Análisis de Datos y desarrollo de un sistema de IA para la detección de la Tasa de Aprobación de Visas a Estados Unidos.

Data mining and development of software with artificial intelligence focused on detecting visa approval profiles for the United States

Leonardo Andrés Novoa Peralta[1], Víctor Hugo Medina[2]

[1]MSc., Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, grupo de investigación internacional de informática comunicación y gestión del conocimiento GICOGE, Bogotá, Colombia.
[2]Ph.D., Universidad Pontificia de Salamanca, grupo de investigación internacional de informática comunicación y gestión del conocimiento GICOGE, Bogotá, Colombia.

Cite this article as:L Novoa Peralta, V. Medina "Análisis de Datos y desarrollo de un sistema de IA para la detección de la Tasa de Aprobación de Visas a Estados Unidos.", Prospectiva, Vol 23, N° 2, 2025.

Recibido: 02/11/2024 / Aceptado: 16/12/2024

http://doi.org/ 10.15665/rp.v23i2.3693

RESUMEN

Este artículo presenta un análisis de los perfiles de personas a quienes les aprueban o niegan la visa de turista en Colombia usando minería de datos a través de la metodología CRISP-DM e inteligencia computacional, teniendo en cuenta 2 factores cruciales. 1. El índice de negaciones es aproximadamente un 46% lo que significa que los colombianos gastan casi dos millones de dólares en visas rechazadas, 2. Las citas para la entrevista de visa se encuentran retrasadas por dos años después de la emergencia del COVID 19. Apoyado también en otras investigaciones como las realizadas por Prateek y Karun quienes utilizaron algoritmos de clasificación para detectar perfiles y predecir los resultados de visas en su estudio, nuestro análisis obtiene información sobre los patrones y características comunes entre los solicitantes que han obtenido la aprobación de su visa de turismo, tales como su edad, género, nacionalidad, estado civil, profesión, entre otros aspectos relacionados en los formularios de aplicación (DS-160). El análisis culmina con el desarrollo de una calculadora IA, capaz de predecir la probabilidad de aprobación con una efectividad de más de un 85%. Ideal para solicitantes quienes podrían ver que puntos son relevantes para mejorar o simplemente no presentarse y esperar el momento idóneo.

Palabras-clave: Visa turismo; minería de datos; inteligencia artificial; migración a Estados Unidos.

ABSTRACT

This article presents an analysis of the profiles of individuals who are approved or denied tourist visas in Colombia using data mining through the CRISP-DM methodology and computational intelligence, taking into account 2 crucial factors. 1. The denial rate is approximately 46%, which means that Colombians spend nearly two million dollars on rejected visas. 2. Visa interview appointments are delayed by two years after the emergence of COVID-19. supported too by other investigations like Prateek and Karun who used classification algorithms to detect profiles and predict results of study visas , our analysis obtains information about the patterns and common characteristics among applicants who have obtained approval for their tourist visa, such as their age, gender, nationality, marital status, profession, among other aspects related in the application forms (DS-160). The analysis concludes with the development of an AI calculator capable of predicting the approval probability with an effectiveness of over 85%. Ideal for applicants who could see which points are relevant to improve or simply not show up and wait for the right moment.

***Keywords**:* Tourism visa; data mining; artificial intelligence; migration to the United States.

1. Introducción

Cada año, los colombianos gastan al menos 1,7 millones de dólares en sus intentos de obtener una visa de turista con un porcentaje de rechazo de un 46% según el Informe de la Oficina de Visas **[1]**. Así mismo Colombia es el segundo país en América Latina, después de Venezuela, con el mayor volumen de solicitudes de visas de turista. Estos datos subrayan la necesidad crítica de métodos eficientes y efectivos para determinar si vale la pena o no presentarse, así como el análisis de los perfiles o datos determinantes para la aprobación de una visa, empleando análisis de datos e inteligencia artificial.

Colombia es el país número 10 en el top de países que más solicitan visas de no inmigrante en el mundo lo que nos muestra un interés importante por parte de los colombianos por querer migrar a otros países y principalmente a Estados unidos, además de acuerdo con la información contenida por el Dane este mismo es el país número 1 entre los destinos turísticos que visitan los colombianos. **[1]**

Figura 1 – Tabla de países con sus solicitudes.

Figure 1 – Table with countries and their visa applications.

Posición	País	Cantidad de solicitudes
1	México	4,096,341
2	China	2,080,936
3	India	1,931,546
4	Brasil	573,669
5	Argentina	567,273
6	Perú	562,901
7	Venezuela	555,732
8	República Dominicana	550,281
9	Filipinas	548,773
10	Colombia	545,960

Debido a la pandemia de COVID-19, los puestos de las embajadas encargados de la emisión de visas suspendieron su operación hasta nueva orden lo cual generó un retraso importante en la asignación de citas para la solicitud de visas, por esta razón y teniendo en cuenta que la cantidad de solicitudes de visas va en aumento cada año se presentará aún más un inconveniente de agendamiento de citas.

