UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

ESCUELA SUPERIOR DE TLAHUELILPAN

LICENCIATURA EN INGENIERÍA DE SOFTWARE

TESIS

ANÁLISIS DE DATOS POBLACIONALES PARA EL DESARROLLO DE LA PLATAFORMA DATAMEX

Para obtener el título de

Licenciado en Ingeniería de Software

PRESENTA

Uriel Navarrete Garamendi

Director

Dr. Gabriel Sánchez Bautista

Comité tutorial

Dra. Julia Yazmín Arana Llanes
Dr. Gustavo Padrón Rivera
Dr. Gabriel Sánchez Bautista
Mtra. Matilde Reyes Fuentes

Tlahuelilpan, Hidalgo., junio 2025

Universidad Autónoma del Estado de Hidalgo
Escuela Superior de Tlahuelilpan
Campus Tlahuelilpan

11 de junio de 2025

Asunto: Autorización de impresión formal.

M.C. MIGUEL ÁNGEL DE LA FUENTE LÓPEZ
Director de la Escuela Superior de Tlahuelilpan

Manifestamos a usted que se autoriza la impresión formal del trabajo de investigación del pasante Uriel Navarrete Garamendi, bajo la modalidad de tesis individual, cuyo título es: "ANÁLISIS DE DATOS POBLACIONALES PARA EL DESARROLLO DE LA PLATAFORMA DATAMEX" debido a que reúne los requisitos de decoro académico a que obligan los reglamentos en vigor para ser discutidos por los miembros del jurado.

"AMOR, ORDEN Y PROGRESO"

Nombre de integrantes del jurado	Cargo	Firma
Dra. Julia Yazmín Arana Llanes	Presidente	
Dr. Gustavo Padrón Rivera	Secretario	
Dr. Gabriel Sánchez Bautista	Vocal	
Mtra. Matilde Reyes Fuentes	Suplente	

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
ESCUELA SUPERIOR DE TLAHUELILPAN

Ex-Hacienda de San Servando S/N, Col. Centro, Tlahuelilpan, Hidalgo, México; C.P. 42780
Teléfono: 771 71 720 00 Ext. 50601 y 50603
esc_sup_tlahuelilpan@uaeh.edu.mx
uaeh.edu.mx

Agradecimientos

Agradezco a Dios, por darme la oportunidad de vivir y poder demostrar todo lo que puedo ser emocional y profesionalmente, le agradezco por guiarnos y nunca soltarnos de la mano a mí y a mi familia.

Agradezco a mi Madre, por apoyarme en toda mi trayectoria como estudiante e inculcarme cada valor que ahora me permiten tener ese gran compromiso con el trabajo y en todo lo que hago, ha sido la mayor fuente de inspiración en mi vida, ese ejemplo de salir adelante con esfuerzo y dedicación siempre lo tendré presente, y en algún momento espero ser esa fuerza para mis hijos, estoy feliz de haber concluido esta etapa para poder devolverle todo lo que ha hecho por nosotros, te amo.

Agradezco al Dr. Gabriel Sánchez Bautista, por la atención brindada y ese gran apoyo en todo momento durante la realización de este trabajo, sin su orientación en los temas no hubiera sido posible la conclusión del mismo.

Resumen

El presente trabajo de investigación se centró en el análisis de datos en México, abordando aspectos críticos como seguridad, educación, economía, salud y empleo, con el objetivo realizar un análisis de datos para facilitar la toma de decisiones. Se emplearon técnicas de ciencia de datos para analizar esta información, cuyos resultados fueron integrados en una plataforma web denominada DATAMEX, desarrollada para permitir a los usuarios consultar datos de interés y realizar predicciones útiles. La plataforma web, desarrollada con tecnologías como HTML, CSS y JavaScript, se creó para ser accesible y comprensible, adaptándose a diferentes dispositivos. Este proyecto subraya la importancia de contar con herramientas que permitan a la población, empresas e instituciones gubernamentales acceder a información crítica de manera clara y efectiva. A lo largo de esta investigación, se aplicaron diversas técnicas y modelos predictivos entre los que se encuentran regresión lineal, árboles de decisión y regresión de vectores de soporte, proporcionando una base sólida para futuras investigaciones e incentivar el desarrollo de políticas públicas más efectivas. Por otro lado, para la parte de visualización de datos se utilizaron bibliotecas de Python como seaborn y matplotlib. Cabe mencionar que a pesar de las limitaciones encontradas, como la dependencia de datos históricos o la recopilación de datos de distintas fuentes, este trabajo destaca la necesidad de realizar un análisis de datos en México para ofrecer una visión actualizada. La plataforma DATAMEX puede tener un impacto significativo en diversos sectores, contribuyendo al bienestar y desarrollo de México.

Abstract

This research focused on data analysis in Mexico, addressing critical aspects such as security, education, economy, health and employment, with the aim of performing data analysis to facilitate decision making. Data science techniques were used to analyze this information, the results of which were integrated into a web platform called DATAMEX, developed to allow users to consult data of interest and make useful predictions. The web platform, developed with technologies such as HTML, CSS and JavaScript, was created to be accessible and understandable, adapting to different devices. This project highlights the importance of having tools that allow people, companies and government institutions to access critical information in a clear and effective way. Throughout this research, various techniques and predictive models were applied such as linear regression, decision trees and support vector regression, providing a solid basis for future research and encouraging the development of more effective public policies. On the other hand, for the data visualization, Python libraries such as seaborn and matplotlib were used. Despite the limitations found, such as the dependence on historical data or the collection of data from different sources, this work highlights the need to perform data analysis in Mexico to offer an up-to-date view. The DATAMEX platform can have a significant impact on various sectors, contributing to the well-being and development of Mexico.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Construcción del objeto de estudio

1.1. Introducción

Este trabajo de investigación se basa en el análisis de datos poblacionales en México, utilizando técnicas avanzadas de ciencia de datos para identificar patrones, relaciones y tendencias que impactan la vida cotidiana de las personas, ya que el análisis de datos puede ser utilizado para tomar decisiones, por ejemplo, cuando se identifica que una variable socioeconómica se ve influenciada por el valor de otra variable. A través del desarrollo de una plataforma web, DATAMEX, se busca no solo facilitar el acceso a estos datos, sino también presentar los resultados de manera clara y comprensible, permitiendo a ciudadanos, empresas e instituciones gubernamentales tomar decisiones informadas y estratégicas. La creciente preocupación por temas económicos y sociales en México refuerza la necesidad de tener herramientas que ofrezcan un análisis profundo y accesible de los datos poblacionales. Este proyecto no solo aborda esta necesidad, sino que también busca contribuir al conocimiento sobre cómo los diferentes factores sociales y económicos interactúan y afectan el bienestar general de la población. Por lo anterior se prevé que el análisis de datos en conjunto con la plataforma sean una estrategia adecuada para la toma de decisiones por parte de las personas.

1.2. Planteamiento del Problema

En cada país, existen diferentes factores tales como seguridad, salud, medio ambiente, educación, economía, entre otros, que la población considera importantes debido a que tienen un impacto directo en su vida cotidiana. De acuerdo con lo anterior las personas pueden revisar los informes respectivos a cada ámbito emitidos a través de internet, una vez que estos sean publicados, y así puedan por ejemplo, tomar una mejor decisión al momento de elegir su lugar de residencia. México no es la excepción, tal como en otros países hay temas que son de gran interés. En la web es posible consultarlos de diferentes sitios, un ejemplo de ello es el Instituto Nacional de Estadística y Geografía (INEGI), los temas que aborda de acuerdo con su sitio web son: demografía y sociedad, economía y sectores productivos, geografía y medio ambiente, gobierno, seguridad y justicia. Sin embargo, en la actualidad no existe un sitio web que presente los resultados del análisis de datos en distintas áreas de México, y que muestre la información de manera clara y sea capaz de realizar predicciones con base a los datos. Todo esto en conjunto sería una herramienta de mucha utilidad para distintos panoramas en México.

1.3. Justificación

Actualmente estamos viviendo un cambio radical, la tecnología está en constante crecimiento, cada día surge algo nuevo e innovador que ayuda al ser humano a realizar sus tareas o simplemente enfocado al entretenimiento. En tal sentido, a lo largo de nuestra cultura la innovación es una parte que se ha destacado debido a las diferentes tendencias que han surgido, lo nuevo siempre sobresale, artefactos como el teléfono, la máquina de escribir, la televisión y la computadora son herramientas innovadoras, las cuales han facilitado el trabajo, lo que da cuenta que día a día hay un constante cambio. En esta era digital los datos son el nuevo recurso, gracias a su análisis pueden surgir áreas de oportunidad en las cuales sea posible encontrar tendencias y resolver problemas mediante ellos [13]. Hoy en día la gran cantidad de datos a la que se accede, creada de manera continua por parte de usuarios, entidades o servicios, ha dado lugar al desarrollo de métodos científicos e ingenieriles para contar con sistemas y procedimientos con la capacidad de almacenar, procesar y analizar gran cantidad de datos, generando información y conocimiento en áreas de suma importancia como la banca, finanzas, el marketing, redes sociales, la industria, el comercio electrónico, salud, medicina, ciudades inteligentes, entre otras [14]. Las personas están cada vez más preocupadas de cómo viven y si la situación económica y social mejorará en algún futuro. De acuerdo con el

estudio de Ipsos [15], el porcentaje de interés por parte de los mexicanos sobre algunos factores es: "Crimen y violencia (53 %), Desempleo (35 %), Pobreza y Desigualdad Social (30 %), Inflación (30 %), Corrupción (25 %), Educación (23 %), Cambio climático (22 %), Costo de la Salud (16 %), Programas Sociales (8 %), Covid-19 (8 %), entre otros. Es por ello que es necesario tener una herramienta que les brinde la facilidad de consultar información relevante sobre algunos temas de interés, tomando diferentes fuentes de información para llevar a cabo un análisis, mostrando la relación entre cada factor para la predicción de datos. Esta herramienta es una aportación para la sociedad ya que de acuerdo con [16, 17, 18], es importante que las personas se apoyen de la ciencia de datos para la toma de decisiones.

Como parte del trabajo de investigación fue aplicada una encuesta a 103 personas, la cual muestra en cada sección el interés de los usuarios acerca de contar con un análisis de datos más profundo sobre temas que ellos consideran de mayor preocupación. Los datos demográficos de las personas encuestadas se muestran en la Tabla 1.1

Entidad Federativa	**Rango de edad**	**Sexo**
Las personas encuestadas son de los 32 estados de la República Mexicana. Sin embargo, la mayoría son del Estado de Hidalgo (67 %), seguido por el Estado de México (7.8 %) y la Ciudad de México (5.8 %)	La población encuestada esta entre 15 y 69 años. Sin embargo, la mayoría esta en el rango de entre los 20 y 29 años (59.2 %), seguido por el rango de entre 15 y 19 años (16.5 %), el rango de entre 30 y 39 años (14.6 %) y el rango entre 40 y 49 años (7.8 %)	Mujer 40.8 % Hombre 59.2 %

Tabla 1.1: Datos demográficos de las personas encuestadas.

Los resultados de las preguntas de la encuesta se explican a continuación.

Pregunta: ¿Conoces alguna organización en México que recolecta datos estadísticos?

El objetivo de esta pregunta es saber qué tanto está informada la población acerca de la recolección de datos en México y de las organizaciones que

se encargan de ello. Los resultados muestran que la mayor parte de las personas encuestadas, para ser exactos un 81.6 % de la muestra, conocen por lo menos una organización que se encarga de recolectar datos estadísticos a nivel nacional. En la Figura 1.1 se muestran los porcentajes de respuesta para esta pregunta.

Figura 1.1: Organizaciones de recolección respecto a los datos estadísticos en México.

Pregunta: En tu opinión ¿consideras que el resultado del análisis de datos recolectados en México es presentado a la población de forma clara?

El objetivo de esta pregunta es saber qué tan bien se muestran los resultados del análisis a la población respecto a los datos recolectados en México.

En los resultados se puede observar que el 59.2 % de la muestra no está conforme con la manera en la que el resultado del análisis de datos se presenta a la población, mientras que el resto de la muestra, el 40.8 %, considera que el resultado del análisis se presenta de forma entendible. Los resultados de esta pregunta se muestran en la Figura 1.2.

Figura 1.2: La manera en la que se muestran los datos recolectados en México.

Pregunta: Selecciona los 5 temas que tienen mayor importancia para ti y sobre los que te gustaría que se realice un análisis de datos en México.

El objetivo de esta sección de la encuesta es identificar cuales con algunos de los temas que la población considera importantes en México para realizar un análisis de datos sobre ellos.

En los resultados aparecen una serie de temas que la población considera interesantes, entre los que se encuentran:

- La seguridad se posiciona como el tema de mayor interés. En México la inseguridad va en aumento año con año. La población reporta que se siente menos segura en los estados donde radican [19, 20]. La inseguridad se ha convertido en una de las más grandes problemáticas de México, tanto así que México puede ser considerado como un "estado fallido" debido a que en muchos casos se pierde el control de la situación y el crimen organizado se apodera del estado mediante la corrupción lo que conlleva que la población pierda la confianza en la policía, la justicia y los gobernantes recurriendo en algunos casos a la autoprotección [21].
- La educación ocupa el segundo lugar como tema de mayor importancia de acuerdo con la encuesta. La educación es parte fundamental para el

desarrollo personal y profesional de las personas. Con una buena educación las personas pueden alcanzar otros niveles de empleo, ayudándoles a un crecimiento social y económico. Especialmente, los jóvenes están en busca de una educación de calidad con el objetivo de ampliar sus oportunidades dentro de la ciencia, la tecnología y la innovación. Por esa razón el tema de la educación no se debe dejar a un lado debido al gran impacto en nuestro país [22].

- La economía de igual manera es un tema que genera un nivel de preocupación en algunas personas en México. Las personas en México y en otras naciones, día con día tienen diferentes necesidades tales como alimentación, educación, vivienda, salud, entre otras. Estos bienes y servicios dependen del desarrollo económico de cada país, algunos aspectos que contribuyen con la economía son la agricultura, la ganadería, el comercio, la industria, entre otros [23, 24]. A partir del año 2019 México pasó por una caída en su economía debido a los precios bajos del petróleo y a la crisis económica global ocasionada por la pandemia de COVID-19 en el año 2020. Sin embargo, se espera que en los próximos años la economía mexicana crezca gracias a las estrategias de desarrollo económico del Gobierno Mexicano. Sin un buen crecimiento económico las personas no pueden disponer de los bienes y servicios necesarios para subsistir, de acuerdo con lo anterior la economía es un tema importante para todos.

- Otro aspecto que se mantiene como tema de gran interés a nivel nacional es la salud. La preocupación por la salud personal siempre ha estado, pero aumentó a partir del año 2019 cuando surgió la enfermedad por coronavirus (COVID-19), la cual es una enfermedad infecciosa causada por el virus SARS-CoV-2. La preocupación sobre los posibles casos de infección por COVID-19 trajo consigo la prevalencia global de ansiedad y depresión. En la actualidad aún persisten brechas y preocupaciones [25]. Lo que aumenta más el impacto que tendrá la salud sobre la población es que en los últimos años el gobierno mexicano recortó alrededor de 15,000 millones de pesos del sector salud. En los siguientes años se prevé que el sistema enfrente una población con un nivel significativo de adultos mayores, además presentará un gran número de enfermos crónicos que soliciten tratamientos costosos y tardados [26].

- Empleo, el panorama laboral en México es complicado desde hace tiempo, los niveles de pobreza no disminuyen tanto, debido a un factor de riesgo que la provoca, la informalidad laboral. Sin contar con un factor externo que influyó mucho en su momento como fue la pandemia. Se estima que alrededor del 38.3 % de la población en México se encuentra en un nivel

de pobreza laboral, es decir, 49.2 millones de mexicanos viven en una situación en la que los ingresos laborales de su vivienda no son suficientes para adquirir los alimentos básicos para todos los integrantes [27, 28].

Los resultados de la encuesta de los temas de interés en la población mexicana se muestran en la Figura 1.3. Estas respuestas son relevantes ya que son una muestra de lo que opina la sociedad mexicana, ya que en este estudio se incluyeron personas de distintas entidades federativas y de distintas edades.

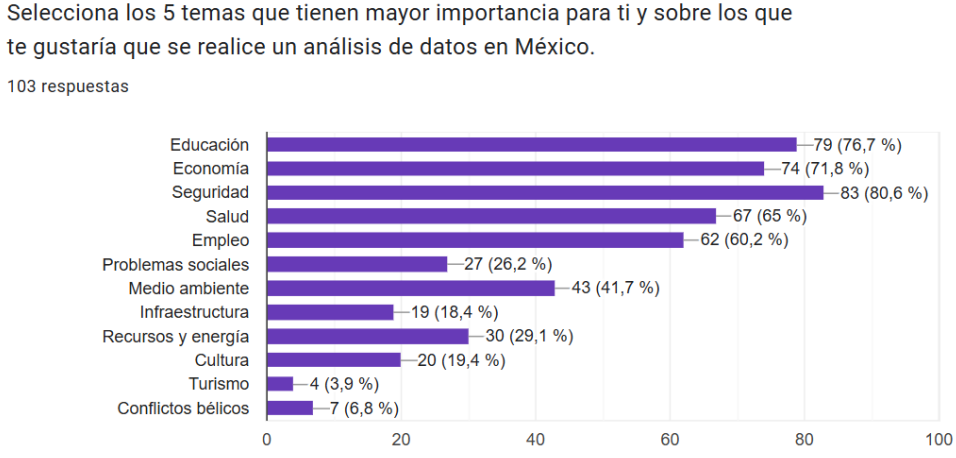


Figura 1.3: Los temas de mayor importancia para los mexicanos.

