar la categoría migratoria. Por lo cual, su relevancia es trascendental en lo que respecta al otorgamiento del visado de residencia permanente o temporal. Posterior en importancia se encontraron las variables referentes a la edad de los ciudadanos, si el estado civil es soltero/divorciado, y el género femenino.

Figura 1

Top 5 importancia de variables



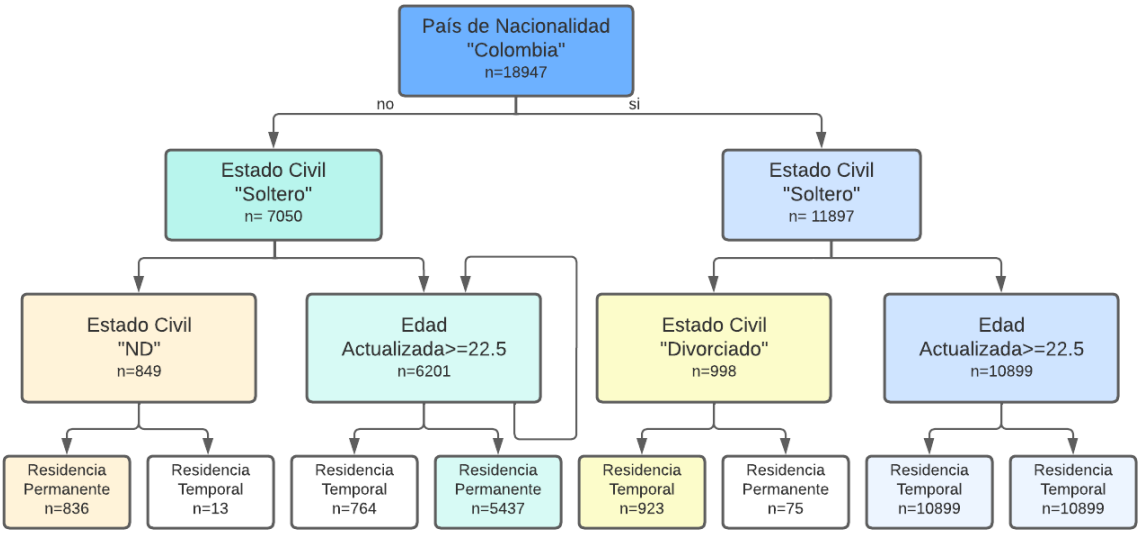
A continuación, en la Figura 2 se muestra las reglas de decisión optimizadas mediante el entrenamiento del árbol.

Mediante un análisis detallado de los resultados determinado por el árbol de clasificación mostrados en la Figura 2, se observan las siguientes características en cuanto a la categoría migratoria:

- Residente temporal:
  - Nacionalidad colombiana y estado civil soltero.
  - Nacionalidad colombiana y su estado civil no es soltero ni divorciado.
  - Nacionalidad venezolana, estado civil soltero y edad menor a 22 años.
  - Nacionalidad venezolana, su estado civil no es soltero, pero no está determinado en el sistema.
- Residente permanente:
  - Nacionalidad colombiana, su estado civil no es soltero, pero si consta como divorciado.
  - Nacionalidad venezolana, estado civil soltero y edad mayor a 22 años.
  - Nacionalidad venezolana y su estado civil es distinto de soltero y no determinado en el sistema.

Figura 2

Árbol de decisión



Nota: Para una lectura de las reglas de decisión mediante la visualización del árbol de decisión, se analiza al interior de cada nodo la condición evaluada. A la derecha se encuentra la condición verdadera y a la izquierda la condición falsa.

Para contrarrestar los resultados obtenidos a partir del árbol de decisión, se emplea un modelo de regresión logística (Salcedo & Poma, 2002) utilizado para evaluar el efecto de las variables consideradas en el estudio sobre la probabilidad de solicitar una visa de tipo permanente o temporal.

Figura 3

Resumen Regresión logística (Todas las variables)

Dep. Variable:	CategoriaMigratoria_Residente Temporal			**No. Observations:**		18947
Model:	Logit			**Df Residuals:**		18938
Method:	MLE			**Df Model:**		8
Date:	Aug 2022			**Pseudo R-squ.:**		inf
Time:				**Log-Likelihood:**		-4.7037e+05
converged:	True			**LL-Null:**		0.0000
	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Edad_actualizada	0.0009	0.001	0.651	0.515	-0.002	0.004
Género_Femenino	0.0751	0.038	1.956	0.051	-0.000	0.150
PaisNacionalidad_COLOMBIA	2.5734	0.040	64.689	0.000	2.495	2.651
EstadoCivil_Casado	-1.3928	0.094	-14.871	0.000	-1.576	-1.209
EstadoCivil_Divorciado	-2.5083	0.235	-10.685	0.000	-2.968	-2.048
EstadoCivil_ND	0.3825	0.505	0.758	0.449	-0.607	1.371
EstadoCivil_Soltero	-0.3324	0.058	-5.722	0.000	-0.446	-0.219
EstadoCivil_UniónDeHecho	-1.7749	0.238	-7.472	0.000	-2.241	-1.309
EstadoCivil_Viudo	-1.2469	0.292	-4.272	0.000	-1.819	-0.675