Para el caso de Colombia de acuerdo con los comunicados internos de la misma embajada se estiman 688 días calendario, aproximadamente 23 meses". **[3]**

El análisis de datos desempeña un papel fundamental en la extracción de ideas valiosas y facilita la toma de decisiones informadas. Al examinar e interpretar conjuntos de datos extensos, se pueden identificar patrones y tendencias, revelando información oculta y permitiendo análisis predictivos.

"La ciencia de datos, un campo que abarca el big data, la analítica empresarial y la inteligencia de negocios, sirve como término general para esta práctica" **[2]**.

2. Estudio de dominio

2.1. La importancia del análisis de datos en la emisión de visas

El análisis de datos implica extraer ideas de los datos, aprovechando el big data y analizándolo para obtener información valiosa para la toma de decisiones **[4]**. Un dominio específico donde el análisis de datos resulta beneficioso es en el proceso de solicitud de visa.

Este artículo se benefició de la colaboración con una firma colombiana llamada Visas Gómez y Asociados, que además de ser un referente en temas de migración amablemente nos proporcionó acceso a los formularios de sus clientes previa autorización del uso de los mismos por parte de sus clientes a través de convenios de confidencialidad, cabe aclarar igualmente que se realizó un proceso previo de anonimización.

La emisión de visas representa un procedimiento complejo y crucial, especialmente en lo que respecta a la aprobación de visas para ingresar a los Estados Unidos.

Los Estados Unidos tienen criterios y requisitos específicos que las personas deben cumplir para adquirir una visa. Estos criterios abarcan elementos como estabilidad financiera, intención de regresar a su país de origen y posibles preocupaciones de seguridad. Al emplear el análisis de datos, específicamente la detección de perfiles, los solicitantes pueden tener una idea previa de aspectos a mejorar antes de una presentación aumentando así la tasa de éxito en la obtención de una visa de ingreso a los Estados Unidos, por que el aspirante podría abstenerse de la presentación o mejorar su perfil mientras es analizado su caso.

A través de este análisis, se desarrolló una calculadora basada en inteligencia artificial que genera una probabilidad de éxito en la aprobación de su visa evaluando el estado actual del solicitante permitiendo principalmente dos cosas. 1. Que una persona que no tenga un perfil adecuado pueda tomar la decisión de mejorar y aplicanr las recomendaciones entregadas. 2. Desistir de la presentación o aguardar al momento idóneo para realizar su presentación, lo que podría permitir reducir las presentaciones de los solicitantes. Esta calculadora está en constante actualización de acuerdo con la información obtenida de procesos exitosos o no de los clientes nuevos de la compañía.

2.2. El papel de la detección de perfiles en la emisión de visas

"A raíz de los cierres prolongados en las embajadas de EE. UU. en todo el mundo, el Departamento de Estado acumuló tiempos de espera muy largos para las visas de no inmigrante, que se promedian en 250 días en varios países. Estos son particularmente largos en Bogotá, Colombia, y Ciudad de México, México, donde los tiempos de espera que se manejan en las solicitudes superan los dos años. ". **[3]**

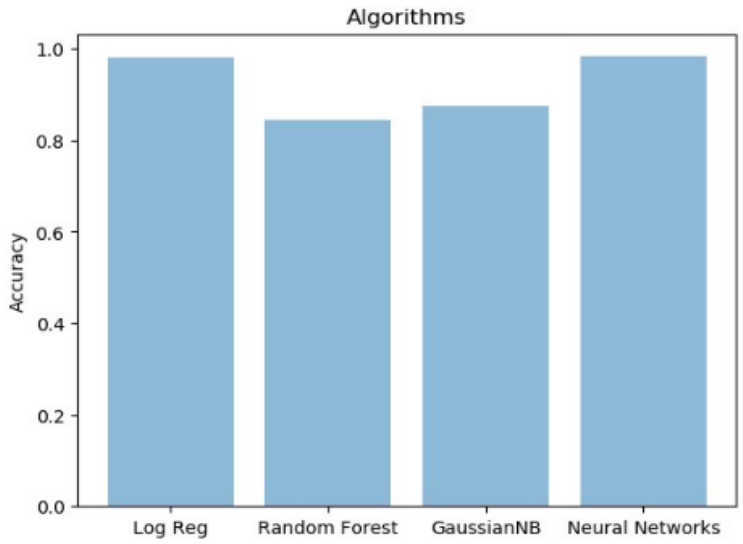
La detección de perfiles desempeña un papel fundamental en el proceso de emisión de visas, ya que ayuda a reconocer patrones y características de individuos con mayor probabilidad de obtener una visa. Teniendo en cuenta que la emergencia global debido a COVID-19 afectó desde hace ya varios años significativamente el procedimiento de emisión de visas, lo que resultó en el cierre indefinido de operaciones consulares y que como se mencionó como consecuencia las citas para entrevistas de visa ahora enfrentan un período de espera de hasta dos años. Comprender este escenario indica la importancia de facilitar el proceso para un aspirante, así como el desarrollo de una herramienta de inteligencia artificial para estimar la probabilidad de obtener una visa de turista, reduciendo potencialmente la afluencia de solicitantes a este proceso prolongado.