Pregunta: ¿De cuantas fuentes distintas de información te gustaría que se obtengan los datos para llevar a cabo el análisis?

En los resultados se observa que más de la mitad de los encuestados, el 62.1 %, está de acuerdo con que se usen distintas fuentes de información para obtener los datos que servirán para el análisis. Entre más fuentes de datos se utilicen, existirá un análisis más completo ya que considera información proveniente de diversos medios. Los resultados de esta pregunta se muestran en la Figura 1.4

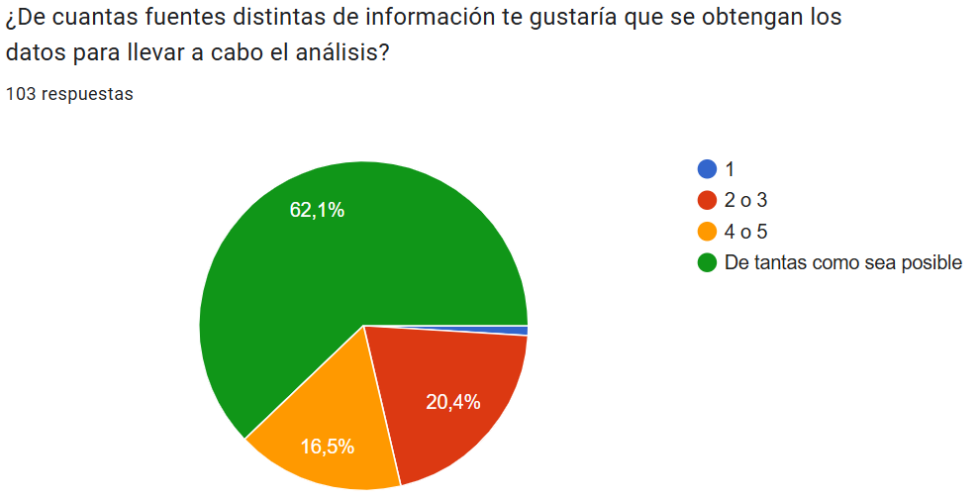


Figura 1.4: Número de fuentes de información.

Pregunta: ¿Te gustaría que el resultado del análisis de datos te muestre si existe una relación entre dos o más factores?

El objetivo de esta pregunta es conocer si a las personas les interesa saber si existen correlaciones que puedan existir entre los diferentes factores que se tomarán para el análisis de datos. Es decir, a través de esta pregunta se busca saber si a las personas les interesa identificar correlaciones entre variables dependientes e independientes, es decir cuando una variable se ve influenciada por otra de forma positiva o negativa.

En los resultados de la encuesta se observa que el porcentaje es significativo, con un 99 % a favor de que se identifiquen las correlaciones entre variables, notándose que las personas encuestadas consideran necesario identificar si existe una relación entre dos o más factores. Los resultados de esta pregunta con los respectivos porcentajes se observan en la Figura 1.5.

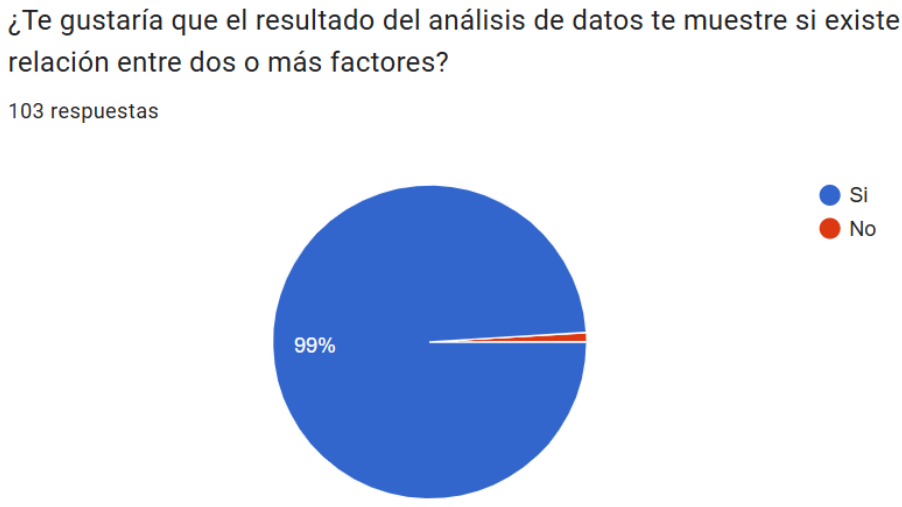


Figura 1.5: Correlación de factores.

Pregunta: ¿Te gustaría que el análisis de datos te muestre gráficas para la visualización de los resultados?

El objetivo de la pregunta anterior es para saber de qué manera se pueden presentar los resultados del análisis. La mayoría de los usuarios al momento de presentarles algún tipo de análisis respecto a datos estadísticos prefieren que los resultados obtenidos se puedan visualizar en gráficas que a su vez faciliten el entendimiento de los datos presentados. Es por ello que en esta sección de pregunta es importante identificar de qué forma es más adecuado presentar los resultados del análisis de datos, de tal forma que la información mostrada sea de utilidad para las personas que la consulten.

De acuerdo con la encuesta, al 100 % de las personas le gustaría que se puedan visualizar los resultados de manera gráfica. El resultado de esta pregunta se observa en la Figura 1.6

Figura 1.6: Una mejor forma para mostrar los resultados del análisis de datos.

Pregunta: ¿Te gustaría que además del análisis de datos históricos, también se incluya una predicción de indicadores para el futuro?

De acuerdo con los resultados, al 97.1 % de la muestra les parece buena idea la implementación de una predicción basada en el análisis de datos históricos. Los resultados de esta pregunta se muestran en la Figura 1.7.

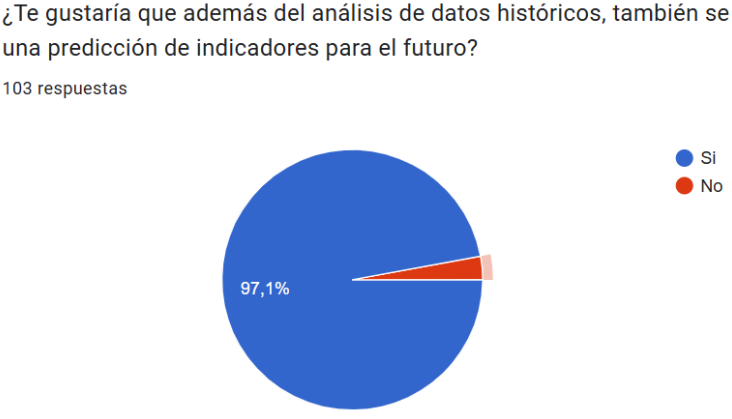


Figura 1.7: Análisis de datos históricos para predicción.

Pregunta: ¿Conoces alguna plataforma que realice análisis de datos acerca de los temas que a ti más te interesan, que tome distintas fuentes de información y que además incorpore predicción de datos?

El objetivo de la pregunta anterior es obtener el dato de que tantas plataformas son reconocidas a nivel nacional que realicen análisis de datos y predicciones con base en la recopilación de datos de distintas fuentes. Los porcentajes de respuesta de esta pregunta se muestran en la Figura 1.8. Se observa que el 77.7 % de la muestra dice no conocer alguna plataforma que implemente análisis de datos acerca de temas de interés para la población mexicana, incluyendo características como la predicción con base en los datos recolectados de distintas fuentes, por lo que se justifica el desarrollo de una plataforma que tenga estas características.

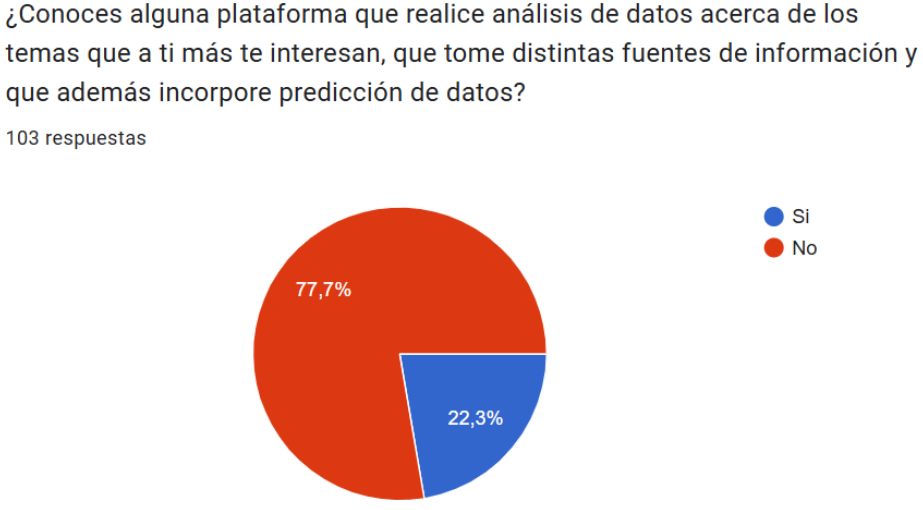


Figura 1.8: Competitividad.

Pregunta: ¿Te gustaría que existiera esta plataforma?

La pregunta final consiste en saber si a la población le gustaría tener una plataforma web con un análisis de datos poblacionales sobre distintos rubros, donde se muestren los resultados del análisis a los usuarios, y además la plataforma tenga la capacidad de incorporar predicción de datos. En la Figura 1.9 se pueden observar los resultados en los cuales, el 100 % de las personas a las

que se les aplicó la encuesta está a favor que exista una plataforma web con ese análisis de datos.

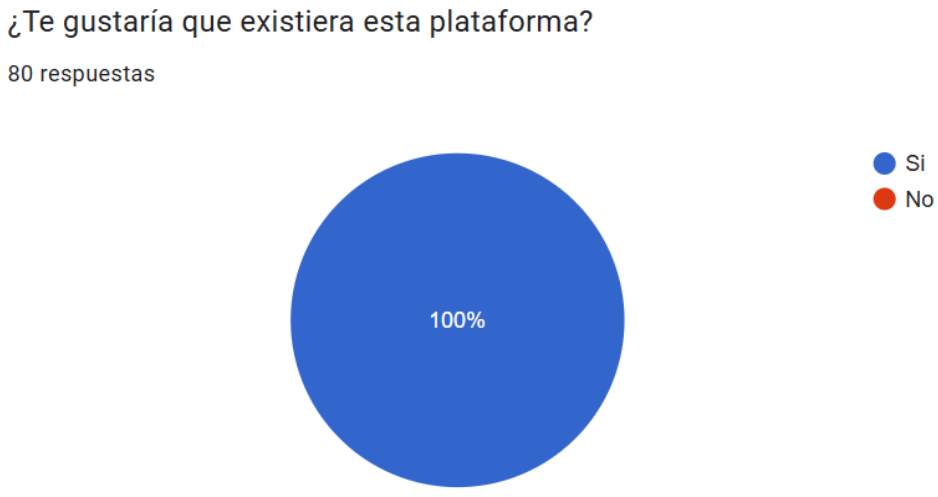


Figura 1.9: Demanda de la plataforma.

Como se puede observar, todas las personas encuestadas manifestaron que les gustaría que existiera esta plataforma.

1.4. Objetivos de la investigación

Este trabajo consta de un objetivo general y cinco específicos que ayuden a desarrollar una plataforma que presente el resultado del análisis de datos.

1.4.1. Objetivo general

Desarrollar una plataforma web interactiva que muestre los resultados del análisis de datos poblacionales en México en los temas de educación, economía, empleo, seguridad y salud, implementando técnicas de ciencia de datos para que en la plataforma se puedan consultar las predicciones en áreas de interés con base en las correlaciones encontradas entre distintas variables.

1.4.2. Objetivos específicos

1. Identificar y recabar datos poblacionales relevantes de diferentes fuentes de información.
2. Limpiar los datos recabados para su correcto análisis, esto implica la eliminación de datos atípicos.
3. Desarrollar un modelo predictivo para saber cómo se comportan los datos poblacionales en los siguientes años.
4. Evaluar y validar el modelo, a través de la validación cruzada, la comprobación con otros modelos y la evaluación de la precisión.
5. Crear un sistema web que integre los resultados del análisis de una manera intuitiva para que los usuarios puedan consultar datos relevantes y realizar predicciones en sus áreas de interés.

1.5. Pregunta de investigación

¿Es posible realizar predicciones de interés para la población mexicana, basadas en correlaciones de variables dependientes e independientes en áreas de economía, salud, empleo, seguridad y educación, mostradas en una plataforma web para la toma de decisiones?

1.6. Hipótesis

La realización de un análisis de datos sobre temas de interés para la población mexicana puede permitir identificar variables dependientes e independientes, así como sus correlaciones, las cuales podrían utilizarse para generar predicciones que aporten información útil y relevante para la sociedad.

1.7. Alcances y limitaciones

Los alcances son los siguientes:

- Recopilación de datos poblacionales de diferentes fuentes, datos en formato CSV o datos semiestructurados.
- Limpieza y estructuración de datos.
- Desarrollo y validación de un modelo predictivo.

- Creación de un sitio web capaz de integrar los resultados del análisis poblacional.

Por otra parte, las limitaciones en este proyecto son las siguientes:

- Dependencia en datos históricos.
- Falta de actualización en las fuentes de datos.
- Los conjuntos de datos existentes para ciertos temas no están estructurados.
- Conjuntos de datos con demasiados valores nulos o con datos faltantes.
- Acceso únicamente a repositorios de datos abiertos.

1.8. Organización del documento

Este documento está organizado en cinco capítulos. El Capítulo 2 presenta el marco teórico en el análisis de datos, haciendo una revisión de distintas técnicas utilizadas para ciencia de datos. De igual forma, se presenta una revisión de distintos trabajos relacionados, así como de metodologías y tipos de variables. El Capítulo 3 presenta la recopilación, análisis y visualización de datos. De igual forma, este capítulo presenta el diseño de la plataforma DATAMEX, un sitio web donde se presentan los resultados del análisis de datos. En el Capítulo 4, se muestran las correlaciones y los resultados de las predicciones para distintas variables poblacionales. Además, se presentan las interfaces desarrolladas de la plataforma y los resultados de las pruebas de aceptación de los usuarios. Finalmente, en el Capítulo 5 se presentan las conclusiones y el trabajo a futuro.

Capítulo 2

Marco teórico en el análisis de datos

2.1. Técnicas de ciencia de datos

En esta sección se describen diversas técnicas utilizadas en ciencia de datos, entre las que se encuentran regresión lineal, análisis bayesiano, árboles de decisión y máquinas de vectores de soporte, ya que estas técnicas permiten generar predicciones con base en el análisis de datos.

2.1.1. Regresión Lineal

La regresión lineal se puede definir como una técnica paramétrica principalmente utilizada para predecir variables continuas, dependientes dando un conjunto de variables independientes, es decir, la finalidad de esta técnica es predecir la variable dependiente y en función de los valores que emplean las variables independientes x. Regresión lineal es usada principalmente en los casos donde se quiera predecir alguna cantidad continua, por ejemplo, se puede utilizar para predecir el tráfico en una tienda minorista, al igual que predecir el tiempo de permanencia de un usuario. Se dice que es paramétrica ya que hace algunas suposiciones que se basan en el conjunto de datos. En el caso que el conjunto de datos sigue esas suposiciones, los resultados que muestre la regresión serán resultados concretos, en el peor de los casos tiene dificultades para proporcionar una precisión adecuada [1].

Variables independientes y dependientes. Dentro de la regresión lineal se utilizan dos tipos de variables, las cuales son variables dependientes e indepen-

dientes.
Variable dependiente (y): Variable que se desea predecir.
Variable independiente (x): Variable que se está utilizando para predecir el valor de la otra variable, es decir, de la variable dependiente (y) [29, 61].

Representación matemática: La regresión lineal se divide en simple o múltiple, la regresión lineal simple se refiere a cuando solamente existe una sola variable independiente para la predicción, por otro lado, en la regresión lineal múltiple se emplean múltiples variables independientes. La regresión lineal simple, como su nombre lo indica, utiliza una función lineal para predecir la variable dependiente representada de la siguiente manera:

$$y = mx + b \tag{2.1}$$

Donde:
y es la variable dependiente.
x es la variable independiente.
m es la pendiente de la línea.
b es la intersección con el eje y.

La Figura 2.1 muestra un ejemplo de regresión lineal simple, donde se muestra cómo predecir el número de accidentes de tráfico mortales en un estado, representado como la variable de respuesta (y), en función de la población del estado, que actúa como la variable predictora (x). Dicho de otra forma, el número de accidentes es la variable dependiente y la población del estado es la variable independiente.

Figura 2.1: Ejemplo de Regresión Lineal Simple [1].

En la regresión lineal múltiple la gran diferencia es que se pueden incluir n variables independientes con su respectivo coeficiente (a), por lo que en esta regresión se involucran múltiples coeficientes.

Cabe destacar que la regresión lineal múltiple es la que se utiliza más que la regresión lineal simple debido a que para realizar un análisis normalmente se cuenta con múltiples variables independientes [2]. A continuación, en la Figura 2.2 se muestra una gráfica que ejemplifica una regresión lineal múltiple utilizada para predecir las millas por galón (MPG) de varios automóviles como variable dependiente (y), tomando como variables independientes el peso y la potencia (x).