La Figura 3 contiene el resumen estadístico obtenido para el modelo de regresión logística, al considerar los estadísticos asociados a la significancia de cada una de las variables (P>|z|). Se tiene que el valor asociado a la edad actualizada, género femenino y estado civil ND es mayor a 0.05 (Shaffer, 1995), por lo que no son decisivos al momento de analizar la categoría migratoria del solicitante. A continuación, se muestra el resumen del modelo al omitir las variables nombradas anteriormente.

Enfocando el análisis en el valor de los coeficientes mostrados en la Figura 4 (coef), se poseen las siguientes características respecto al efecto de las variables seleccionadas para explicar la categoría migratoria:

- El valor positivo en el coeficiente asociado al País de Nacionalidad Colombia, indica que las personas de esta nacionalidad poseen una mayor propensión a solicitar un visado temporal, en comparación a aquellas de nacionalidad venezolana.
- Los valores negativos asociados a los coeficientes de los distintos tipos de Estado Civil indican que la propensión a solicitar un visado temporal de las personas que conforman estos grupos es menor respecto a aquellas que poseen un estado civil no definido. Adicionalmente, se puede observar que aquellos perfiles que pueden ser asociados con poseer un núcleo familiar tienen una probabilidad menor de solicitar un visado temporal frente a la posibilidad de uno permanente (el valor de los coeficientes tiende a ser menor).

Figura 4

Resumen Regresión logística (Variables significantes)

Dep. Variable:	CategoriaMigratoria_Residente Temporal			**No. Observations:**		18947
Model:	Logit			**Df Residuals:**		18941
Method:	MLE			**Df Model:**		5
Date:	Aug 2022			**Pseudo R-squ.:**		inf
Time:				**Log-Likelihood:**		-4.6967e+05
converged:	True			**LL-Null:**		0.0000
	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
PaisNacionalidad_COLOMBIA	2.5730	0.040	65.009	0.000	2.495	2.651
EstadoCivil_Casado	-1.3077	0.066	-19.885	0.000	-1.437	-1.179
EstadoCivil_Divorciado	-2.4110	0.222	-10.880	0.000	-2.845	-1.977
EstadoCivil_Soltero	-0.2617	0.025	-10.447	0.000	-0.311	-0.213
EstadoCivil_UniónDeHecho	-1.6925	0.230	-7.367	0.000	-2.143	-1.242
EstadoCivil_Viudo	-1.1230	0.275	-4.076	0.000	-1.663	-0.583

Finalmente, se emplearon las siguientes métricas calculadas a partir de la matriz de confusión que sirven para evaluar tanto el rendimiento del modelo y la precisión de sus clasificaciones (Fernández Casal & Costa, 2021).

Tabla 4

Matriz de confusión

	1	**0**
1	Verdaderos Positivos (TP)	Falsos Negativos (FN)
0	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Las métricas utilizadas para medir el desempeño del modelo se obtienen a partir de los valores mostrados en la Tabla 4. Estos son:

- Exactitud: Porcentaje de datos clasificados correctamente.

$$Exactitud = \frac{TP + TN}{TP + FN + TN + FP}$$

- Tasa de error: Porcentaje de datos clasificados incorrectamente.

$$Tasa\ de\ error = \frac{FP + FN}{TP + FN + TN + FP}$$

- Tasa de verdaderos positivos (sensibilidad)]: El porcentaje de datos clasificados correctamente cuando pertenecen al grupo 1.

$$sensibilidad = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- Especificidad: El porcentaje de datos clasificados correctamente cuando pertenecen al grupo 0.

$$especifidad = \frac{TN}{TN + FP}$$

- Precisión: Mide los datos es clasificado como 1s con qué frecuencia es etiquetado correctamente.

$$precisión = \frac{TP}{TP + FP}$$

- Curva AUC-ROC: El área bajo la curva ROC sirve para medir el rendimiento de los problemas de clasificación binaria al variar el umbral (Cifuentes, 2012). La ROC es una curva de probabilidad que permite representar gráficamente a la sensibilidad frente a la razón de falsos positivos (1-especificidad), y el AUC representa el grado de separabilidad, indicando en qué medida el modelo es capaz de diferenciar entre grupos. Cuanto más alto sea el AUC, mejor será el modelo para distinguir entre clases.