La afectación se ha venido manteniendo con los años e incluso han aumentado los tiempos de espera, sin embargo la embajada realiza diferentes acciones para cambiar este panorama **[5]** como se puede evidenciar en sus últimos comunicados donde indican incluso que duplicarán el esfuerzo de los oficiales consulares y de su personal en general para mitigar y garantizar unos tiempos de respuesta óptimos, más aun teniendo en cuenta que se aproxima el mundial de fútbol lo que normalmente tiene un auge turístico para los países anfitriones.

Prateek y Karun **[2]** utilizaron algoritmos de clasificación para detectar perfiles y predecir los resultados de visas en su estudio. Analizando datos de receptores de visas H-1B, aplicaron estos algoritmos para identificar aquellos perfiles con mayores probabilidades de éxito en la obtención de visas. Además, otros estudios como el realizado por Dombe **[6]** los dos algoritmos utilizados para su estudio que más tuvieron relevancia superando el 96% fueron la regresión logística y una red neuronal, podríamos tomar relevancia a cualquiera de estos dos para realizar nuestro estudio. (pp.193-202), aunque para este proyecto se realizó un análisis más completo incluyendo 5 algoritmos y de acuerdo con la efectividad de los mismos se tomó como resultado uno de ellos.

Figura 2 – Efectividad de algoritmos para perfiles de visa.

Figure 2 – "Effectiveness of algorithms for visa profiles"



Esto subraya el potencial de examinar diversos factores de los solicitantes, incluida la historia laboral, calificaciones, estabilidad financiera y antecedentes penales, para crear un perfil capaz de predecir con precisión la probabilidad de aprobación de visa.

2.3. Visión General

Basándonos en investigaciones previas sobre la detección de perfiles para resultados de visas, el trabajo propuesto en Colombia implica recopilar datos de solicitantes de visa y realizar un análisis de datos exhaustivo. Esta colaboración de investigación se lleva a cabo con Visas Gómez y Asociados, una empresa colombiana especializada en ayudar a personas que buscan visas de turista para ingresar a los Estados Unidos. Notablemente, aprovechan ideas psicológicas para mejorar el rendimiento en las entrevistas. Al emplear más de 100 variables obtenidas del formulario DS-160 completado por los solicitantes, estos datos son cruciales para evaluar la elegibilidad de las personas para las visas.

Entender conceptos clave en el entorno empresarial es fundamental:

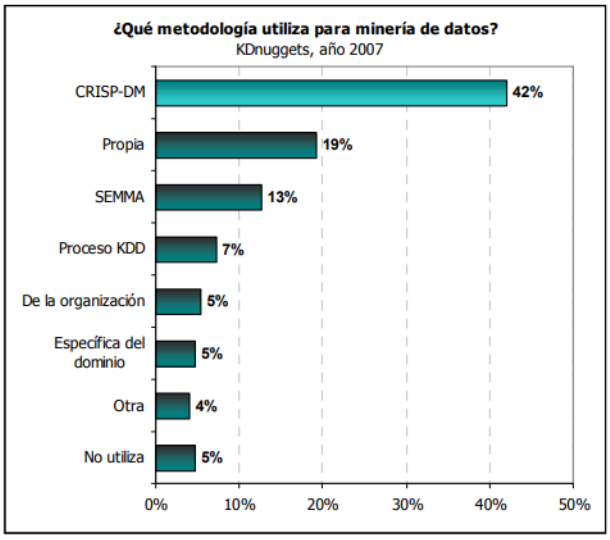
- Visa B1/B2: Visas de no inmigrante que otorgan a los visitantes la entrada a los Estados Unidos.
- Formulario DS-160: Un formulario desarrollado por la Embajada de EE. UU. para recopilar datos esenciales que verifiquen el cumplimiento de los requisitos de visa de turista.
- Entrevista consular: Una cita crucial en la cual los funcionarios consulares verifican e indagan sobre la información proporcionada en el Formulario DS-160. Actualmente, en Colombia, estas citas se programan con un período de espera de 2 años.

3. Definición de la metodología

Para determinar la metodología que se va a utilizar en el desarrollo del proyecto realizó la revisión de las metodologías que se usan en la minería de datos, dentro de las más usadas en la actualidad se encuentran KDD, CRISP-DM y SEMMA. **[7],** Aunque se deben tener en cuenta algunos estudios comparativos entre las metodologías, entre ellos se encuentra uno realizado por miembros de la Red de Universidades con Carreras en Informática (RedUNCI), titulado “Estudio comparativo de metodologías para minería de datos” **[8]** muestra una breve reseña de varias técnicas de minería de datos, además de que muestra una gráfica del uso de las metodologías.