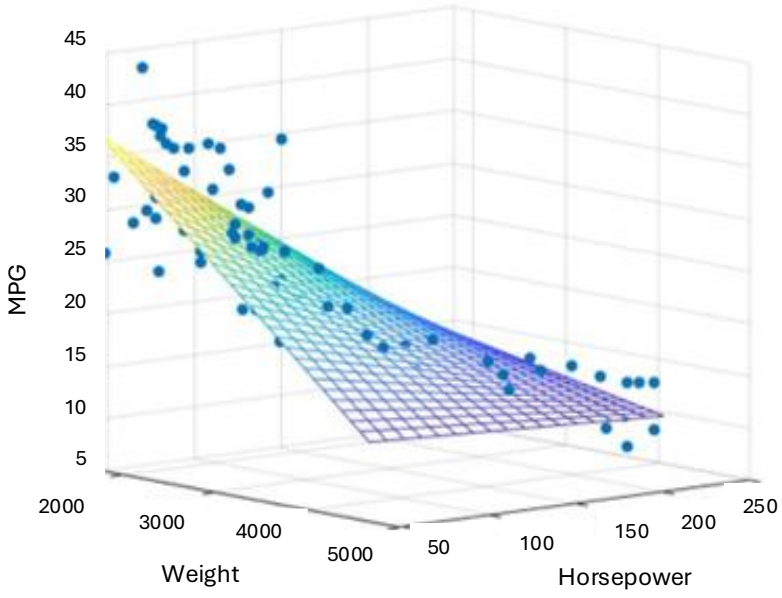


Figura 2.2: Ejemplo de Regresión Lineal Múltiple [2].

2.1.2. Árboles de Decisión

Un árbol de decisión se puede definir como un algoritmo supervisado no paramétrico que se utiliza para dos tipos de tareas, de clasificación y de regresión. La estructura que posee es de tipo jerárquica, que se forma de un nodo raíz, ramas, nodos internos y nodos hoja. En el siguiente diagrama se puede observar un árbol de decisión que consta de un nodo raíz al principio, las ramas salientes a los extremos del nodo raíz dan origen a los nodos internos y por último los nodos hoja representan todos aquellos resultados posibles que se pueden generar dentro del conjunto de datos [3]. Un ejemplo de un arból de decisión se muestra en la Figura 2.3

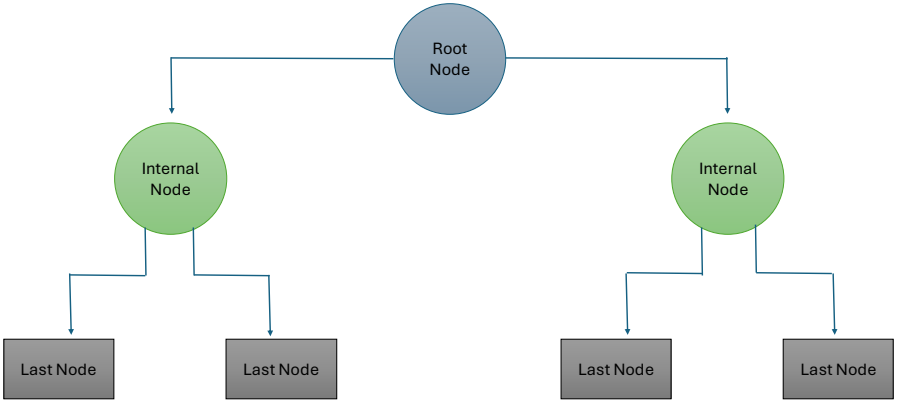


Figura 2.3: Árbol de decisión [3].

Árbol de decisión y sus terminologias

- Nodo raíz: Representa a toda la muestra.
- División: Proceso de división de uno o más nodos denominados subnodos.
- Nodo de decisión: Se crea cuando un subnodo se divide en subnodos adicionales.
- Nodo de hoja: También llamados nodos terminales y son los nodos sin hijos.
- Poda: Proceso cuando se disminuye el tamaño de los árboles de decisión eliminando nodos.
- Rama: Es una subsección del árbol de decisión, conocida también como subárbol.
- Nodo padre e hijo: El nodo principal es un solo nodo que se divide en subnodos, mientras que los hijos son subnodos del nodo principal.

La Figura 2.4 muestra la terminología utilizada en un árbol de decisión, donde se pueden identificar los Leaf Node (Nodos Hoja) y los Decision Node (Nodos de Decisión), asi como el Sub-árbol (Sub-Tree).

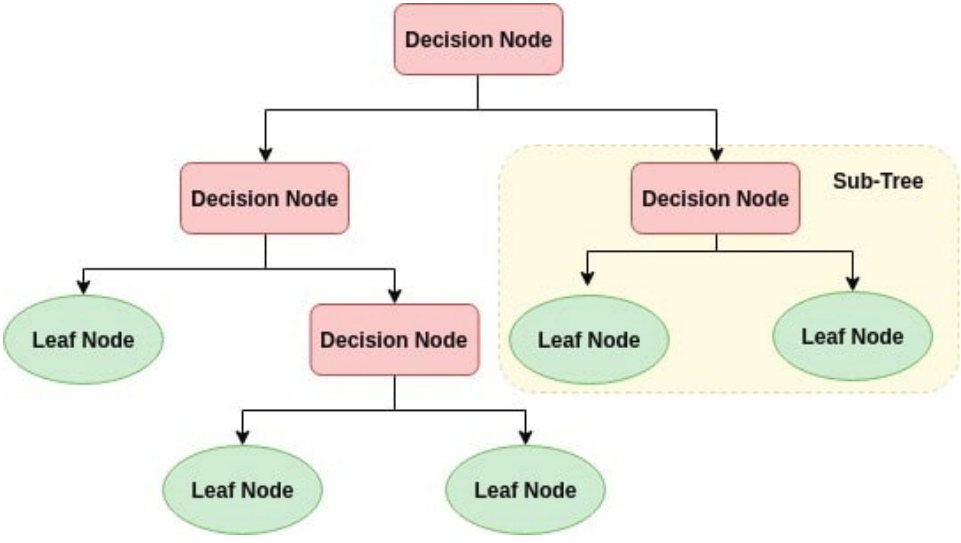


Figura 2.4: Árbol de decisión, terminología [4].

Un árbol de decisión tiene una estructura muy parecida a la de un diagrama de flujo, en donde un nodo interno hace referencia a una característica o atributo [4]. Por otra parte, la rama representa una regla de decisión y cada hoja se refiere al resultado, tal como se puede observar en la Figura 2.5 en donde se presentan varias condiciones con las opciones de Si y No y donde se observan distintos tipos de nodos (raíz, de pruebas y de decisión).

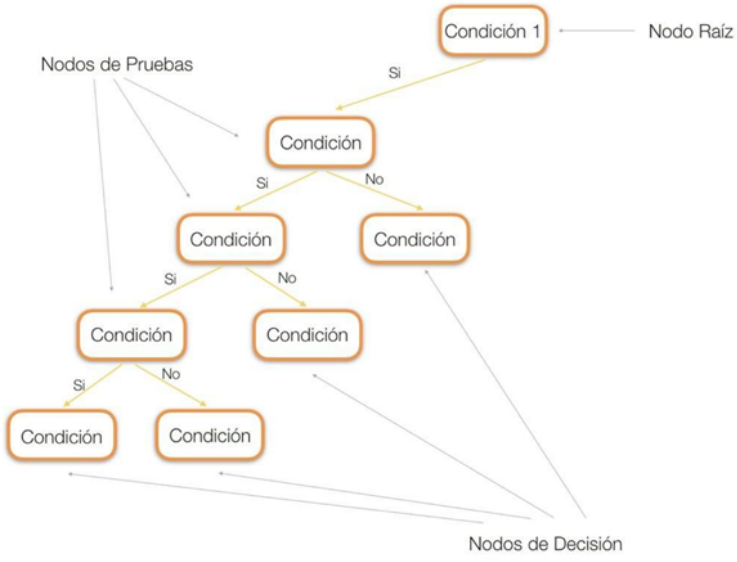


Figura 2.5: Árbol de decisión [4].

La diferencia con los modelos lineales, es que los árboles de decisión mapean de manera correcta las relaciones no lineales, además se adaptan para resolver problemas de clasificación o regresión.

Problemas de tipo regresión. Este tipo de problemas son aquellos en los que se intenta predecir los valores de una variable continua a partir de una o más variables predictoras categóricas.

Problemas de tipo clasificación. Por otro lado, en los problemas de clasificación se intenta predecir los valores de una variable dependiente categórica a partir de una o más variables predictoras continuas.

Medidas de Selección de atributos. La medida de selección de atributos se entiende como una heurística para seleccionar el criterio de división que divide los datos de la forma que mejor se acople, o también la medida de selección es llamada como reglas de partición debido a que facilita a encontrar puntos de ruptura para un conjunto en un nodo dado. Entre las medidas de selección que más sobresalen se encuentran la ganancia de información, la relación de ganancia y el índice de Gini [30].

Ganancia de información. La ganancia de información es una medida que surge cuando la entropía cambia, y esto se da al momento de utilizar un nodo en un árbol de decisión con la finalidad de particionar las instancias de for-

mación de pequeños subconjuntos. La entropia se define como la medida de la incertidumbre de una variable de tipo aleatoria. Cuando mayor sea la entropia, mayor será el contenido de la información.

Existen una serie de pasos a tener en cuenta para construir un árbol de decisión con la medida de selección ganancia de información.

- Empezar con cada una de las instancias de formación en conjunto al nodo raíz.
- Usar la ganancia de información y así seleccionar qué atributo estará con cada nodo.
- Construir de manera recursiva cada uno de los subárbol en el subconjunto de instancias de capacitación que a su vez se clasificarían en el camino del árbol.

Indice Gini. Por otra parte, el índice de Gini se define como una métrica que sirve para medir la frecuencia con la que un elemento seleccionando aleatoriamente sería identificado incorrectamente. Esto quiere decir que se debe seleccionar un atributo con un índice más bajo [30].

Ventajas

- Se pueden capturar de manera más sencilla patrones no lineales.
- No es tan complicado interpretar y visualizar los árboles de decisión.
- Los árboles de decisión al ser de naturaleza no paramétrica no cuentan con suposiciones sobre la distribución.
- Se requiere menos preprocesamiento de datos lo que para el usuario es de gran ayuda debido a que no es necesario normalizar las columnas.

Desventajas

- A causa de una ligera variación en los datos se puede generar un árbol de decisión diferente.
- Cuenta con datos sensibles al ruido.
- Antes de crear el árbol de decisión es recomendable equilibrar el conjunto de datos porque los árboles de decisión están sesgados con un conjunto de desequilibrio.

2.1.3. Máquinas de Vectores de Soporte (SVM)

Una máquina de vectores de soporte se basa en un tipo de aprendizaje automático supervisado. Se define como un algoritmo que usualmente se utiliza para la solución de problemas de clasificación al igual que problemas de regresión pero su aplicación se da más en problemas de clasificación. Además, SVM (Máquina de vectores de soporte) son perfectos para conjuntos de datos pequeños con menos valores atípicos [5]. Los vectores de soporte son aquellos puntos de datos que se encuentran más cerca al hiperplano. En la Figura 2.6 se muestra una representación de ellos.

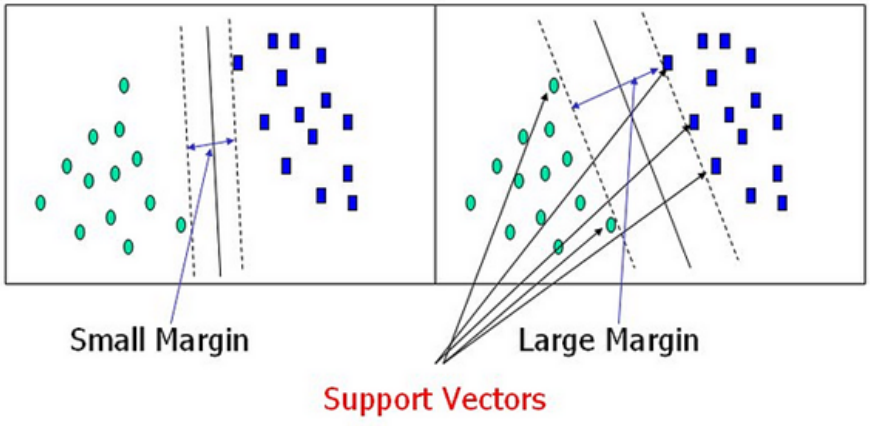


Figura 2.6: Vectores de soporte [5].

El SVM emplea un hiperplano, que se define como un subespacio de una dimensión menos que su espacio ambiental, es decir, divide el espacio en dos dimensiones. El hiperplano generado por el SVM es óptimo y de forma iterativa, busca minimizar el error. La máquina de vectores de soporte se centra en hallar un hiperplano marginal máximo que separe de la mejor manera el conjunto de datos en las diferentes clases [5]. En la Figura 2.7 se muestran los hiperplanos como superficies de decisión.

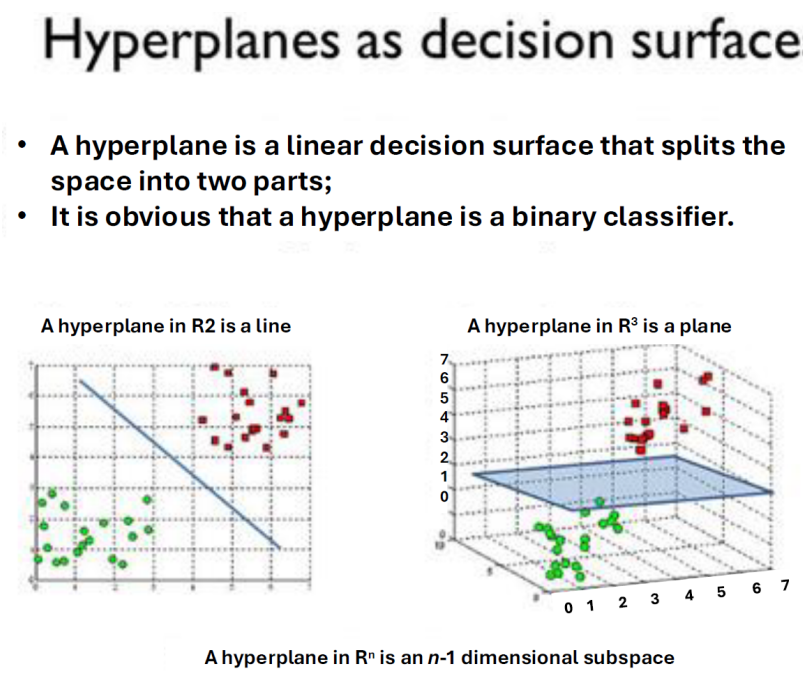


Figura 2.7: Hiperplanos como superficies de decisión [5].

En la Figura 2.8 se observan los elementos, los cuales se explican a continuación.

- Vectores de soporte: Puntos de datos que se encuentran más cercanos al hiperplano.
- Hiperplano: Plano de decisión, su función es separar un conjunto de objetos que se encuentran en clases diferentes dentro de un espacio multidimensional.
- Margen: Espacio entre las dos líneas en los puntos que están más cerca de la clase. Se considera un buen margen si este es mayor entre las clases, si el margen es menor entonces es un mal margen.

Función del algoritmo de Máquinas de Soporte de Vectores: El objetivo es emplear un hiperplano que cuente con el máximo margen posible entre los vectores de soporte dentro del conjunto de datos establecido. Una manera de hallar el hiperplano correcto es con los siguientes pasos:

Generar diferentes hiperplanos que segregan las clases lo mejor que se pueda. Por ejemplo, en la Figura 2.8 se muestra que se generaron 3 hiperplanos, pero únicamente un hiperplano está dividiendo ambas clases de tal manera que la clasificación sea correcta.

Figura 2.8: El mejor hiperplano [6].

Seleccionar el hiperplano adecuado, con la máxima segregación de los puntos de datos. Una manera de resolver esto es con un hiperplano de tipo lineal, pero no funciona en todos los casos. En la Figura 2.9 se observa cómo el algoritmo usa el núcleo para transformar el espacio de entrada en uno dimensional superior.

Figura 2.9: Hiperplano antes de ser lineal [6].

Los puntos de datos se grafican en los ejes X y Z, donde Z es la suma cuadrada de los ejes X y Y. Ahora es posible separar de una manera más sencilla los puntos utilizando la separación lineal, tal como se observa en la Figura 2.10.

Figura 2.10: Hiperplano con la separación lineal [6].

Lo aplicado anteriormente se le conoce como Kernel. Un kernel cambia un espacio de datos de entrada en la forma que se requiera, es decir, el truco kernel permite ampliar los datos en un sinfín de dimensiones [6].

Ventajas: La clasificación con Máquinas de Vectores de Soporte posee una precisión muy buena en comparación con Regresión Lineal y Árboles de Decisión que igual pueden servir como clasificadores. A causa de utilizar un subconjunto de puntos de entrenamiento en la etapa de decisión no ocupa mucha memoria.

Desventajas: Las SVM no son muy aptas para manejar grandes cantidades de datos a causa del tiempo que demora en su formación. Además, su funcionamiento falla con las clases superpuestas y es sensible al tipo de núcleo empleado.

2.1.4. Análisis Bayesiano

La inferencia estadística que regularmente se venía utilizando en los últimos años se encuentra en un momento de declive, debido a otras maneras de inferir que se están poniendo en práctica, como lo son los métodos bayesianos. La inferencia bayesiana constituye otro panorama para el análisis estadístico de datos que tiene mucha relación con los métodos convencionales de inferencia [31].

El análisis bayesiano, fue desarrollado a partir de una fórmula matemática por Thomas Bayes durante el siglo XVIII. Su enfoque en sistemas expertos, una rama de la inteligencia artificial en la década de los 80s, contribuyó a grandes avances, y en la actualidad esta técnica de ciencia de datos está por todas partes, de hecho, es implementada en los teléfonos celulares de hoy [32].