Se obtuvieron los siguientes resultados en cuanto al modelamiento (Tabla 5):

Tabla 5

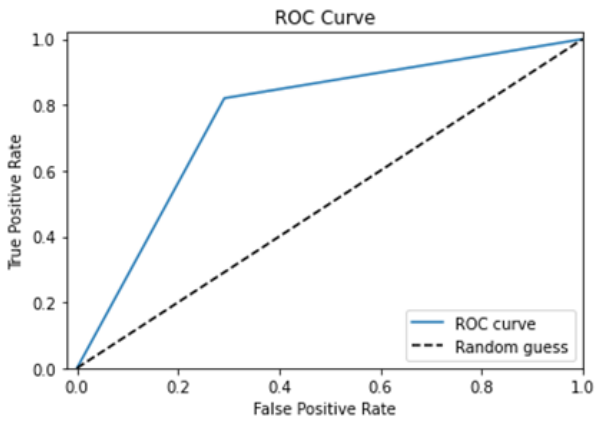
Matriz de confusión modelamiento Árbol decisión

	1	**0**
1	4733	1037
0	685	1666

A partir de la Tabla 5 se procede a calcular las métricas de evaluación en el modelamiento. Los valores obtenidos son: tasa de error (21.20%), exactitud (78.80%), sensibilidad (82.03%), especificidad (70.86%) y precisión (87.36%) indican un desempeño adecuado al momento de clasificar entre los grupos de interés.

Figura 5

Curva AUC-ROC modelamiento Árbol decisión



El comportamiento ideal de la curva se da cuando la línea azul se encuentra cercana al eje superior izquierdo, puesto que indica que el modelo es efectivo al momento de juzgar entre los grupos de clasificación. La Figura 5 muestra el resultado sobre este conjunto de datos obteniendo un AUC de 76.45%, siendo una métrica adecuada por su cercanía al 80%.

Tabla 6

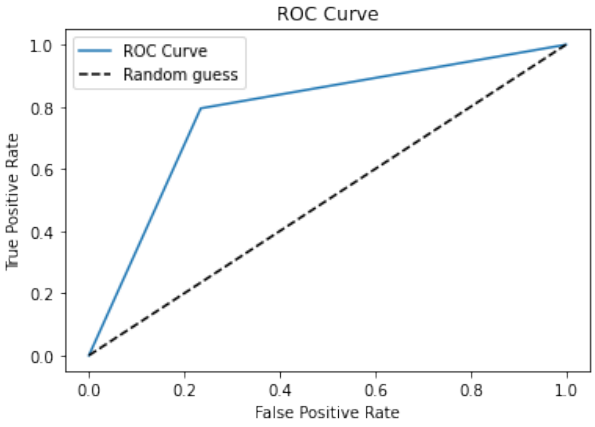
Matriz de confusión modelamiento Regresión

	1	**0**
1	4588	1182
0	551	1800

Al igual que en el caso del Árbol de decisión se calculan las métricas asociadas a su desempeño a partir de los valores de la Tabla 6: tasa de error (21.34%), exactitud (78.66%), sensibilidad (79.51%), especificidad (76.56%) y precisión (89.28%) obteniendo valores bastante similares al caso anterior con excepción de la sensibilidad y especificidad.

Figura 6

Curva AUC-ROC modelamiento Regresión



Al igual que en las métricas de evaluación, la Figura 6 muestra un comportamiento parecido en cuanto a la forma de juzgar entre grupos obteniendo un AUC del 78.04%, por lo que en el modelamiento se tiene que la Regresión logística posee mejor desempeño que el Árbol de decisión.

Se calcularon las métricas para la evaluación del desempeño en cuanto a las clasificaciones realizadas con los datos de validación con la finalidad de mostrar que no existe sobreajuste en el modelo.

Tabla 7

Matriz de confusión validación Árbol de decisión

	1	**0**
1	2780	175
0	481	1277

Los valores mostrados en la Tabla 7 se asocian a una tasa de error (13.92%), exactitud (86.08%), sensibilidad (94.08%), especificidad (72.64%) y precisión (85.25%) que coligen que el modelo basado en un árbol de decisión conserva un buen desempeño al momento de clasificar entre los grupos de interés. Mientras que la Figura 7 se muestra que el comportamiento de la curva ROC es bastante cercano al ideal (cercano al eje), lo que se traduce en un AUC del 83,36%; reforzando las conclusiones de un correcto desempeño en las clasificaciones obtenidas por el modelo.