Figura 3 - Encuesta realizada por la KDnuggets

Figure 3 - Encuesta realizada por la KDnuggets



Esto muestra una gran aceptación de las 3 metodologías anteriormente mencionadas y la utilización de una

metodología propia que cada una de las personas que implementa minería de datos diseña o ajusta a sus necesidades; por tal motivo se requiere el análisis de cada una de estas 3 metodologías para encontrar en sus características propias cual es la que más se ajusta al objetivo de negocio planteado.
En general, la metodología de análisis y minería de datos aplicada sigue un proceso similar, que consta de las siguientes fases:

- Exploración de datos: Se recopilan y exploran datos para identificar patrones y tendencias.

- Preparación de datos: Los datos se limpian y transforman para que sean adecuados para el análisis.

- Minería de datos: Se utilizan técnicas estadísticas y de aprendizaje automático para extraer conocimiento de los datos.

- Interpretación de resultados: Los resultados de la minería de datos se interpretan para obtener conclusiones útiles. En este caso, elegimos CRISP-DM, considerando la relevancia que le entrega al conocimiento del negocio lo cual nos parece crucial para no solamente depender del experto del dominio en el proceso teniendo en cuenta que los datos actuales, aunque existen no están en una base de datos organizada.

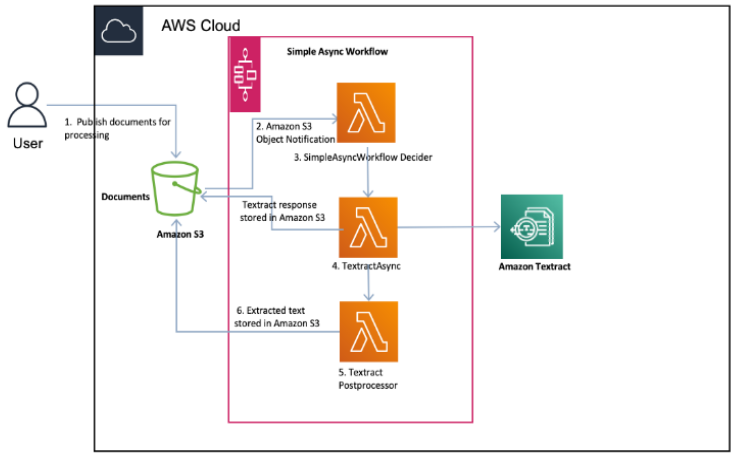
3. Definición del conjunto de datos

3.1. Obtención de Datos

Para facilitar este proceso, fue imperativo emplear software capaz de extraer información OCR (Reconocimiento Óptico de Caracteres) de cada variable completada por un solicitante. A continuación, se presenta un breve resumen de esta implementación. El uso de software fue indispensable debido a la complejidad de manejar hasta 400 variables por usuario. Obtener manualmente esta extensa información era inviable y obstaculizaba la obtención de datos consistentes, se realizó el desarrollo en python que tomaba los documentos de los formularios en PDF y en conjunto con una arquitectura AWS, dejaba los documentos en un servicio S3, donde posteriormente el servicio Textract en este caso puntual el componente de forms realizaba el reconocimiento de OCR realizando un trabajo asincrónico dejando una cola de trabajos pendientes para posteriormente tener la notificación de los elementos finalizados, tomando esta información se construyó una primera base de datos en CSV, importante tener en cuenta que como cada formulario puede tener información aunque similar en algunos caso más extensa, se debe construir un elemento que organice la información recopilada en formato key->value a un formato de tabla, para su posterior manipulación en el software R, a continuación un ejemplo de la arquitectura utilizada (Figura 4).

Figura 4 – Arquitectura utilizada para el reconocimiento OCR.

Figure 4 – Architecture used for OCR recognition.



3.2. Semántica de Datos

Se obtuvieron los datos del formulario DS160 de los clientes de la compañía Visas Gómez y Asociados en Colombia quien presta servicios de asesoría para visas de turista. Al obtener el conjunto de datos el siguiente paso fue determinar la semántica asociada con cada campo de 393 campos obtenidos del formulario, Se puede evidenciar la cantidad de información en diferentes medios como las guias entregadas por la Embajada de los Estados Unidos **[9]**.