Teorema de Bayes: Dada una hipótesis sobre una población, la inferencia Bayesiana lo que hace es actualizarla una vez que se hayan presentado datos mediante:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.2}$$

Donde:
$P(A|B)$: Probabilidad posterior de A dado B.
$P(B|A)$: Verosimilitud de B dado A.
$P(A)$: Probabilidad previa de A.
$P(B)$: Probabilidad marginal de B (evidencia).

Para aplicar este teorema antes que todo es necesario formular el problema en término de una o más hipótesis. Por poner un ejemplo respecto al área de salud, un paciente con COVID-19, en el caso de seleccionar a alguien de manera

aleatoria se desconoce el diagnóstico final, sin embargo, los expertos tendrían una idea inicial sobre el diagnóstico, esto es llamado probabilidad previa y se refiere a la medida de la probabilidad de que un suceso pueda ocurrir. La probabilidad se encarga de medir que tan posible es que se presente un suceso, empleando un valor numérico ya sea 0 o 1. Para obtener una mejor estimación de cualquier probabilidad previa es recomendable consultarlo con un experto al igual que de promedios estadísticos [32].

Lo siguiente es utilizar una evidencia para actualizar la probabilidad previa de la hipótesis. Generalmente, la evidencia empleada se representa con la letra E, y con la expresión P(E). Bayes notó que a partir de la probabilidad previa de una hipótesis e incluyendo las evidencias de lo observado, era posible calcular una probabilidad actualizada conocida como la probabilidad posterior [33].

Sistemas ingenuos de aprendizaje bayesiano

Una de las desventajas que presenta el análisis bayesiano es que cuando en sistemas que se manejan grandes dimensiones, el aumento de hipótesis y pruebas puede llevar a la presencia de demasiadas combinaciones entre la cantidad de hipótesis y de pruebas. La implementación del aprendizaje automático respecto a los datos es posible que evite gran cantidad de codificación manual. Un sistema ingenuo de aprendizaje automático también se puede denominar como una red neuronal de clasificación. Utiliza ejemplos de aprendizaje supervisado, es decir, ejemplos de datos en donde los resultados constituyen al algoritmo de aprendizaje y que funciona mucho mejor con grandes conjuntos de datos. Una vez que los datos se convierten en una tabla de frecuencia estadística, es posible que aprenda valores nuevos utilizados antes para cada clase. Posteriormente, se clasifican los predictores. En este caso se refiere a la hipótesis, de manera descendente para la predicción, incluso cuando no se satisfacen los criterios de independencia. De este modo a esta técnica se le conoce como aprendizaje ingenuo de Bayes [33].

La inferencia Bayesiana se ha posicionado como una técnica muy utilizada para el desarrollo de nuevas aplicaciones, tanto en sistemas computacionales como en sistemas derivados de un modelo de aprendizaje [33].

2.1.5. Comparación entre técnicas

En esta sección se muestra la comparación entre las técnicas de ciencia de datos descritas anteriormente.

Técnica	**Descripción**	**Ventajas**	**Desventajas**	**Casos de uso**

Regresión Lineal	Modelo estadístico que encuentra la relación lineal entre variables.	Fácil de interpretar, rápido de entrenar.	Sensible a valores atípicos, asume linealidad.	Predicción de precios, análisis de tendencias.
Árboles de Decisión	Modelo basado en reglas que divide los datos en ramas según condiciones.	Fácil interpretación, útil para datos no lineales.	Puede sobreajustar si no se poda correctamente.	Clasificación de clientes, diagnóstico médico.
Bosques Aleatorios	Conjunto de múltiples árboles de decisión para mejorar precisión.	Reduce el sobreajuste, maneja datos grandes.	Más lento y menos interpretable que un solo árbol.	Detección de fraudes, clasificación de imágenes.
Máquinas de Vectores de Soporte (SVM)	Encuentra un hiperplano óptimo para separar clases en datos.	Eficiente en espacios de alta dimensión, robusto.	Lento con grandes volúmenes de datos, difícil de interpretar.	Reconocimiento de texto, clasificación de imágenes.
Análisis Bayesiano	Modelo probabilístico basado en el Teorema de Bayes.	Funciona bien con pocos datos, maneja incertidumbre.	Sensible a la calidad de las distribuciones previas.	Filtrado de spam, diagnóstico médico.

Tabla 2.1: Descripción de técnicas.

2.2. Metodologías en ciencia de datos

En este trabajo se consideraron dos metodologías enfocadas a ciencia de datos, descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés) [34, 35, 36] y la metodología BATCH FMDS de IBM [7, 37, 38].

KDD es un proceso utilizado para extraer conocimiento útil y significativo a partir de grandes volúmenes de datos. Esta metodología es fundamental en campos como la minería de datos, la inteligencia artificial y el aprendizaje automático, y se compone de varias etapas que van desde la preparación de los datos hasta la interpretación de los resultados. El objetivo principal de KDD es identificar patrones ocultos, relaciones y tendencias dentro de los datos que puedan ser útiles para la toma de decisiones. Las etapas del proceso KDD son las siguientes.

- Etapa 1: Selección de datos. Se seleccionan los datos relevantes de la base de datos, aquellos que son importantes para el análisis. Esta fase implica definir los objetivos y criterios para seleccionar los subconjuntos de datos que se van a utilizar.
- Etapa 2: Preprocesamiento de los datos. Antes de aplicar técnicas de minería de datos, es necesario limpiar y preparar los datos. Esto incluye la corrección de errores, el manejo de valores faltantes, la normalización y la transformación de datos para que sean consistentes y adecuados para el análisis.
- Etapa 3: Transformación de los datos. Los datos se transforman o consolidan en un formato adecuado para el proceso de minería de datos. Esto puede incluir la reducción de dimensiones, la proyección de variables, o la transformación de variables en nuevas características que faciliten la extracción de patrones.
- Etapa 4: Minería de datos (Data Mining). Esta es la etapa central del proceso KDD, donde se aplican técnicas y algoritmos para descubrir patrones, relaciones, o modelos dentro de los datos. Los métodos pueden incluir algoritmos de clasificación, regresión, clustering, asociaciones, o detección de anomalías.
- Etapa 5: Evaluación de patrones. Una vez que se han identificado los patrones o modelos, es crucial evaluarlos para determinar su validez y utilidad. Esta fase implica medir la calidad del modelo, su precisión, y su relevancia para los objetivos definidos al inicio del proceso.
- Etapa 6: Interpretación y uso del conocimiento. Los resultados obtenidos se interpretan y se traducen en conocimiento accionable. Esta fase implica

presentar los hallazgos de manera que sean comprensibles y útiles para la toma de decisiones. En muchos casos, el conocimiento descubierto puede retroalimentar el proceso para refinar futuras extracciones.

Por otra parte, se presenta la descripción de cada uno de los pasos seguidos en la investigación con base en la metodología de IBM especialmente para el análisis de datos. La metodología BATCH FMDS de IBM [38] proporciona un marco completo y sistemático para el procesamiento eficiente de grandes volúmenes de datos, optimizando los recursos y mejorando la calidad de la toma de decisiones empresariales. Los pasos de esta metodología se muestran en la Figura 2.11.

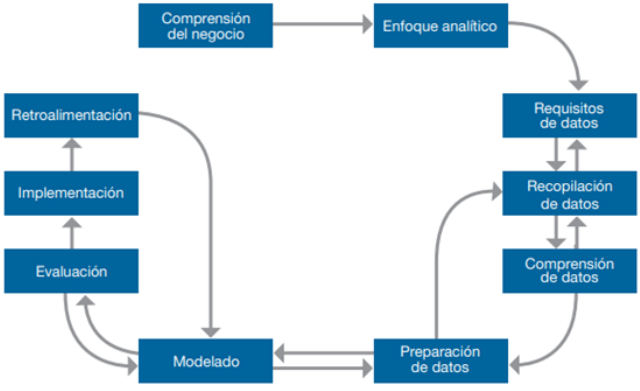


Figura 2.11: Metodología de IBM para ciencia de datos [7].

Los pasos de la metodología IBM se enlistan a continuación:

- Etapa 1: Comprensión del negocio. Para comenzar cualquier proyecto es necesario entender el negocio, debido a que el éxito para resolver el problema empresarial depende de esta primera etapa. Definir el problema, los objetivos del proyecto y los requisitos de la solución son secciones establecidas por los promotores de negocios.
- Etapa 2: Enfoque analítico. Después de definir el problema con claridad, se podrá definir el enfoque analítico por parte del científico de datos, para posteriormente dar solución al problema. En esta segunda etapa se necesita expresar el problema en otro contexto, es decir, bajo el contexto de las técnicas estadísticas y de aprendizaje automático con la finalidad

de que la organización pueda identificar las que mejor se adapten y así obtener el resultado esperado.

- Etapa 3: Requisitos de datos. El enfoque analítico elegido en la etapa anterior, determina lo que es la etapa 3, los requisitos de datos. Los métodos analíticos que se están dispuestos a utilizar requieren de contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio.

- Etapa 4: Recopilación de datos. Para realizar la etapa inicial de recopilación de datos, los científicos de datos se encargan de identificar y reunir todos los recursos de datos disponibles, ya sean datos estructurados, no estructurados y semiestructurados, cualquiera que sea relevante para el dominio del problema. Si durante la recopilación de datos se encuentran lagunas, existe la posibilidad que el científico de datos tenga que regresar a la etapa anterior y revisar los requisitos de datos para recopilar más datos o datos nuevos.

- Etapa 5: Comprensión de datos. Después de completar la recopilación inicial de datos, los científicos de datos deben analizar el contenido, evaluar la calidad de la información y explorar posibles descubrimientos preliminares. Para llevar a cabo esta tarea, suelen emplear estadísticas descriptivas y herramientas de visualización. En esta etapa se puede regresar a la etapa anterior a recopilar datos para llenar algunos huecos.

- Etapa 6: Preparación de datos. La preparación de datos suele ser la etapa más larga de un proyecto de ciencia de datos debido a las actividades que emplea. Las actividades para llevar a cabo la preparación de datos son las siguientes: La limpieza de datos, esta actividad se refiere a tratar con valores no válidos o que faltan, eliminar datos que se repitan y dar un formato correcto, en la limpieza de datos también entra la parte de normalización. Combinar datos de múltiples fuentes, como archivos, plataformas y tablas. Transformar los datos en variables más útiles.

- Etapa 7: Modelado. Para la etapa de modelado se utiliza la primera fase del conjunto de datos previamente preparado y se centra en desarrollar modelos predictivos o descriptivos, esto depende del enfoque analítico definido anteriormente. Para la construcción de los modelos predictivos, los científicos de datos emplean un conjunto de capacitación, es decir, datos históricos en los que se descubre el resultado de interés. Para encontrar el modelo que mejor se acople, los científicos de datos pueden probar diferentes algoritmos, cada uno con sus respectivos parámetros.

- Etapa 8: Evaluación. Mientras el modelo está en su fase de desarrollo y antes de que sea implementado, el científico de datos debe evaluar el modelo para verificar su calidad y asegurarse que el modelo se centra en el problema empresarial de manera correcta y completa. La etapa de evaluación del modelo implica el cálculo de diferentes medidas de diagnóstico y de resultados como tablas y gráficos, con ello el científico de datos puede interpretar el nivel de calidad y que tan capaz es el modelo en la resolución del problema.

- Etapa 9: Implementación. Una vez que el modelo ha sido desarrollado satisfactoriamente y aceptado por los promotores del negocio, procede a implementarse en la fase de producción o en una fase de pruebas comparable, En la mayoría de los casos, no se implementa completamente hasta que su rendimiento se haya evaluado totalmente.

- Etapa 10: Retroalimentación. Cuando el modelo se haya implementado se recopilan los resultados del mismo, la organización obtiene la retroalimentación sobre el funcionamiento del modelo y su impacto en el área en el que se implementó. Esta retroalimentación puede servir para que a los científicos de datos les sea posible ajustar el modelo para aumentar su nivel de precisión y eficiencia. Incluso es posible automatizar algunas o todas las secciones de la evaluación del modelo y de la recopilación de retroalimentación, el ajuste y la reimplementación del modelo con el objetivo de acelerar el proceso de actualización del modelo y así garantizar mejores resultados.

Como resultado de la revisión de ambas metodologías, se concluyen los siguientes puntos.

- Alineación con el negocio: BATCH FMDS es superior en su capacidad para asegurar que todo el proyecto esté alineado con los objetivos estratégicos de la organización, lo que aumenta la probabilidad de que el proyecto tenga un impacto real y positivo en el negocio.

- Ciclo de vida completo: A diferencia de KDD, que se enfoca principalmente en la extracción de conocimiento, BATCH FMDS cubre todo el ciclo de vida del proyecto, desde la comprensión del problema hasta el monitoreo en producción. Esto es esencial para proyectos empresariales que requieren implementación efectiva y sostenible.

- Mejora continua: La inclusión de fases de retroalimentación y monitoreo en BATCH FMDS permite una iteración constante, lo que es crucial en entornos dinámicos donde los modelos pueden degradarse con el tiempo debido a cambios en los datos o el contexto del negocio.

- Preparación para producción: BATCH FMDS está diseñado para asegurar que los modelos no solo sean precisos, sino que también sean desplegables y mantenibles en un entorno de producción, lo cual es una necesidad crítica en aplicaciones empresariales.

Como resultado, se determinó utilizar la metodología de IBM para ciencia de datos en este proyecto.

2.3. Tipos de variables

Las variables estadísticas se pueden definir como características que son propias de las personas, cosas al igual que lugares, y esas características pueden ser medibles. Ejemplos muy claros son el peso, la edad, la estatura, el sexo, la temperatura, entre otros. El estudio de las variables estadísticas tiene como objetivo entender cómo se comportan las variables de un determinado sistema para posteriormente realizar predicciones sobre la manera en la que se comportan en algún futuro. Existe una enorme cantidad de variables para estudiar, sin embargo, algunas pueden ser representadas de diferente manera, ya sea de forma numérica mientras que otras no [39].

2.3.1. Variables cualitativas

Este tipo de variables se emplean para designar cualidades o categorías. Por ejemplo, variables como el estado civil que cubre soltero, casado, divorciado o viudo son cualitativas debido a que ninguna de las categorías anteriores es superior a la otra, solamente expresa una situación diferente. Además, una categoría puede tomar el valor de un número, por ejemplo, número de casa, número de teléfono, de calle o bien de código postal, representando una etiqueta y no un valor numérico diferente. Las variables cualitativas pueden pertenecer a su vez:

- Nominales: Asignan un nombre a la categoría, ejemplo el color.
- Ordinales: Representan un orden, un ejemplo podría ser opiniones sobre alguna propuesta (a favor, indiferente, en contra).
- Binarias: Estas variables solo pueden tomar dos valores, como en el sexo, de igual manera se le puede asignar un valor númerico, como el 1 y el 2, pero sin que represente un valor númerico o algún orden [8].

2.3.2. Variables cuantitativas

Por otro lado, están las variables cuantitativas, a estas variables se les otorga un número, debido a que representan cantidades, ejemplos de variables cuantitativas serían el salario, la edad, las distancias, calificaciones en clase. A comparación de las variables cualitativas, el número de este tipo de variables si puede ser superior a los otros.

Las variables numéricas se dividen en dos categorías:

- Variables discretas: Las variables que pertenecen a esta categoría solamente pueden tomar determinados valores y caracterizan por ser contables, es decir, el número de hijos es una variable discreta ya que puede tomar ciertos valores como 0, 1, 2, 3 o más, pero nunca 3.5 número de hijos.
- Variables continuas: La diferencia con las variables anteriores es que las variables continuas pueden tomar cualquier valor numérico como lo es en el caso del peso y la estatura de las personas, la temperatura, el tiempo, la longitud, entre otros [8].

La Figura 2.12 representa un diagrama de Pareto que compara frecuencia de defecto (variable cuantitativa en el eje vertical) y el porcentaje acumulativo versus cada defecto (variable cualitativa en el eje horizontal).

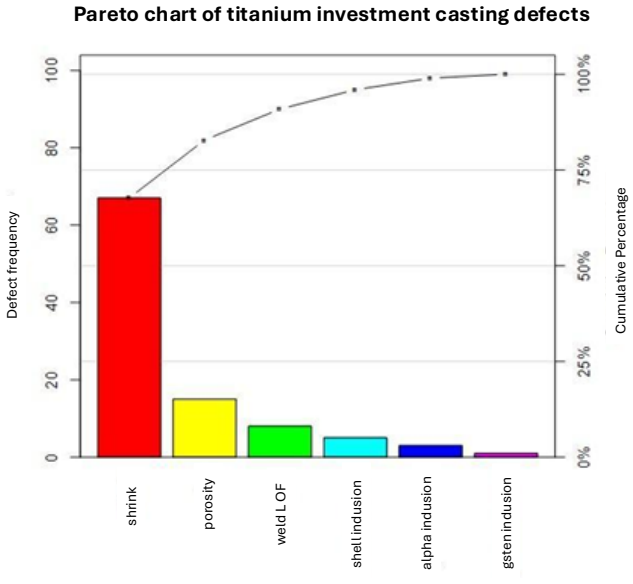


Figura 2.12: Diagrama de Pareto [8].