Figura 7

Curva AUC-ROC validación Árbol de decisión

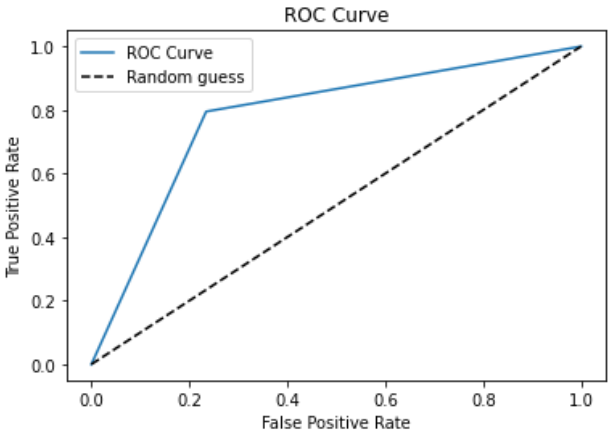


Tabla 8

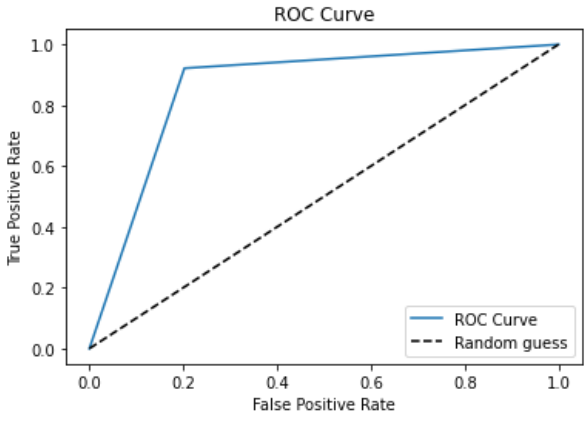
Matriz de confusión validación Regresión

	1	**0**
1	2724	231
0	356	1402

En el caso de la regresión logística, los valores de las métricas asociadas a la Tabla 8 son: tasa de error (12.45%), exactitud (87.55%), sensibilidad (92.18%), especificidad (79.75%) y precisión (88.44%). Al igual que en el caso del Árbol el modelo de Regresión conserva un desempeño bastante similar al del modelamiento, lo que se traduce en un AUC del 85.97% como se muestra en la Figura 8.

Figura 8

Curva AUC-ROC validación Regresión



Resultados y Discusión

Se observa que la mayoría de los ciudadanos colombianos realizan el trámite de visado para una residencia temporal que representa el primer paso dentro del proceso de solicitud. Sin embargo, toman la decisión de mantener esta categoría migratoria dado que posiblemente no buscan radicarse en el país. Para los ciudadanos provenientes de Venezuela, se observa una búsqueda de completar el proceso para solicitar una residencia permanente, con la finalidad de establecerse dentro de Ecuador.

Al momento de comparar los modelos se tiene que el desempeño de la regresión logística es ligeramente mejor que del Árbol de decisión; no obstante, es recomendable fijar el objeto del análisis para poder seleccionar en método a emplearse. Siendo así que si el objetivo fuese analizar aquellas personas que solicitan un visado de tipo permanente el modelo de regresión posee una mejor especificidad (Tabla 8); y, por otro lado, si se pretende estudiar las características de aquellas personas que solicitan un visado temporal el modelo de árbol posee una mayor sensibilidad (Tabla 7).

Sin embargo, dado que el estudio se enfoca en la plausibilidad de emplear un modelo de aprendizaje para la identificación de las características predominantes al momento de solicitar una visa de tipo permanente o temporal, se debe tener en cuenta el efecto de las variables independientes en cada uno de los modelos sobre la variable dependiente (Figura 4 y Figura 2). Siendo así, que las conclusiones halladas en cada modelo son bastante similares, contrastando la validez del método.

Finalmente, se considera al modelo del Árbol de decisión para poder explicar los hallazgos encontrados dentro del conjunto de datos, por la practicidad que este presenta al momento de realizar las clasificaciones. Se concluye que las características predominantes al momento de otorgar una visa de residencia temporal a una persona de nacionalidad colombiana es que sean solteras, que se encuentren con una pareja o una familia establecida. Mientras que para el caso de las personas con nacionalidad venezolana se encuentra principalmente predominada por una población relativamente joven (edad menor a 23 años) y soltera.