A continuación, se presentan algunos atributos del formulario, **los cuales tuvieron relevancia estadística a través del modelo matemático generado al final de este ejercicio y también con la estadística exploratoria y correlaciones corroboradas por los expertos del dominio** con la variable aprobado o no, basado en el análisis que puede ser evidenciado más adelante:

- Estado Civil: Las opciones incluyen soltero, casado, unión libre, divorciado y viudo.
- País/Región de Origen: Principalmente incluye colombiano, venezolano, chileno, República Dominicana, etc.
- Ciudades de llegada: Como Miami, Tampa, Orlando, Nueva York, Los Ángeles, entre otras.
- Acompañado por otras personas: Sí o No.
- Visitas Previas a los EE. UU.: Sí o No.
- Plataformas de Redes Sociales: Ejemplos incluyen Instagram, Facebook, TikTok, LinkedIn, etc.
- Rechazo de Visa Previo de EE. UU.: Sí o No.
- Duración del Último Viaje a EE. UU. (si corresponde): Las categorías incluyen menos de 10 días, entre 10 y 20 días, entre 20 y 30 días, y más de 30 días.
- Patrocinio Financiero para Viajes: Sí o No.

Algunos otros datos como: edad que normalmente se pensaría que pueden llegar a tener relevancia aunque se incluyeron en el análisis no tienen relevancia estadística, de acuerdo con las pruebas realizadas posteriormente con la aplicación del modelo de regresión logística.

Figura 5 – Columnas de algunos datos generados en CSV.

Figure 5 – Columns of some data generated in CSV.

US_Social_!	Full_Name_	City	Other_Name	Sex	Date_of_Birt	Age	Age1	Place_of_Bi	Telecode_Nu	US_Taxpaye	Marital_Stat	National_Ide	Are_you_a_permanent_resid	CountryReg	Do_you_hol	City_State_	Date_of_De
DOES NOT APF	DOES NOT APF	BOGOTA	NO	FEMALE	30 OCTOBER 1	30/10/1969	54	BOGOTA, COL	NO	DOES NOT APF	SINGLE	52015081	NO	COLOMBIA	NO	PUERTO RICO,	20 MARCH
DOES NOT APF	DOES NOT APF	BOGOTA	NO	MALE	27 OCTOBER 1	27/10/1999	24	VILLETA, CUND	NO	DOES NOT APF	SINGLE	1007161451	NO	COLOMBIA	NO	MIAMI, FLORID	29 JUNE 20

4. Preparación de los datos

4.1. Procesamiento de Datos

No obstante, se emplearon técnicas estadísticas descriptivas para comprender las características de los datos y obtener ideas para la detección de perfiles, ya sea aprobados o no como: Medidas de tendencia central, correlaciones, patrones visuales comparados con los expertos del dominio como histogramas, análisis de regresión. Principalmente, nuestro enfoque radica en examinar procesos que han arrojado un resultado concluyente. Específicamente, el objetivo fue analizar los perfiles de individuos que han obtenido con éxito una visa para Estados Unidos desde Colombia, haciendo énfasis en técnicas pertinentes de análisis de datos.

Se importó el conjunto de datos al software R para explorar estadísticamente y validar la relevancia de los datos y variables incluidas en el conjunto de datos con expertos en el campo para la detección de perfiles.

4.2. Análisis Estadístico

Se realizó el análisis de histogramas, por ejemplo, con edad, educación, ingresos, ciudad, estado civil y otras variables relevantes para identificar posibles patrones y tendencias.

En el ejemplo siguiente, es evidente que la mayoría de los clientes que atraviesan este proceso se encuentran en el rango de edad de 20 a 40 años, con un pico a los 40 años.

Figura 6 –Histograma de edad.

Figure 6 – Age histogram.

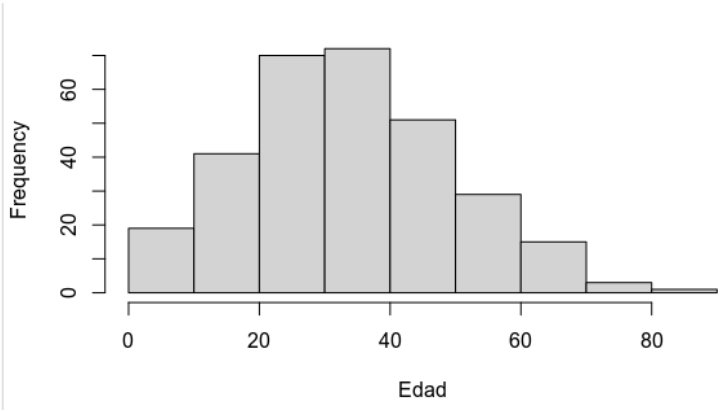


Figura 7 –Histograma de ciudades de nacimiento.

Figura 7 –Born City histogram.

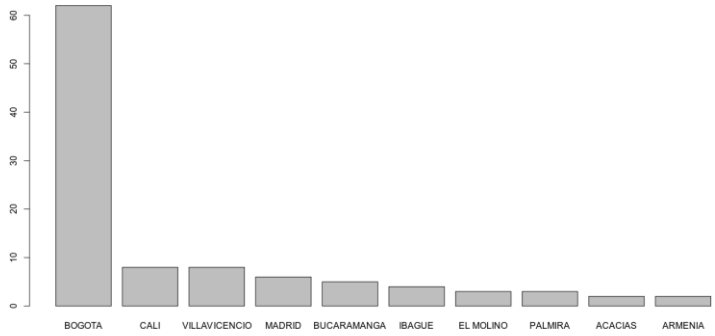


Figura 8 –Histograma de estado civil.