2.4. Estado del arte

En el artículo "Ciencia de datos educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México" [9] se aborda el tema de la deserción educativa en la actualidad es un tema de gran preocupación, ya que además de afectar a los estudiantes también lo hace con el estado y las instituciones educativas. Una manera de enfrentarse a este problema es aplicando técnicas de ciencia de datos que ayuden a identificar las posibilidades de permanencia de alumnos universitarios. La investigación se realizó con la finalidad de predecir la deserción estudiantil en el primer año empleando técnicas de aprendizaje automático y se basa en un caso de estudio práctico enfocado en la educación tomando información de estudiantes de una universidad privada.

Se aplicó la técnica XGBoost que se define como un algoritmo enfocado en aprendizaje automático basado en una amplia gama de aplicaciones como los árboles de decisión para resolver problemas de clasificación, predicción y regresión, además utiliza un marco de potenciación de gradientes. Este algoritmo permite encontrar importantes características para llevar a cabo la predicción

en la deserción escolar de un estudiante, tomando en cuenta los parámetros predeterminados de cada algoritmo en Jupyter.

A continuación, se puede observar en la Figura 2.13 el resultado obtenido del modelo entrenado en conjunto con el modelo XGBoost, llegando a la conclusión de que las variables promedio en el primer periodo, porcentaje de la beca y la región, pero principalmente la variable promedio en el primer periodo es por mucho el predictor que determina si el alumno continúa con sus estudios universitarios.

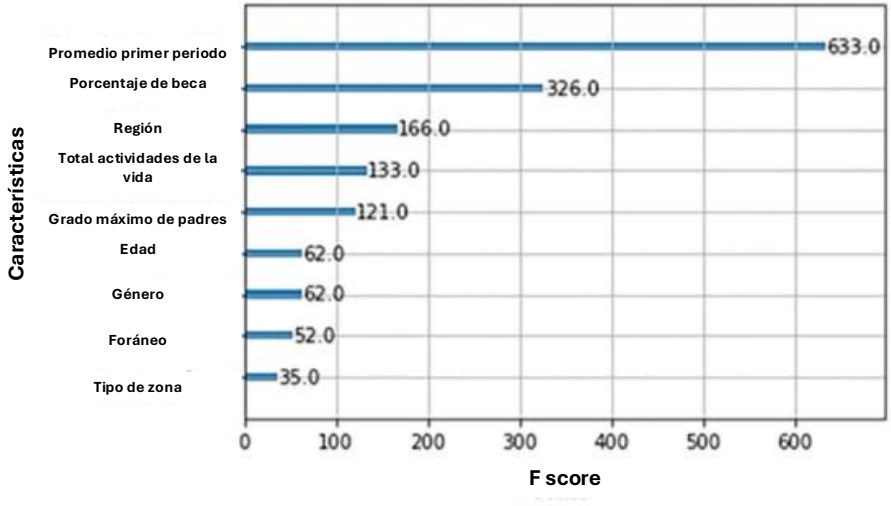


Figura 2.13: Importancia de las características [9].

Después de las pruebas con XGBoost se seleccionaron dos variables debido a que son las mejores opciones para evaluar la regresión logística, árbol de decisión y red neuronal, de igual manera se seleccionó una variable objetivo la cual determina si el estudiante se retiene o se da de baja. La aplicación de las tres técnicas mencionadas anteriormente consta de una serie de pasos entre los cuales está la prueba y el entrenamiento de datos, la creación del algoritmo, la predicción y la evaluación del mismo. Para llevar a cabo la evaluación del algoritmo se necesitan las métricas más usadas, como lo son puntuación, precisión, recuperación y matriz de confusión, la cual se puede observar en la Figura 2.14. La matriz de confusión es una tabla que describe los resultados de un modelo de predicción, la matriz cuenta con dos tipos de valores, los reales y los predichos, de esta forma es posible realizar los cálculos necesarios para

la puntuación de precisión del modelo, en este caso la exactitud del modelo XGBoost es de 0.9928.

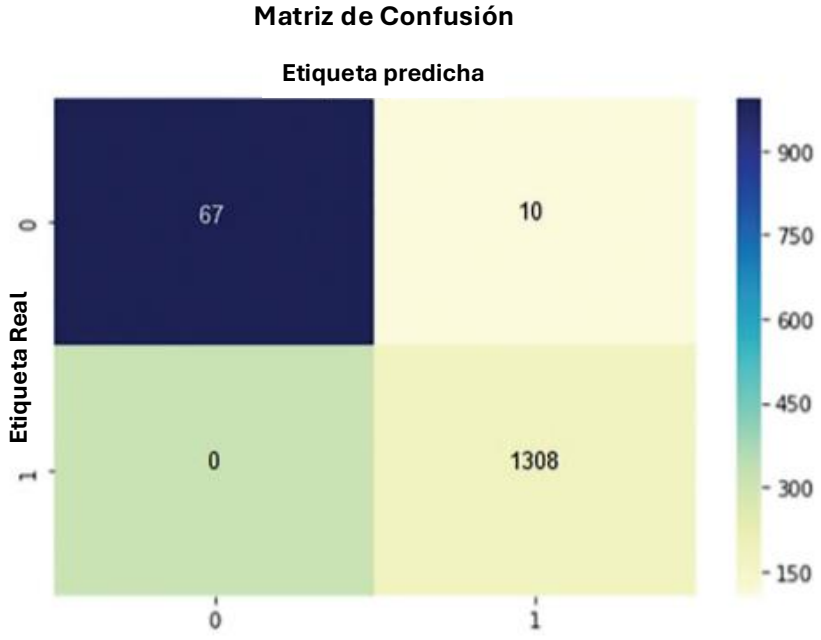


Figura 2.14: Matriz de confusión [9].

Finalmente, los resultados arrojan nueve principales categorías, de las cuales tres sobresalen a causa de su ponderación, promedio estudiantil en el primer periodo, porcentaje de la beca otorgada y región. La exactitud de la técnica XGBoost supera el 99 % mientras que las otras técnicas que se aplicaron con el objetivo de validar los resultados, la exactitud está por debajo del 94 %, a pesar de que la información se depuró y así obtener una base de datos más normalizada.

Por otra parte, en el artículo "Indice de factores que inciden en el desarrollo de las zonas metropolitanas en México" [40] los autores plantean que conforme un país va en crecimiento, se ven influenciados varios factores, especialmente en las ciudades grandes, que están apegados en la concentración de las economías, es decir, este cambio da origen a diferencias salariales y por consecuente costos de vida superiores en las zonas metropolitanas del país.

En México existe una alta notoriedad en sus ciudades grandes, debido a

que, en lugares como la capital, se conoce como una zona de concentración económica, productiva, política y social. Las zonas metropolitanas se conocen por la integración de: servir como punto de enlace con otras regiones del país que se relacionan de manera interdependiente debido al intercambio de bienes y servicios, contar con un número de habitantes bastante alto, concentración de actividades económicas. La industria juega un papel muy importante en el desarrollo de las metrópolis, por el hecho de que para contar con servicios de bienestar básicos para las personas es necesario la construcción de carreteras, puentes, vías ferroviarias, puertos, viviendas, escuelas, hospitales, cines, hoteles, parques, entre otros. La industria de la construcción para poder materializar todos esos servicios necesita de otras industrias que proporcionan la materia prima como lo es el hierro, el acero, la cal, el aluminio, la madera y otros recursos, y con ello la industria se posiciona como un pilar de la economía para el desarrollo de zonas metropolitanas.

Para el análisis de los datos se implementó la técnica de análisis factorial (AF) ya que a diferencia del modelo de regresión lineal esta técnica expresa las variables en términos de factores comunes específicos no observados, mientras que el modelo de regresión lineal no establece relaciones de causalidad entre las variables previamente observadas [41]. En la parte de resultados, el análisis factorial agrupa los valores en seis componentes que más destacan como se visualiza en la Tabla 2.2, y para cada uno de esos seis valores se obtiene un valor ponderado, los valores permiten analizar cada componente por la zona metropolitana que se evaluó anteriormente, para posteriormente el análisis factorial muestra las variables que cuenten con una alta variabilidad en sus resultados.

Componente	**Sumas de extracción de cargas al cuadrado, porcentaje de varianza**	**Sumas de rotación de cargas al cuadrado, porcentaje acumulado**	**Ponderación**
1	26.779	26.779	0.3822
2	13.513	40.292	0.1929
3	9.553	49.845	0.1363
4	8.88	58.725	0.1267

5	5.772	64.497	0.0824
6	5.568	70.065	0.0795

Tabla 2.2: Varianza total explicada.

En el análisis estadístico se demostró que si el nivel de bienestar es mayor, la prosperidad aumenta, asi mismo que existen aspectos negativos que perjudican directamente a la población como lo son la inseguridad, los homicidios, los secuestros, entre otros.

De igual forma, en el artículo "La probabilidad del crimen y su relación con su crecimiento económico en México: un análisis regional" [10] los autores establecen que en los estados de la república mexicana el crimen tiene una relación muy notable con el desarrollo económico. El análisis de esa correlación es importante debido a que entre más presencia criminal la inversión productiva se ve afectada negativamente. En esta investigación se estima la probabilidad del índice delictivo con base a modelos de regresión de respuesta cualitativa, posteriormente, con métodos de panel, se calcula el impacto del crimen en el desarrollo económico. El crimen además de afectar la actividad económica tiene repercusiones negativas sobre la vida social de la población, su aumento perjudica la economía a causa de las diferentes formas que toma el crimen como lo son el robo, la extorsión, el asesinato, el secuestro, entre otros. Todo esto en conjunto genera un ambiente de temor y angustia que desalienta el desarrollo económico.

En los últimos años el crimen ha crecido de una manera que se cree que no es posible su reducción, al mismo tiempo que la productividad económica va en declive, entre el año 2007 y el año 2012 en la república mexicana la tasa de homicidios aumentó a un poco más del 22 % anual, por otro lado, el Productor interno bruto precipita del país lo hizo a tasas demasiadas bajas, al registrar aumentos del 0.23 % anual en promedio. Notablemente, la relación crimen-crecimiento no ha sido en los últimos años de manera homogénea entre las diferentes regiones de México. El análisis estadístico indica que los estados pertenecientes a la frontera norte, golfo y frontera sur han sido perjudicados por las consecuencias del crimen y el bajo desempeño económico con mayor severidad. Dado esto, la investigación parte de esta observación la cual se busca probar, la hipótesis es que la oleada de violencia que enfrenta el país afecta negativamente a la economía y su crecimiento, además de que no es el mismo efecto sobre las regiones, teniendo un mayor impacto en la frontera norte y

sur y golfo en comparación con las zonas centro y occidente. La relación entre la tasa de homiciodios y producto interno bruto per capitá se muestra en la Figura 2.15.

Figura 2.15: Tasa de homicidios y producto interno bruto en México 1997-2012 [10].

En el periodo global el indicador del PIB cambió a una tasa del 1.54 % anual, mientras que el periodo del año 2007 y 2012 su tasa anual disminuyó al 0.23 % [10]. Por otro lado, la tasa de homicidios se mantuvo entre 1997 y 2006 pero comenzó a aumentar desde el año 2007, sus tasas se reportaron con un valor del 22.9 % anual. Y al mismo tiempo que los homicidios aumentaron a tal magnitud en la producción económica redujo notablemente, tal como se muestra en la Figura 2.16.

	1997-2012	1997-2000	2001-2006	2007-2012
Tasa de homicidios	0.80	-8.35	-5.16	22.89
Tasa de crecimiento del PIB	3.54	5.32	2.95	2.17
Tasa de crecimiento del PIB per cápita	1.54	3.97	1.55	0.23

Figura 2.16: Tasa de homicidios y tasa de crecimiento económico en México 1997-2012 [10].

Los resultados obtenidos tienden a corresponder con la hipótesis empleada. Mayores niveles de escolaridad reducen la presencia criminal, lo cual frena el

desempeño económico. Este efecto es más notable en los estados que conforman la frontera norte, la frontera sur y parte del golfo, mientras que en los estados que se encuentran en la zona centro y occidente no hay un impacto tan negativo. La probabilidad de incidentes criminales es más elevada en los estados del sur y golfo en comparación a los del occidente y centro. Los estados del sur y golfo cuentan con indicadores económicos y sociales muy bajos por lo que el nivel de educación también lo es y por consecuente la violencia tiene más presencia en dichas regiones.

Similarmente, en el trabajo denominado "Aplicación de Ciencia de Datos para el análisis de datos de mortalidad por COVID-19 de México" [7] los autores abordan la aplicación de ciencia de datos en diferentes áreas del conocimiento, como lo es en la epidemiología. En esta investigación fue necesario la aplicación de la metodología BATCH FMDS de IBM, debido a que es una metodología de Ciencia de Datos orientada al dominio epidemiológico y la cual garantiza resultados favorables, en conjunto con un caso práctico que cuente con datos reales y la intervención de expertos en el tema abordado. La metodología fundamental para la Ciencia de Datos se basa en 10 etapas, tal como se muestra en la Figura 2.17.

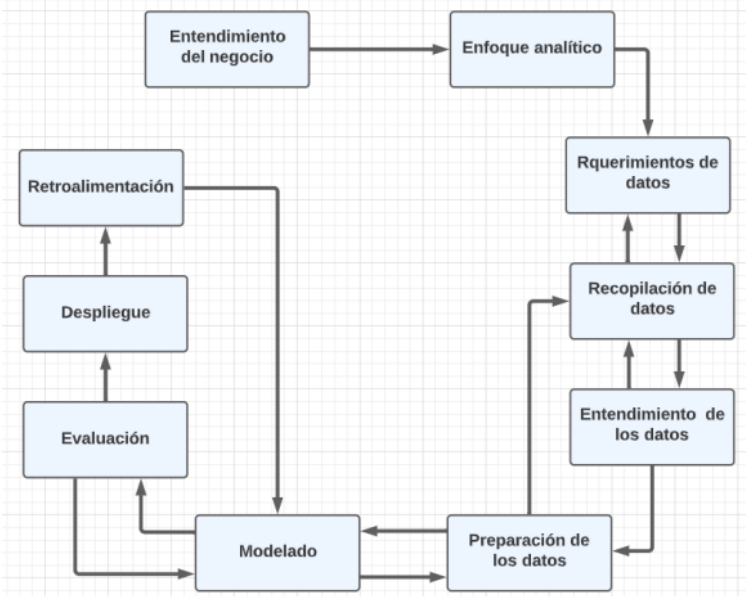


Figura 2.17: Metodología Funcional para la Ciencia de Datos [7].

El caso práctico se basa en el análisis de datos referentes a la mortalidad por COVID-19 en México, se busca dar respuesta a la siguiente pregunta de investigación ¿Cuáles factores sociodemográficos tienen en común los municipios con tasas de mortalidad por COVID-19 similares? La sección de caso práctico está conformada por diferentes etapas, como lo son las etapas de: entendimiento del negocio, enfoque analítico, requerimiento de datos, recopilación de datos, entendimiento de los datos, preparación de los datos, modelado, evaluación, despliegue, retroalimentación. Una de las etapas que más destaca en esta investigación es la etapa de preparación de los datos, la cual se divide en los siguientes puntos. Cabe destacar que los datos de entrada utilizados fueron datos poblacionales recopilados de instituciones oficiales de México.

- Selección, limpieza y transformación de las bases de datos.
- Datos población INEGI.
- Datos catálogo de enfermedades CEMECE.
- Datos de mortalidad DGIS.
- Datos de coordenadas geográficas AGEE.
- Datos de información municipal SNIM.
- Datos de desarrollo social CONEVAL.
- Datos de tasa de mortalidad.
- Datos de densidad poblacional.
- Normalización de datos.
- Almacén de datos.

Una vez concluida la etapa de preparación de los datos se comienza con la de modelado, esta etapa del caso práctico está basada en desarrollar modelos predictivos o descriptivos, esto va a depender del enfoque analítico definido anteriormente. Para desarrollar un modelo que cumpla con los requerido es posible que se tengan que implementar múltiples algoritmos con el fin de probar cada uno de ellos y determinar cuál algoritmo se desempeña mejor de acuerdo a los parámetros. Algunos algoritmos utilizados fueron K-medoids, Fuzzy C-means y K-means, el algoritmo de agrupación K-means se encuentra entre los más utilizados debido a la facilidad que proporciona para interpretar los resultados, por eso mismo para el llevar a cabo el modelado se aplicó una variante hibrida del algoritmo K-means, conocida como OK-means ++, por el

motivo que en pruebas computacionales está por encima de algoritmos estándar en donde el número de iteraciones es mayor, es decir, el tiempo computacional.

Para probar que se ha creado un buen modelo es necesario implementar la etapa de evaluación del modelo la cual consta de la creación de tablas y gráficos los cuales permiten medir su calidad e interpretar su eficiencia en una problemática.

A continuación, se presentan los primeros resultados que se generaron a partir del análisis de grupos, dos atributos sobresalen, densidad poblacional y el porcentaje de personas en situación de pobreza fueron de gran importancia para obtener grupos cuyos elementos tuvieran algunos valores parecidos de tasa de mortalidad por COVID-19.

Se generó el gráfico presentado en la Figura 2.18 para visualizar la distribución de los municipios con base en elporcentaje de pobreza y la densidad poblacional, los municipios están representados por puntos. Los valores de ambos atributos están normalizados en un rango del 0 al 1. En el gráfico se observa que los valores pertenecientes a la densidad poblacional son bajos, por otro lado, los puntos de pobreza están más dispersos.