En el caso de las visas para residencia permanente en su mayoría se encuentran concentradas en personas de nacionalidad venezolana solteras y con una edad mayor a 23 años, asociando a estas características a personas que podrían pertenecer a la fuerza de trabajo dentro del país. Seguidas por personas de la misma nacionalidad que poseen una pareja o familia establecida, y en menor cantidad por personas de nacionalidad colombiana cuyo estatus civil consta como divorciado.

El conjunto de datos fue separado en subconjunto de modelamiento y otro de validación para realizar un contraste mediante métricas de evaluación. Se determinó en datos que no fueron entrenados con el modelo se obtuvo una exactitud del 86%; lo que representa, que el modelo de clasificación tiene un error únicamente del 14% con respecto al total de casos analizados para el conjunto de datos de validación.

Es importante aclarar la forma en que el modelo de aprendizaje automático empleado para el presente estudio se acopla al análisis presentado, dado que el funcionamiento de un Árbol de decisión permite analizar las características conjuntas de las personas que se les ha otorgado una visa de tipo permanente o temporal y cuya nacionalidad corresponde a las de interés.

Cabe recalcar que se ejecutó individualmente un modelo de Árbol para Colombia y otro para Venezuela. No obstante, los resultados no determinaron una correcta separación de la variable objetivo lo cual conlleva a identificar que la nacionalidad ejerció un aporte fundamental con regla de decisión primaria, validando lo obtenido al analizar la importancia de las variables.

Finalmente, se recomienda que al momento de considerar trabajar con datos abiertos el investigador debe estar consiente que deberá atenerse a la información brindada por la respectiva fuente y que en varios casos el tratar de acceder a datos adicionales puede conllevar procesos burocráticos que no aseguran que se pueda conseguir la información solicitada.

Referencias

Beverly Park Woolf, Editor(s): Beverly Park Woolf, Building Intelligent Interactive Tutors, Morgan Kaufmann, 2009, ISBN 978-0-12-373594-2, https://doi.org/10.1016/B978-0-12-373594-2.X0001-9

Gandini, Luciana & Prieto Rosas, Victoria & Lozano-Ascencio, Fernando. (2019). El éxodo venezolano: migración en contexto de crisis y respuestas de los países latinoamericanos.

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación. Eustat. Available in: http://www. eustat. es/documentos/datos/ct, 4.

Jaime Cerda y Lorena Cifuentes. Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. Revista chilena de infectología, 29:138 – 141, 04 2012.

J. R. Quinlan. 1996. Learning decision tree classifiers. ACM Comput. Surv. 28, 1 (March 1996), 71–72. https://doi.org/10.1145/234313.234346

Liberona Concha, N. (2020). Fronteras y movilidad humana en América Latina. Nueva sociedad, (289), 49-58.

Loor Valeriano, Katherine (2012). Estadísticas y distribución espacial de la migración en el Ecuador según censo 2010. Dspace espol. http://www.dspace.espol.edu.ec/handle/123456789/24824

M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," in IEEE Access, vol. 9, pp. 114381-114391, 2021, doi: 10.1109/ACCESS.2021.3104357

M. R. A. Iqbal, S. Rahman, S. I. Nabil and I. U. A. Chowdhury, "Knowledge based decision tree construction with feature importance domain knowledge," 2012 7th International Conference on Electrical and Computer Engineering, 2012, pp. 659-662, doi: 10.1109/ICECE.2012.6471636.

Pedregosa, Fabian & Varoquaux, Gael & Gramfort, Alexandre & Michel, Vincent & Thirion, Bertrand & Grisel, Olivier & Blondel, Mathieu & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Edouard & Louppe, Gilles. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12.

Proyecto (In)Movilidad En Las Américas, Ficha de Ecuador. Covid-19 e (In)movilidad en las Américas, 2020, enlace: https://www.inmovilidadamericas.org/ecuador.

Salcedo Poma, Celia Mercedes, Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación ,Informe Profesional (Lic.) Universidad Nacional Mayor de San Marcos. Facultad de Ciencias Matemáticas. EAP. de Estadística, 2002.

J P Shaffer, Multiple Hypothesis Testing, Journal Article, 1995, Annual Review of Psychology, 561-584, https://www.annualreviews.org/doi/abs/10.1146/annurev.ps.46.020195.003021

Rojo, J. M. (2006). Análisis descriptivo y exploratorio de datos. Laboratorio de Estadística del Instituto de Economía y Geografía Consejo Superior de Investigaciones Científicas, Madrid.

Rubén Fernández Casal y Julián Costa. Aprendizaje estadístico. GitHub, 2020.

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. Learning From Data. AMLBook.