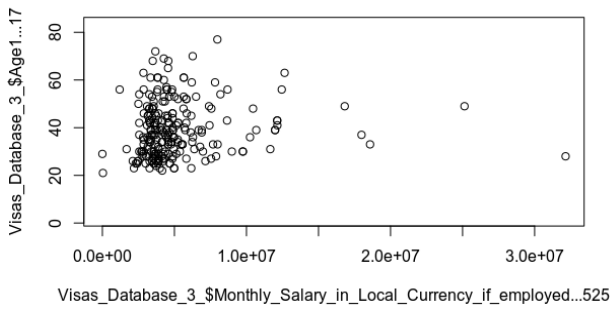
Figure 8 – Marital status histogram.

4.3. Correlaciones

Se obtuvieron datos adicionales significativos a través de comparaciones y correlaciones entre variables como edad e ingresos. En este análisis, es notable que solo hay algunas instancias de ingresos excepcionalmente altos, mientras que la mayoría de los solicitantes poseen ingresos moderados.

Figura 9 –Gráfica de relación entre edades e ingresos.

Figure 9 – Graph of the relationship between ages and incomes.



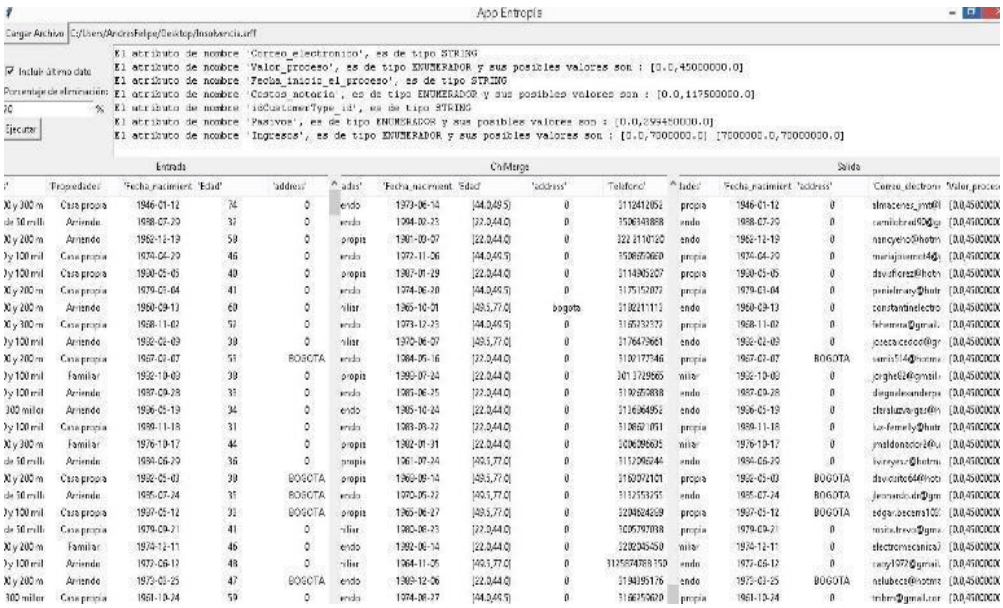
Como se mencionó anteriormente, se mantiene constante la consideración de que los datos atípicos habían sido corregidos por un psicólogo durante el proceso de creación del formulario, garantizando la precisión de la información proporcionada. Sin embargo, sólo una cantidad limitada de datos necesitó corrección.

4.4. Depuración de Entropía

Haciendo referencia al artículo titulado 'Un Estudio Comparativo entre Algoritmos de Selección de Características' en el Capítulo 4.2, desarrollamos un software personalizado destinado a reducir la dimensionalidad mediante la implementación de algoritmos como Chi Merge. Este proceso permite el refinamiento de datos y mejora la efectividad del proceso de minería de datos.

Figura 10 –Resultado software entropía aplicado a unas columnas.

Figure 10 – Result of entropy software applied to some columns.



En este escenario, es crucial discretizar ciertas variables, ya que hacerlo podría mejorar significativamente el modelo estadístico final. Por ejemplo, la regresión logística tiende a funcionar mejor con variables discretizadas.

El software proporciona retroalimentación que sugiere la eliminación de ciertas variables, como números de teléfono, zonas de códigos postales o instancias donde la dirección de correo electrónico coincide con la dirección de aplicación. Estas recomendaciones, basadas en la opinión de expertos, se consideraron viables para su eliminación, lo que resultó en un conjunto de datos mejorado.