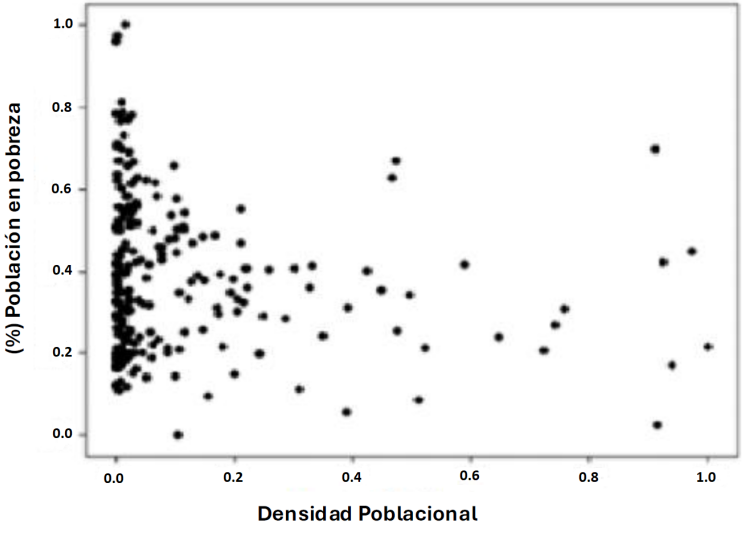


Figura 2.18: Distribución de los municipios [7].

En la Figura 2.19 se muestran los resultados obtenidos del agrupamiento de 233 municipios que se dividen en 16 grupos debido a que 16 fue la cantidad

de grupos con la que se mostraba una mejor distribución, para las pruebas se seleccionaron los grupos (6, 12, 16 y 18). Las tres primeras filas corresponden a la mortalidad más alta, y las tres últimas a la mortalidad más baja.

En la parte de las columnas se encuentran el identificador del grupo, la segunda y la tercera columna los centroides de los grupos, que tienen como atributos la densidad poblacional y el porcentaje de personas en estado de pobreza. La siguiente columna pertenece al número de municipios en cada grupo.

Grupo	Promedio de densidad poblacional	Promedio % de población en pobreza	Número de municipios	Tasa de mortalidad promedio por grupo
0	0.9138	0.0264	3	0.7970
12	0.7223	0.2059	6	0.5524
7	0.1995	0.1509	7	0.2471
9	0.1947	0.3491	21	0.2463
8	0.4734	0.6696	2	0.2420
3	0.4250	0.4042	8	0.2365
11	0.9717	0.4515	2	0.2103
14	0.0345	0.1630	25	0.2037
13	0.1393	0.3910	18	0.1911
1	0.0399	0.2401	30	0.1889
15	0.0032	0.3271	30	0.1571
5	0.0059	0.4185	25	0.1437
6	0.0270	0.6145	33	0.1316
2	0.9108	0.6982	1	0.0714
10	0.0089	0.7676	19	0.0579
4	0.0009	0.9582	3	0.0080

Figura 2.19: Resultado de la agrupación [7].

La Figura 2.20 muestra la distribución de los centroides del grupo y los municipios más próximos a los centroides los cuales están representados por cruces, y los municipios están representados por puntos.

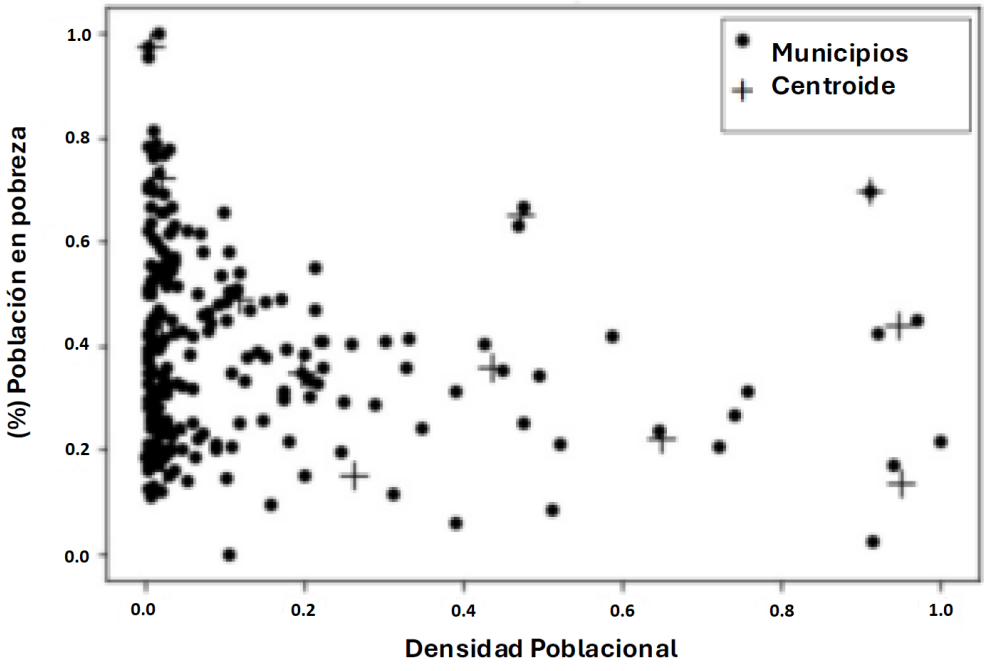


Figura 2.20: Distribución de los centroides en los grupos [7].

Para una mejor visualización de los centroides y de la distribución de los municipios por grupos externos se generó un gráfico presentado en la Figura 2.21, donde cada color pertenece a un grupo, tal como se muestra en la parte superior derecha de la figura.

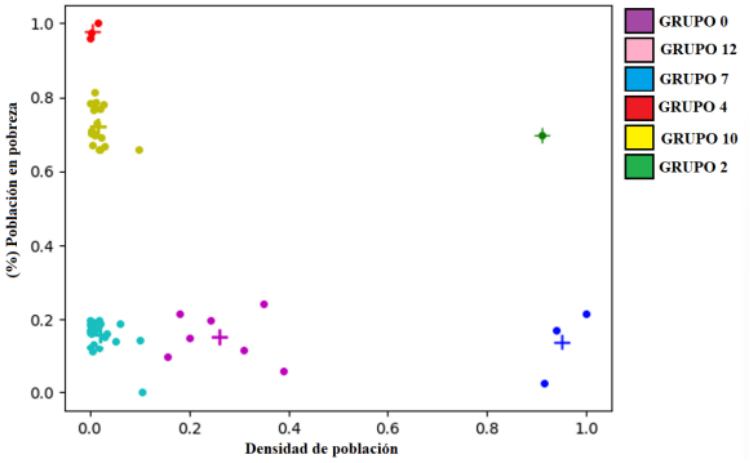


Figura 2.21: Distribución de los municipios de los grupos externos [7].

Una vez que se haya seleccionado el mejor modelo, se proyectaron los grupos con sus municipios, con su mayor y su menor tasa de mortalidad por COVID-19, tal como se indica en la Figura 2.22 y Figura 2.23.

Municipios / Alcaldías	Densidad poblacional	% de población en pobreza	Mortalidad promedio por municipio
Benito Juárez	16079.74	7.90	935.845
Iztacalco	17595.43	25.20	696.327
Cuauhtémoc	16541.94	20.90	634.933
Azcapotzalco	12711.91	24.20	915.769
Miguel Hidalgo	9010.22	13.50	858.687
Coyoacán	11378.65	27.10	471.318
Gustavo A. Madero	13333.53	33.80	410.278
Guadalajara	9176.35	24.80	392.962
Venustiano Carranza	13050.12	30.00	95.334
San Nicolás de los Garza	6869.98	10.80	575.935
Ciudad Madero	4290.27	23.40	393.332
Monterrey	3516.90	19.20	379.530
Guadalupe	5450.36	15.80	151.444
Apodaca	2746.71	14.20	113.030
General Escobedo	3186.84	25.00	20.157
San Pedro Tlaquepaque	6135.06	27.40	11.206

Figura 2.22: Municipios con mayor tasa de mortalidad COVID-19 [7].

Municipios / Alcaldías	Densidad poblacional	% de población en pobreza	Mortalidad promedio por municipio
Chimalhuacán	16027.11	68.90	68.634
Huejutla de Reyes	321.78	65.40	131.723
Comitán de Domínguez	169.92	68.80	125.769
Taxco de Alarcón	162.19	75.00	113.651
Ixtlahuaca	476.60	76.40	105.533
San Felipe del Progreso	392.75	75.40	84.872
Macuspana	65.29	69.30	76.923
San Martín Texmelucan	1730.42	65.30	62.926
Chilapa de Álvarez	164.96	75.20	54.154
San Andrés Tuxtla	169.73	79.30	48.637
Huauchinango	414.13	68.40	47.140
San Cristóbal de las Casas	547.90	66.10	45.397
Palenque	45.80	69.90	38.559
Centla	40.00	76.80	38.058
Villaflores	57.62	69.50	21.911
Almoloya de Juárez	1269.55	26.60	19.475
San José del Rincón	205.09	77.00	18.984
Papantla	109.83	69.70	11.882
Hidalgo	109.98	66.30	8.750
Villa Victoria	255.18	71.90	6.470

Figura 2.23: Municipios con menor tasa de mortalidad COVID-19 [7].

En cada fase del análisis se involucraron expertos en la validación y en la interpretación de los resultados obtenidos del análisis. Desde los resultados del caso práctico y las observaciones de la parte epidemiológica se identificaron que los valores respecto a los indicadores de densidad poblacional y porcentaje de personas en situación de pobreza contaban con una notable correlación con los valores de la tasa de mortalidad a causa de COVID-19. Por otro lado, desde la parte del trabajo computacional se observó con base en los pasos aplicados conforme a la metodología BFMDS fue posible la conclusión del caso práctico y se dio respuesta a la cuestión de investigación planteada anteriormente.

De igual forma, en el trabajo "Efectos del endeudamiento de los hogares mexicanos en su ahorro y consumo: Un enfoque de Ciencia de Datos" [11] se menciona que las finanzas en los hogares mexicanos juegan un papel importante especialmente en función de diversos factores, para llevar a cabo un análisis sobre ello es indispensable recabar datos, datos que se encuentran en diferentes encuestas oficiales, en esas encuestas se puede hallar información acerca de las fuentes de ingreso de las familias, de qué forma controlan su consumo, incluyendo atributos sociodemográficos que los caracterizan. El objetivo de este trabajo de investigación fue agrupar diferentes tipos de muestras respecto a hogares mexicanos que se encuentren en un estado económico de endeudamiento además de que compartan atributos sociodemográficos con alguna similitud

aplicando el algoritmo k-means de tal forma que se estimen modelos de tipo no lineales con el fin de medir los efectos de la deuda de los grupos en su consumo y en su ahorro. Entre las variables económicas y socioeconómicas que forman parte de la muestra.

- Gasto Corriente Monetario.
- Ingreso Corriente.
- Tasa de Ahorro.
- Tasa de Endeudamiento.
- Localidad.
- Educación Formal del Jefe de Hogar.
- Edad del Jefe de Hogar.
- Sexo del jefe de hogar.

El nivel educativo del jefe de familia se define como una variable importante debido a que la tasa de endeudamiento según el nivel educativo causa una tendencia que va en aumento, por otro lado, los hogares en donde el jefe de familia no cuenta con educación formal se puede observar una tasa de ahorro del 6 %, el cual es mayor a los jefes de familia que cuentan con estudios de posgrado, tal como se muestra en la Figura 2.24.

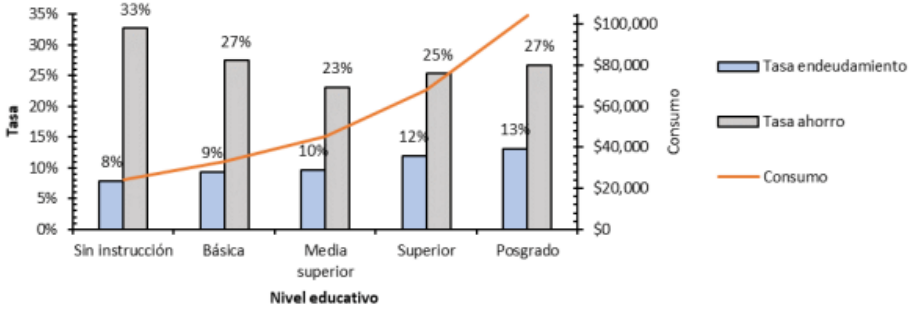


Figura 2.24: Distribución de las tasas de ahorro promedio, tasa de endeudamiento promedio y consumo por nivel educativo en hogares endeudados [11].

La Figura 2.25 presenta un nivel constante de la tasa de endeudamiento promedio conforme al rango de edad, en tanto se observa una tendencia que va en aumento en la tasa de ahorro de los hogares.

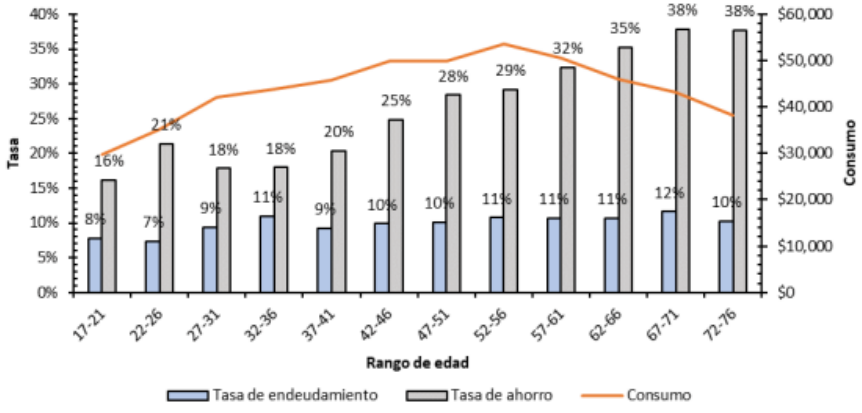


Figura 2.25: Distribución de la tasa de ahorro promedio, tasa de endeudamiento promedio y consumo por rangos de edad [11].

Para segmentar cualquier conjunto de datos en diferentes grupos se implementó el algoritmo K-means, el algoritmo forma parte de las técnicas de agrupamiento de tipo no supervisadas del aprendizaje automático. En las técnicas de agrupamiento no supervisadas, a comparación del agrupamiento supervisado, solo hay un conjunto de p características $X_1 ... X_p$ medidas para n observaciones. Además, no busca la predicción del comportamiento de una variable y. El método empleado para hallar el número más adecuado de clústeres en el algoritmo k-means fue el método elbow, que es un método iterativo gráfico que se basa en la búsqueda de la reducción de la suma total de cuadrados dentro de cada clúster, mediante el proceso iterativo que se muestra a continuación:

1. Poner en marcha el algoritmo K-means para diferente número de clústeres.
2. Calcular la suma promedio de cuadrados en cada clúster.
3. Graficar el valor de la suma total de cuadrados dentro de los clústeres contra el número de k clústeres para cada ejecución del algoritmo.
4. Identificar el punto de inflexión en la gráfica y seleccionar el número más adecuado de k clústeres como el número óptimo de grupos.

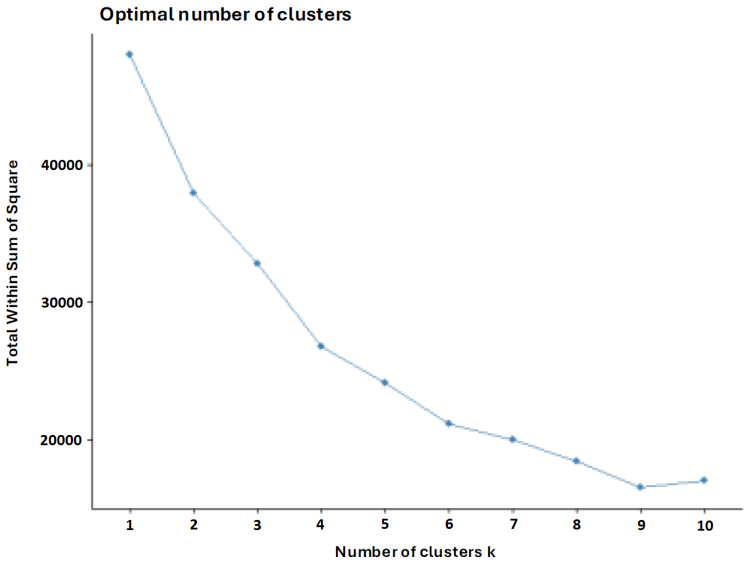


Figura 2.26: Implementación del método elbow para determinar el número óptimo de clústeres [11].

En los resultados se identificaron 4 clústeres, destacando uno de ellos que representó el 3.4 % de la muestra. No obstante, la tasa promedio de endeudamiento en este clúster supera los 53 puntos porcentuales en comparación con los otros clústeres. Los clústeres 1,3 y 4 muestran tasas de endeudamiento del 7.2 %. 10.1 % y 8.1 %, por otro lado, el clúster 2 destaca entre los demás ya que está conformado por hogares que cuentan con altas tasa de endeudamiento, reflejando una tasa de endeudamiento del 61.5 %, tal como se muestra en la Tabla 2.3.

Clúster	**Tasa de endeudamiento**
1	7.2 %
2	61.5 %
3	10.1 %

4	8.1 %

Tabla 2.3: Tasa de endeudamiento por clúster.

Esta investigación concluye que los hogares con un nivel mayor de sobreendeudamiento se conforman por aquellos donde el jefe de familia cuenta con estudios superiores, es decir, se muestra que tener un nivel educativo alto, o ser parte al decil de ingresos más pobres de la población no significa que se mantengan niveles de endeudamiento sanos, sin embargo es probable que en los hogares endeudados con un nivel de ingreso bajo empleen la deuda como complemento de los pocos recursos que poseen, y los hogares que cuentan con un nivel de ingreso alto es posible que excedan los niveles sanos de endeudamiento a causa de la sobreoferta de productos de crédito disponibles.

De igual forma, los autores del artículo "Grandes Datos, Google y Desempleo" [12] analizan la relación entre la tasa de desempleo con otras variables. En el trabajo de investigación se emplearon datos de búsqueda en Google acerca del empleo para posteriormente pronosticar el nivel de la tasa de desempleo en México. Además, se consideraron una gran cantidad de datos (Big Data) y algoritmos de aprendizaje que son de gran ayuda para seleccionar el modelo adecuado para una predicción concreta. Algoritmos de aprendizaje: Los algoritmos de aprendizaje automático se definen como fragmentos de código que les permiten a los usuarios analizar grandes cantidades de datos complejos y tratar de encontrar algún significado en ellos. Cada algoritmo se basa en una secuencia de instrucciones finitas que puede seguir una máquina para realizar un determinado proceso. Para un modelo de aprendizaje automático la finalidad es detectar patrones que puedan servir para realizar predicciones o clasificar información.

- Algoritmos de clústeres: Buscan similitudes, fragmentan los datos en diferentes grupos determinando que tanto es la semejanza entre los puntos de datos. Saber a qué tantos espectadores les gusta el mismo género de película. Los modelos de impresoras o cualquier hardware generan errores similares.

- Algoritmos de clasificación: Identificar cuál de los correos no es deseado.

Big Data: Engloba datos que contienen una variedad superior, sus volúmenes se presentan de manera creciente y con una velocidad a gran escala. Al

Big Data lo conforman un conjunto de datos de mayor tamaño y con un nivel de complejidad mayor, comúnmente procedentes de diferentes fuentes de datos [42]. Conforme pasa el tiempo Google Trends ha tenido gran fama debido a la gran cantidad de temas de interés para algunos usuarios, la gran mayoría de personas utilizan este motor de búsqueda. La interpretación de la información que presenta Google Trends se puede definir como sencilla, a causa del índice que contiene, va de 0 a 100 en el periodo de tiempo establecido. Los números representan el mayor valor de interés, es decir, un valor de 100 indica la popularidad más alta, mientras que 50 y 0 se refieren a una popularidad a la mitad o menor al 1 %. Para obtener el índice respectivo, Google Trends emplea una medida de interés de los términos de búsqueda, en lugar del total de búsquedas del término anterior. Google Trends se apoya en la siguiente ecuación para calcular el índice de interés.

$$interes = \frac{Busquedas \quad empleo}{Total \quad de \quad busquedas \quad Google} \tag{2.3}$$

En la Figura 2.27 se presenta una matriz de correlación en forma numérica y mediante un gráfico de dispersión, utilizando datos de Google Trends junto con la tasa de desocupación a nivel nacional, desglosada por hombres, mujeres y el promedio general. En el primer renglón y columna se encuentra IGT, en el siguiente las tasas de desocupación, en el tercer y cuarto renglón da lugar a la tasa para hombres y mujeres, respectivamente.

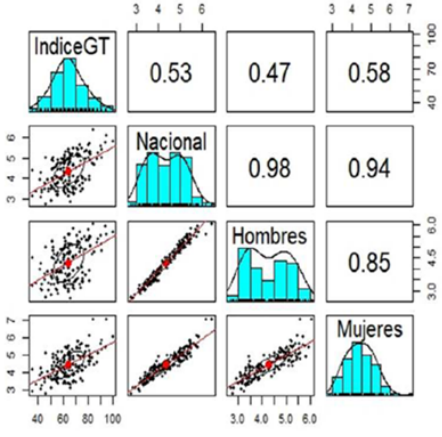


Figura 2.27: Matriz de correlación [12].

Las correlaciones que se obtuvieron en las tasas de desocupación como se mostró en la Tabla 2.1, la desocupación de los hombres tiene una mayor similitud a la nacional (corr=0.98) que la de las mujeres (corr=0.94), la correlación entre las dos anteriores es más baja (corr=0.85). En índice de Google Trends cuenta con una correlación con la desocupación femenina (corr=0.58) a comparación con la masculina (corr=0.47), sin embargo, ambos valores se encuentran cercanos a la correlación que la tasa de nacional presenta (corr=0.53).

Método LASSO: La regresión LASSO (Least Absolute Shrinkage and Selection Operator) es un modelo muy útil en la aplicación del aprendizaje automático y la estadística. Lasso se destaca por el manejo de prevención del sobreajuste y el manejo de grandes cantidades de datos [43]. El método LASSO, es posiblemente el algoritmo de aprendizaje de máquina más empleado por economistas, a causa de su similitud con la regresión lineal clásica [44]. Para el problema de predicción se implementó el método LASSO con el objetivo de predecir el valor de la tasa de desocupación tomando en cuenta valores pasados tanto de la variable tasa de desocupación (desde t - 1 hasta t – 12) como del IGT (desde t hasta t – 12). Además, se incorporó al análisis las interacciones cúbicas de las variables y así explorar relaciones no lineales.

Bosque Aleatorio: El segundo método implementado fue un bosque aleatorio, se conoce como un método de predicción bastante utilizado debido a su gran capacidad predictiva sin tener que afinar tanto los hiperparámetros [44]. Para la investigación se tomó en cuenta el bosque aleatorio para predecir el valor de la tasa de desocupación nacional, se utilizaron regazos trimestrales del IGT (IGT_t, IGT_{t-1}, IGT_{t-3}, IGT_{t-6}, IGT_{t-9}, IGT_{t-12}). A diferencia de LASSO no se generan interacciones no lineales de las variables para posteriormente ingresarlas al modelo. Al probar diferentes modelos predictivos con la finalidad de encontrar el que mejor resultados generen respecto a la tasa de desempleo a nivel nacional, se detectó que el método LASSO si generó ganancias predictivas a un método propio de la econometría, es decir los modelos autorregresivos (AR).

También, los autores del artículo “La costumbre al envenenamiento: el caso de los contaminantes atmosféricos de la ciudad de Guadalajara, en México” [45] utilizan un modelo de regresión para predecir valores. En este trabajo se aborda el problema de contaminación en varias ciudades de Latinoamérica, las medidas de contaminación que se registran no son normales y su vez afecta a gran escala el estado de salud de los ciudadanos.

El objetivo principal de la investigación es inferir sobre la presencia de enfermedades pulmonares en la ciudad de Guadalajara, Jalisco, México además de su correlación con los niveles de contaminación que se presentan en el ambiente. En un modelo de regresión se incluyen datos tanto cualitativos como cuantitativos. Cuantitativos se refiere a los datos numéricos, mientras que los

cualitativos se refieren a los que no se pueden representar de manera numérica o de magnitud. Para el análisis de datos se implementó el modelo de regresión múltiple, que es un modelo que parte de la regresión lineal simple, a diferencia del anterior modelo, el de regresión lineal múltiple contiene diferentes variables independientes x asociadas a una variable dependiente y. A continuación se muestra la ecuación del modelo de regresión múltiple.

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{2.4}$$

Donde:
y es la variable dependiente.
$x_1, x_2, ..., x_n$ son las variables independientes.
b_0 es la intersección con el eje y.
$b_1, b_2, ..., b_n$ son los coeficientes de las variables independientes.

En problemas complejos o situaciones reales de fenómenos naturales es más conveniente trabajar con el modelo de regresión múltiple ya que su análisis se aproxima mejor a ese tipo de problemas. El análisis de varianza presentado en la Figura 2.28, determina el nivel de relación entre las variables dependientes (contaminantes atmosféricos) y la variable independiente (total de casos IRAs).

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor *p*
Regresión	6	138 653 566	23 108 928	13.61	0.000
CO (ppm)	1	17 838 709	17 838 709	10.50	*0.007
O_3 (ppm)	1	9 290 284	9 290 284	5.47	*0.037
SO_2 (ppm)	1	6 310 392	6 310 392	3.72	0.078
NO_2 (ppm)	1	1 270 270	1 270 270	0.75	0.404
$PM_{2.5}$ (µg/m3)	1	10 631	10 631	0.01	0.938
PM_{10} (µg/m3)	1	23 683 907	23 683 907	13.95	*0.003
Error	12	20 379 001	1 698 250		
Total	18	159 032 567			

*Los valores muestran un nivel de significancia del 95 % (p<0.05).

Figura 2.28: Análisis de varianza [12].

Se hace una comparación del valor p (nivel de significancia), para visualizar el impacto que las variables tienen con un porcentaje de correlación del 95 %, es decir, existe un riesgo del 5 % de error. Tienen un valor p para monóxido de carbono (CO) con 0.007, de ozono (OZ) con 0.037, partículas menores a 10 micras (PM10) con 0.003 catalogándose como el más significante en el modelo. Con el modelo de regresión múltiple se obtuvieron resultados que apuntan hacia la presencia de una correlación entre las variables independientes, incluyendo el ajuste (R2, R2 ajustado, R2 predictivo), en donde se hace una comparación

de los resultados adquiridos mediante la ecuación de regresión representada en la Figura 2.29.

S	R-cuad.	R-cuad.(ajustado)	R-cuad.(pred)
1303.17	87.19 %	80.78 %	61.91 %

Figura 2.29: Bondad del ajuste y ecuación de regresión [12].

El valor S se utiliza para evaluar que tan bien describe el modelo el resultado, normalmente utilizado cuando no existen coeficientes de R o R2, cuanto mayor sea el valor S, menor será la descripción de respuesta del modelo. En la parte de resultados se obtuvo una correlación notable (valor $p < 50$) para el monóxido de carbono (CO) = 0.007, de ozono (O3) =0.037 y partículas menores a 10 micras (PM10) = 0.003. Son valores significativos de agentes contaminantes en el aire que se deberían de tomar a consideración por política ambiental y salud pública. Los principales contaminantes que arrojan valores significantes en la IRAs en Guadalajara fueron el ozono, seguido del monóxido de carbono y en último lugar se encontraron las partículas menores a 10 micras.

En la Tabla 2.4 se puede observar la comparación entre los artículos revisados.

Título	**Descripción**	**Área**	**Resultado**
Ciencia de datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México [9].	Se aplicó la técnica XGBoost, se define como un algoritmo enfocado en aprendizaje automático basado en una amplia gama de aplicaciones como los árboles de decisión para resolver problemas de clasificación, predicción y regresión.	Educación	Las variables promedio en el primer periodo, porcentaje de la beca y la región, pero principalmente la variable promedio en el primer periodo es por mucho el predictor que determina si el alumno continuará con sus estudios universitarios.

Índice de factores que inciden en el desarrollo de las zonas metropolitanas en México [40].	Para el análisis de los datos se implementó la técnica de análisis factorial (AF) ya que a diferencia del modelo de regresión lineal esta técnica expresa las variables en términos de factores comunes específicos no observados.	Economía	En el análisis estadístico se demostró que, si el nivel de bienestar es mayor, la prosperidad aumenta.
La probabilidad del crimen y su relación con su crecimiento económico en México: un análisis regional [10].	Para la investigación se estima la probabilidad del índice delictivo con base a modelos de regresión de respuesta cualitativa, posteriormente, con métodos de panel, se calcula el impacto del crimen en el desarrollo económico.	Seguridad	Mayores niveles de escolaridad reducen la presencia criminal, lo cual frena el desempeño económico.
Aplicación de Ciencia de Datos para el análisis de datos de mortalidad por COVID-19 de México [7].	Para la investigación fue necesario la aplicación de la metodología BATCH FMDS de IBM, debido a que es una metodología de Ciencia de Datos orientada al dominio epidemiológico y la cual garantiza resultados favorables.	Salud	Los valores respecto a los indicadores de densidad poblacional y porcentaje de personas en situación de pobreza contaban con una notable correlación con los valores de la tasa de mortalidad a causa de COVID-19.

Efectos del endeudamiento de los hogares mexicanos en su ahorro y consumo: Un enfoque de Ciencia de Datos [11].	El objetivo del trabajo de investigación es agrupar diferentes tipos de muestras respecto a hogares mexicanos que se encuentren en un estado económico de endeudamiento.	Economía	Los hogares con un nivel mayor de sobreendeudamiento se conforman por aquellos donde el jefe de familia cuenta con estudios superiores, es decir, se muestra que tener un nivel educativo alto, o ser parte al decil de ingresos más pobres de la población no significa que se mantengan niveles de endeudamiento sanos.
Grandes Datos, Google y Desempleo [12].	En el trabajo de investigación se emplearon datos de búsqueda en Google acerca del empleo para posteriormente pronosticar el nivel de la tasa de desempleo en México.	Empleo	Al probar diferentes modelos predictivos con la finalidad de encontrar el que mejor resultados genere respecto a la tasa de desempleo a nivel nacional, se detectó que el método LASSO si generó ganancias predictivas a un método propio de la econometría.
La costumbre al envenenamiento: el caso de los contaminantes atmosféricos de la ciudad de Guadalajara, en México [45].	El objetivo principal de la investigación es inferir sobre la presencia de enfermedades pulmonares en la ciudad de Guadalajara, Jalisco, México además de su correlación con los niveles de contaminación que se presentan en el ambiente.	Salud	Los principales contaminantes que arrojan valores significantes en la IRAs en Guadalajara fueron el ozono, seguido del monóxido de carbono y en último lugar se encontraron las partículas menores a 10 micras.

Tabla 2.4: Comparación de trabajo relacionado.

Capítulo 3

La plataforma DATAMEX

3.1. Herramientas

En esta sección se describen las herramientas que se utilizaron para el desarrollo de este proyecto.

Para el diseño se utilizó Figma, que es un editor gráfico que se destaca por su funcionalidad de colaboración en tiempo real, permitiendo que los diseñadores, programadores y gerentes trabajen juntos en el diseño de un proyecto. Esta herramienta permite crear prototipos que actúan como planes de acción priorizando la comodidad del usuario al navegar por una aplicación o sitio web. Estos prototipos conectan las diferentes pantallas, modelando el comportamiento del usuario y facilitando la presentación y edición del borrador del producto con los clientes, así como la entrega a los desarrolladores. Figma se utiliza para diseñar la interfaz visual de un sitio o aplicación, incluyendo iconos, botones, formularios y otros elementos. Cada uno de estos componentes se integran en un sistema de diseño coherente para garantizar una experiencia de usuario consistente. Este editor gráfico facilita la manipulación y modificación de elementos en múltiples pantallas simultáneamente. La maquetación de un sitio web es fundamental para trasmitir confianza y causar buena impresión, lo cual es crucial para garantizar la usabilidad. Este proceso implica organizar y distribuir visualmente elementos, tales como menús, imágenes, botones, encabezados, párrafos, y seleccionar paletas de colores y tipografías. Una maqueta web bien diseñada facilita la creación de un diseño único que refleje la identidad del sitio y al mismo tiempo mejore la experiencia del usuario. La maquetación no solo organiza el contenido, sino que también permite visualizar el sitio web antes de su programación, lo que puede reducir en un 20 % el tiempo de desarrollo. Esto es comparable a crear planos antes de la construcción de un

edificio, asegurando que sea coherente, atractivo y funcional. Además, facilita las modificaciones sin alterar el código fuente, mejorando la colaboración entre los equipos de diseño y desarrollo [46].

Respecto al entorno de desarrollo integrado (IDE, por sus siglas en inglés), que es una aplicación diseñada para proporcionar un conjunto de herramientas completas que faciliten el desarrollo de software a los programadores. Un IDE típico incluye un editor de código fuente, herramientas de construcción automáticas y un depurador. La mayoría de los IDE también ofrecen autocompletado inteligente de código (IntelliSense). Algunos IDE, como NetBeans y Eclipse, integran un compilador, un intérprete o ambos, mientras que otros, como SharpDevelop y Lazarus, no los incluyen [47]. Existen IDEs dedicados a un solo lenguaje de programación, optimizando sus características para ajustarse al paradigma de dicho lenguaje. Por otro lado, hay numerosos IDE que soportan múltiples lenguajes de programación, como Eclipse, ActiveState Komodo, IntelliJ IDEA, MyEclipse, Oracle JDeveloper, NetBeans, Codenvy y Microsoft Visual Studio. IDEs como Xcode, Xojo y Delphi están orientados a lenguajes específicos o a ciertos tipos de ajustes de lenguajes de programación.

Visual Studio Code, desarrollado por Microsoft, es un editor de código fuente de software libre disponible para Windows, GNU/Linux y macOS. VS Code se destaca por su excelente integración con Git, su capacidad para depurar código y una amplia variedad de extensiones, permitiendo escribir y ejecutar código en prácticamente cualquier lenguaje de programación.

Multiplataforma: Esta es una característica crucial para cualquier aplicación, especialmente en el ámbito del desarrollo. Visual Studio Code está disponible en Windows, GNU/Linux y macOS.

IntelliSense: Esta funcionalidad se vincula con la edición de código, el autocompletado y el resaltado de sintaxis, facilitando una escritura de código más eficiente. Tal como su nombre lo sugiere, ofrece sugerencias y completaciones inteligentes de código basadas en tipos de variables, funciones, entre otros. Con el apoyo de extensiones, es posible personalizar IntelliSense y hacerlo más completo para cualquier lenguaje de programación.

Visual Studio Code cuenta con una potente función de depuración que facilita la detección de errores en el código, eliminando la necesidad de revisar manualmente línea por línea en busca de fallos. Además, VS Code es capaz de identificar automáticamente pequeños errores antes de ejecutar el código o iniciar el proceso de depuración.

Visual Studio Code integra Git, lo que te permite verificar diferencias (conocido como git diff), organizar archivos, realizar commits directamente desde el editor, y gestionar push y pull con cualquier servicio de control de código fuente (SCM). Además se puede acceder a otros SCMs a través de extensiones disponibles.