5. Desarrollo del software calculadora basado en un modelo estadístico

5.1. Modelo estadístico basado en una regresión logística

Considerando el objetivo del proceso, que gira en torno a una variable dependiente con solo dos opciones: aprobado o no aprobado y teniendo en cuenta que de acuerdo con el conjunto de datos de prueba y de entrenamiento basados en el histórico de la compañía el modelo que más efectividad tiene es el de la regresión logística con un 88% como se puede ver a continuación optamos por utilizar este algoritmo.

Figura 11 –Análisis de diferentes modelos estadísticos.

Figure 11 – Analysis of different statistical models.

```
# Definir modelos
models = {
    'Logistic Regression': LogisticRegression(),
    'Support Vector Machine': SVC(),
    'Gradient Boosting': GradientBoostingClassifier(),
    'K-Nearest Neighbors': KNeighborsClassifier()
}

# Entrenar y evaluar cada modelo
for name, model in models.items():
    classifier = Pipeline(steps=[('preprocessor', preprocessor), ('classifier', model)])
    classifier.fit(X_train, y_train)
    accuracy = classifier.score(X_test, y_test)


    print(f'{name} Accuracy: {accuracy}')
```

```
Logistic Regression Accuracy: 0.8809836065573771
Support Vector Machine Accuracy: 0.7918032786885245
Gradient Boosting Accuracy: 0.8737704918032787
K-Nearest Neighbors Accuracy: 0.8737704918032787
```

Posteriormente, procedimos a realizar análisis separados entre nuestra variable objetivo y otras. Este enfoque buscaba lograr concordancia a través de correlaciones y análisis previos, ayudándonos a discernir evaluando la hipótesis sobre qué variables son genuinamente relevantes en este proceso, incluyendo entre otras las variables listadas a continuación:

Figura 12 – Verificación hipótesis relación estadística entre edad y aprobación de visa.

Figure 12 – Hypothesis verification of the statistical relationship between age and visa approval.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.5465     0.2679  -2.040   0.0413 *
factor(df$AgeRange)20-39   0.1762     0.3177   0.555   0.5792
factor(df$AgeRange)40-59   0.5966     0.3490   1.709   0.0874 .
factor(df$AgeRange)60-79   1.5021     0.5905   2.544   0.0110 *
factor(df$AgeRange)80-99  15.1126   882.7434   0.017   0.9863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Figura 13 – Verificación hipótesis relación estadística entre haber permanecido diferentes rangos de días y aprobación de visa.

Figure 13 – Hypothesis verification of the statistical relationship between staying in different ranges of days and visa approval.

```
Console   Terminal ×   Background Jobs ×
R 4.3.0 · ~/
> cat("Porcentaje de casos donde la columna binaria es 1 y los dias son mayores a x:", relacion, "%\n")
Porcentaje de casos donde la columna binaria es 1 y los dias son mayores a x: 41.52824 %
>
> # Paso 5: Verificar relación entre la columna binaria y la cantidad de dias
> datos_filtrados <- datos_modificados %>%
+     filter(datos_modificados$Aproved == 1 & datos_modificados$dias > 1000)
>
> # Paso 6: Realizar análisis de la relación
> relacion <- nrow(datos_filtrados) / nrow(datos_modificados) * 100
>
> # Paso 7: Imprimir resultados
> cat("Porcentaje de casos donde la columna binaria es 1 y los dias son mayores a x:", relacion, "%\n")
Porcentaje de casos donde la columna binaria es 1 y los dias son mayores a x: 44.51827 %
~
```

Figura 14 – Verificación hipótesis relación estadística entre tener planes de viajes y aprobación de visa.

Figure 14 – Hypothesis verification of the statistical relationship between having travel plans and visa approval.

```
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                          -1.0986     1.1547  -0.951    0.341
Visas_Database_3_$Have_you_made_specific_travel_plans...45null   1.3749     1.1888   1.157    0.247
Visas_Database_3_$Have_you_made_specific_travel_plans...45YES    0.8038     1.1619   0.692    0.489

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 414.08  on 300  degrees of freedom
Residual deviance: 409.98  on 298  degrees of freedom
AIC: 415.98
```

5.2. Implementación de código en python:

- Importación de las librerías para su uso.

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
```

- Carga de datos:

```
import pandas as pd
# leer la base de datos, deberá ser actualizada por la bbdd de la plataforma.
visas  data = pd.read  csv(os.path.join(settings.STATIC  ROOT, 'tablaback.csv'))
visas  data.head()
```

- Generación del modelo

```
features = ['AgeRange', 'CountryRegion_of_Origin_Nationality...24', 'haveinus', 'social', 'personpay']
X = visas_data[features]
y = visas_data['Aproved']  # Variable objetivo
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
categorical_features = ['AgeRange', 'CountryRegion_of_Origin_Nationality...24', 'haveinus', 'social', 'personpay']
categorical_transformer = Pipeline(steps=[
   ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
preprocessor = ColumnTransformer(
   transformers=[
      ('cat', categorical_transformer, categorical_features)
   ])
model = Pipeline(steps=[('preprocessor', preprocessor),
                        ('classifier', LogisticRegression())])
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print(f'Accuracy: {accuracy}')
```

Verificando la precisión, se alcanza una tasa de **precisión de más del 88%** lo que proporciona una estimación sólida para este modelo, lo que nos permitiría proceder con la creación de una solución calculada.

- Envío de datos insertados por el usuario

```
new_data = pd.DataFrame({'AgeRange': [self.object.AgeRange], 'CountryRegion_of_Origin_Nationality...24':
[self.object.originNationality], 'haveinus': [self.object.haveinus], 'social': [self.object.social], 'personpay': [self.object.personpay]
})
```

- Realización de la predicción probabilística.

```
prediction = model.predict(new_data)
print(prediction)
# Realiza la predicción de probabilidad
probabilities = model.predict_proba(new_data)
# Imprime la probabilidad de la clase positiva
positive_probability = probabilities[:, 1]
positive_probability = positive_probability * 100
```

- Resultado final.

Figura 15 – Calculadora de probabilidad en producción.

Figure 15 – Probability calculator in production.

Finalmente, el solicitante puede utilizar la calculadora para evaluar si desea mejorar ciertos aspectos de su vida o considerar esperar hasta que la situación sea diferente y más favorable, la alimentación de la información de nuevos clientes así como si el proceso fue exitoso o no realizará el entrenamiento del algoritmo nuevamente permitiendo mejorar su efectividad. Esta calculadora se puede ver funcionando en la siguiente URL: https://visasgomezyasociados.com/calculadora/

6. Conclusiones

- Se puede evidenciar la utilidad de integrar el análisis de datos y la inteligencia artificial para crear una herramienta que permite a los postulantes de visa conocer la probabilidad de aprobación y recibir sugerencias personalizadas. A través de un modelo predictivo basado en datos históricos, la calculadora ayuda a los usuarios a tomar decisiones informadas y mejorar sus posibilidades de éxito.
- Se demostró cómo el análisis de datos e inteligencia artificial pueden ayudar a los postulantes de visa a evaluar sus probabilidades de aprobación y mejorar su solicitud. Si la herramienta se masifica, podría no solo hacer que algunos

desistan, sino también que otros decidan esperar el tiempo que actualmente brinda la embajada, aprovechando ese periodo largo para optimizar su perfil y aumentar sus posibilidades de éxito.

- Continuar almacenando nuevos datos no solamente con el histórico realizado este proyecto sino con los clientes nuevos es crucial, lo que permite una mejora continua del modelo con cada iteración que permitirá entrenar la IA que realiza la predicción.

Referencias

[1] Report of the Visa Office 2022. (2023). *Travel.gov*. Available: https://travel.state.gov/content/travel/en/legal/visa-law0/visa-statistics/annual-reports/report-of-the-visa-office-2022.html [Accessed: Feb. 16, 2023]

[2] A. Prateek and S. Karun, "A survey on data mining techniques for customer churn prediction," *International Journal of Data Mining and Knowledge Management Process*, vol. 7, no. 3, pp. 1-25, 2017.

[3] Embajada de los Estados Unidos en Colombia, "Tiempo de espera para una cita de visa", *Embajada de los Estados Unidos en Colombia*, 2025. [Enlace]. Disponible en: https://co.usembassy.gov/es/visas-es/tiempo-de-espera-para-una-cita-de-visa/. [Accedido: 24-ene-2025].

[4] S. Triche, F. Goncalves, and J. Gama, "A survey on data preprocessing for classification tasks," *Data Mining and Knowledge Discovery*, vol. 30, no. 1, pp. 1-36, 2016.

[5] Embajada de los Estados Unidos en Colombia, "Durante los próximos seis meses, duplicaremos nuestras horas de trabajo," Facebook, 17 de enero de 2025. Disponible en: https://www.facebook.com/reel/1627548918141285 [Accedido: 15 de enero de 2025].

[6] A. Dombe, *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, D. Swain, P. K. Pattnaik, and P. K. Gupta, Eds. Singapore: Springer Nature, 2020.

[7] E. L. Guzmán, «Minería de Datos,» [En línea]. Available: https://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf. [Último acceso: 01 04 2020].

[8] J. M. Moine, A. S. Haedo y S. Gordillo, «Estudio comparativo de metodologías para minería de datos,» XIII Workshop de Investigadores en Ciencias de la Computación, pp. 278-281, 2011.

[9] Embajada de los Estados Unidos, "Guía para completar el Formulario DS160 para la Visa de No Inmigrante," YouTube, 24 de enero de 2025. [Enlace: https://www.youtube.com/watch?v=SdIb3USweyk]