Las extensiones son un elemento crucial para dicho IDE, Visual Studio Code se destaca como un editor robusto en gran parte gracias a estas extensiones. Ellas posibilitan la personalización y la adición de funcionalidades adicionales de manera modular y separada. Por ejemplo, facilitan la programación en diversos lenguajes, la incorporación de nuevos temas al editor y la integración con otros servicios. Las extensiones realmente enriquecen nuestra experiencia de uso y, lo más significativo, no afectan el rendimiento del editor, ya que operan en procesos independientes [48].

El desarrollo web es el proceso de crear y mantener un sitio web funcional en internet, utilizando diferentes lenguajes de programación según el modelo y la parte específica de la página. Un sitio web puede clasificarse de varias maneras. En términos de desarrollo web, se divide principalmente en dos componentes.

Frontend: Esta es la sección del sitio web que interactúa directamente con el usuario, tanto en términos de apariencia como de funcionalidad. Por lo tanto, está estrechamente relacionada con la experiencia del usuario (UX) y la interfaz de usuario (IU).

Backend: Hace referencia a la sección que interactúa directamente con el servidor, donde se escribe el código de programación que crea la estructura del sitio. Opera en segundo plano, gestionando la accesibilidad, las actualizaciones, las bases de datos y las modificaciones del sitio.

El diseño web está vinculado al frontend, ya que se encarga de definir la apariencia estética del sitio web. En cambio, el desarrollo web está asociado con el backend, asegurándose de que el código que sustenta la estructura del sitio sea tanto funcional como lógico [49, 50]. Para el diseño y el desarrollo web del proyecto se implementaron las siguientes herramientas fundamentales para conformar una página web funcional. También, se utilizó HTML, un lenguaje que utiliza etiquetas para definir la jerarquía de los elementos en una página. HTML actúa como el esqueleto de una página web: organiza el contenido, como encabezados, párrafos, tablas de datos, enlaces, insertando imágenes y videos en la página, asegurando que todo esté ordenado y en su lugar [57].

Por otra parte, para llevar a cabo la manipulación de datos, fue necesario la implementación de una herramienta que se especializa en ello, la librería Pandas cuenta con las características para manejar datos de alto nivel, con ayuda del lenguaje de programación Python es posible el proceso de grandes cantidades de datos. Pandas fue desarrollada por Wes Mckinney y construida basándose en Numpy lo que permite emplear el análisis de datos que contiene estructuras de datos necesarias para el proceso de limpieza de los mismos y al final el conjunto de datos seleccionado sea adaptado para el análisis [53, 54, 55].

De igual forma, una de las herramientas más eficaces para visualizar las relaciones entre múltiples variables son los gráficos de correlación. Estos gráficos facilitan el análisis de dichas relaciones. En este sentido, Seaborn es una de

las bibliotecas de visualización de datos más destacadas en Python, ofrece dos funciones principales para crear estos gráficos: los mapas de calor (heatmap) y los gráficos de pares (pairplot) [64, 65].

3.2. Implementación de la metodología de IBM

En esta sección se describe cómo se implementaron las etapas de la metodología IBM para ciencia de datos en este proyecto de investigación.

3.2.1. Etapa 1: Comprensión del negocio

En esta etapa se aplicó una encuesta a una muestra de 100 personas, la cual evidencia en cada sección el interés de los usuarios acerca de contar con un análisis de datos más profundo sobre temas que ellos consideran de mayor preocupación para la población mexicana. De acuerdo con la encuesta, se observó que el 61 % de la muestra no está conforme con la manera en la que el resultado del análisis de datos se presenta a la población, mientras que el resto de la muestra, el 39 %, considera que el resultado del análisis se presenta de forma entendible. Entre los temas que se consideraron de mayor importancia para la población se encuentran:

Seguridad (81 %) Educación (76 %) Economía (73 %) Salud (64 %) Empleo (60 %)

Con base en los resultados anteriores fue posible que el análisis se centrara solo en los temas seleccionados por los usuarios. En una de las preguntas se observa que más de la mitad de los encuestados, el 63 %, está de acuerdo con que se usen distintas fuentes de información para obtener los datos que servirán para el análisis. Entre más fuentes de datos se utilicen, existirá un análisis más completo ya que considera información proveniente de diversos conjuntos de datos. La respuesta de una cuestión detonante indica que el 78 % de la muestra dice no conocer alguna plataforma que implemente análisis de datos acerca de temas de interés para la población mexicana, incluyendo características como la predicción con base en los datos recolectados de distintas fuentes, por lo que se justifica el desarrollo de una plataforma que tenga estas herramientas en conjunto. La pregunta final consiste en saber si a la población le gustaría tener una plataforma web con un análisis de datos poblacionales sobre distintos rubros, donde se muestren los resultados del análisis a los usuarios, y además la plataforma tenga la capacidad de incorporar predicción de datos. A lo cual, el 100 % de las personas a las que se les aplicó la encuesta está a favor que exista una plataforma web con ese tipo de análisis de datos.

3.2.2. Etapa 2: Enfoque analítico

Una vez definido el problema con claridad, se procedió con la investigación de diversas técnicas de ciencia de datos, tales como:

Regresión lineal Árboles de decisión Bosques aleatorios Máquinas de vectores de soporte Análisis bayesiano

Con la finalidad de identificar las que mejor se adaptaran al modelo y de esta manera cumplir con los objetivos planteados. En este caso se optó por trabajar con regresión lineal, debido a que es una técnica que se especializa en la predicción de variables continuas a comparación de los casos de uso de las otras técnicas consideradas.

3.2.3. Etapa 3: Requisitos de datos

En esta fase se definieron los datos necesarios para la recopilación de los mismos.

Identificación de fuentes de datos: Se consultaron principalmente conjuntos de datos con un formato CSV (Valores separados por comas).

Definición de variables: Los conjuntos de datos debían contar con registros históricos de los temas de interés para la población (Seguridad, Educación, Economía, Salud, Empleo) referente a cada entidad federativa de la república mexicana.

Formato y estructura: El tipo de dato considerado para el análisis fue numérico, debido a que las variables se representaron gráficamente para posteriormente medir el nivel de correlación sin ningún problema.

Volumen y calidad: Para garantizar el buen funcionamiento del modelo se optó por tomar en cuenta bases de datos con una cantidad suficiente de registros, y con la mínima existencia de datos nulos o inconsistentes.

3.2.4. Etapa 4: Recopilación de datos

Los datos son extraídos, almacenados y preparados para su posterior procesamiento y análisis.

Fuentes de datos: Se obtuvieron los datos correspondientes de archivos, tablas, encuestas de algunos sitios web dedicados a la recolección y representación de datos, tales como instituto Nacional de Estadística y Geografía (INEGI), México como vamos, periódico el economista, el diario oficial de la federación, la biblioteca digital del senado, statista, data world bank, Kaggle.

Extracción de datos: Las bases de datos con el formato adecuado se descargaron, por otro lado, los datos con información no estructurada y semiestructurada se estructuraron en manera de tabla.

Trasformación de datos: Ya contando con los conjuntos de datos de manera estructurada se procedió a la trasformación de los mismos al formato CSV (Valores separados por comas), con la ayuda del editor de texto Sublime.

Almacenamiento: Para el resguardo de los datos previo a su análisis, con la herramienta Jupyter notebook que implementa Python, un lenguaje de programación con las características suficientes para la manipulación de grandes cantidades de datos se cargó la información.

3.2.5. Etapa 5: Comprensión de datos

Ya contando con los conjuntos de datos correspondientes, se cargaron para un análisis preliminar, es decir, se evaluó la calidad de la información, para ello se empleó estadística descriptiva y herramientas de visualización

Para esta etapa se aplicaron una serie de pasos los cuales ayudaron a tomar decisiones sobre su procesamiento y transformación antes de construir el modelo predictivo o realizar análisis más avanzados.

Exploración general: Se consideraron diferentes puntos a corregir propios de la estructura del conjunto de datos que estaban interviniendo con el análisis descriptivo, tales como el tipo de dato de algunas variables, columnas con contenido irrelevante, datos faltantes y duplicados.

Limpieza preliminar: Se identificaron datos nulos que podrían ocasionar problemas al momento de la preparación de datos.

Análisis descriptivo: En algunos casos se agregó una nueva columna que concentraba el valor general de homicidios, secuestros, robos y desaparecidos efectuados en cada mes por entidad federativa y año.

Visualización de datos: Las librerías que se utilizaron para generar elementos visuales con estilo fueron matplotlib y seaborn, el tipo de gráficos obtenidos fueron de pastel, de líneas, de barras, e histogramas. Para realizar la representación se tomaron variables cuantitativas y cualitativas.

3.2.6. Etapa 6: Preparación de datos

En un trabajo de ciencia de datos consiste en transformar, limpiar y estructurar los datos para que sean adecuados para el análisis y modelado. La fase de preparación fue crítica para el proyecto, ya que la calidad del modelo depende en gran medida de la calidad de los datos utilizados.

Limpieza de datos: Los procesos que emplearon en este paso sirvieron para corregir o en otro caso eliminar, registros incompletos, inexactos, irrelevantes o corruptos.

Selección de características: Existen columnas que cuentan con contenido no apto para el análisis, de esta manera se le fue dando la estructura deseada

al conjunto de datos, con la finalidad de no incluir información poco relevante.

Transformación de datos: Con columnas excluidas, en algunos casos, se generó un nuevo dataframe con registros repetidos, entonces para evitar la redundancia de datos se agruparon las filas, con el objetivo de juntar esos registros, indicando que se agrupara con base en el año y la entidad federativa.

Generación de nuevas variables: En los conjuntos de datos enfocados al tema de seguridad se incluyó una nueva columna llamada Sumatoria, así fue posible obtener un valor que involucrara el número total de incidentes delictivos por cada mes según el año y la entidad federativa.

Codificación de datos categóricos: Antes de comenzar con el modelado, se realizó una conversión de variables en formatos numéricos, de esta manera fue posible medir el nivel de correlación entre ellas.

Detección de relaciones: En esta parte se hizo uso del coeficiente de correlación, el cual implica que, para medir el grado de relación entre dos variables, éstas necesariamente deben ser continuas y cuantitativas.

Las variables consideradas para aplicar el análisis de regresión se basaron en las siguientes categorías.

Para visualizar la manera en la que se relacionan los valores de las variables descritas anteriormente, se consideró el uso de matrices de correlación, en la cual se muestra el coeficiente de correlación para cada par de variables, y en conjunto con la función heatmap() utilizada, que representa un mapa de calor, facilitó el análisis de las relaciones obtenidas. Seaborn, una de las librerías de visualización de datos más destacadas en Python, ofreció dicha función para crear el mapa de calor.

Resultó una cantidad de valores que abarcaron del -1 hasta el 1, Entre más cerca del -1 y del 1 indica una correlación mucho mayor, la diferencia es que para -1 es correlación negativa y para 1 es correlación positiva, mientras que para valores cercanos al 0 la correlación es baja o incluso hasta nula.

3.2.7. Etapa 7: Modelado

Se utilizó la primera fase del conjunto de datos previamente preparado. Para la construcción del modelo se probaron diferentes algoritmos, con la finalidad de conocer el funcionamiento, revisar la precisión e identificar la resistencia al sobreajuste de cada modelo con base en los conjuntos de datos seleccionados, para finalmente verificar cuál algoritmo se adaptó mejor al entrenamiento y ofreció una mejor precisión.

1. Selección del modelo: Cada algoritmo tiene un caso de uso diferente:

Regresión, si el objetivo es predecir valores continuos.

Clasificación, si el objetivo es predecir categorías.

Agrupamiento, cuando se quiere encontrar grupos naturales en los datos. Para los casos de prueban los algoritmos enfocados para clasificación, se adaptaron para predecir valores continuos.

2. Entrenamiento del modelo: Para cada algoritmo se indicó que el 20 % de los datos originales se reservara para pruebas, y el 80 % restante se usó para entrenar el modelo.

3. Ajuste de hiperparámetros: Se optimizaron los parámetros del modelo para mejorar su desempeño.

4. Evaluación del modelo: Se mide el rendimiento con métricas como:

Para regresión: Coeficiente de determinación (R^2), error cuadrático medio (MSE).

Para clasificación: Precisión, Recall, F1-score, matriz de confusión.

5. Comparación de modelos: Se probaron diferentes modelos para elegir el mejor con base en los resultados obtenidos en la siguiente etapa.

3.2.8. Etapa 8: Evaluación

Se midió el rendimiento con las métricas para regresión, los resultados obtenidos fueron los siguientes.

Algoritmo	**Métrica**	**Precisión**
Árboles de decisión	MSE	0.66
Bosques aleatorios	MSE	0.33
Máquinas de vectores de soporte	R^2	0.54
Teorema de Bayes	R^2	0.53
Regresión lineal	R^2	0.74

Tabla 3.1: Precisión de los modelos entrenados.

En la última columna de la tabla anterior se logra apreciar el nivel de precisión de los modelos entrenados con cada algoritmo, el modelo entrenado con regresión lineal obtuvo una precisión del 74 %, lo que indica el valor más alto. Las evidencias de las precisiones obtenidas se muestran en la Seccion 4.1.

3.2.9. Etapa 9: Implementación

Con base en la etapa anterior, la cual permitió evaluar cada algoritmo considerado, se optó por trabajar con regresión lineal, debido al nivel de precisión que arrojo el modelo entrenado con dicho algoritmo. Cada predicción empleada en la investigación se realizó con regresión lineal, para posteriormente incluir los resultados en una plataforma fácil de acceder y consultar.

Se desarrollo un sitio web, con el objetivo de que la población pudiera tener acceso al análisis de datos completo de temas relevantes, así como lo expuso en la encuesta aplicada.

En la parte de "¿Qué nos dicen los datos?" se puede apreciar un análisis prescriptivo, el cual demuestra el comportamiento de los datos, sobre algunos temas, al pasar de los años.

El usuario puede tener acceso a la regresión, en la sección "Predicción de datos", donde se explica como primer punto la importancia de la correlación, la cual es la base para realizar predicciones precisas, posteriormente se muestra el par de variables correspondientes, que darán lugar al valor predicho con base en la otra.

3.2.10. Etapa 10: Retroalimentación

Como parte del proyecto se aplicaron pruebas de usabilidad, en la cual los usuarios expresaron sus opciones sobre la manera de navegar, la calidad del contenido que ofrece DATAMEX, y otros aspectos clave.

3.3. Preparación de datos

Para comenzar el análisis de datos, es primordial obtener los datos a analizar, es decir, identificar el origen y recopilarlos. Los datos recopilados se transforman para posteriormente cargarlos. Como primer punto se debe contar con un dataset que incluya la información adecuada y así dar inicio a el análisis. En este sentido, la Tabla 3.3 muestra las fuentes de datos para el tema de Seguridad.

Subtemas	**Fuentes de datos**
Robos	datos.gob.mx inegi.org.mx dof.gob.mx

Secuestros	cdmx.gob.mx datos.gob.mx fgjcdmx.gob.mx datamx.io
Homicidios	datos.gob.mx inegi.org.mx es.statista.com causaencomun.org.mx
Grupos Criminales	politicadedrogas.org causaencomun.org.mx datos.gob.mx seguridadviacivil.ibero.mx mexicosocial.org
Personas desaparecidas	bibliodigitalibd.senado.gob.mx datos.gob.mx

Tabla 3.2: Subtemas y fuentes de datos para el tema de Seguridad.

De igual forma, los subtemas y fuentes de datos para el tema de Educación se muestran en la Tabla 3.4.

Subtemas	**Fuentes de datos**
Becas	datos.gob.mx inegi.org.mx datamx.io
Abandono escolar	datos.gob.mx inegi.org.mx es.statista.com
Escolaridad	datos.gob.mx

Tabla 3.3: Subtemas y fuentes de datos para el tema de Educación.

Respecto a los subtemas y fuentes de datos del tema Economía, estos se muestran en la Tabla 3.5.

Subtemas	**Fuentes de datos**
Inflación canasta alimentaria	data.worldbank statista.com
Producto interno bruto	datamx.io
Apoyo adultos mayores	datamx.io

Tabla 3.4: Subtemas y fuentes de datos para el tema de Economía.

Por otra parte, los subtemas y fuentes de datos del tema Salud se muestran en la Tabla 3.6.

Subtemas	**Fuentes de datos**
Pandemias	kaggle.com who.int datos.gob.mx datos.covid-19.conacyt.mx datamx.io
Adicciones	who.int datos.gob.mx inegi.org.mx

Tabla 3.5: Subtemas y fuentes de datos para el tema de Salud.

Respecto a los subtemas y fuentes de datos para del tema Empleo, estos se muestran en la Tabla 3.7.

Subtemas	**Fuentes de datos**
Desempleo	proyectosmexico.gob.mx datos.bancomundial.org
Salario	datos.gob.mx datamx.io datamexico.org

Asegurados	datos.gob.mx statista.com datos.cdmx.gob.mx

Tabla 3.6: Subtemas y fuentes de datos para el tema de Empleo.

Cabe mencionar que los archivos analizados son los siguientes, mostrados respecto al número de filas por columnas.

Subtemas	**Filas por columnas**
Homicidios	25088 x 19 5 x 34 6049 x 18 257 x 17 22177 x 21
Robos	24865 x 18 108998 x 6
Secuestros	1280 x 19 129 x 17 673 x 18 12321 x 21
Desaparecidos	36265 x 15 2465 x 21 35157 x 15
Desempleo	19 x 2 298 x 6 17 x 33
Escolaridad	32 x 2 2545 x 45
PIB	32 x 2 270 x 78 1500 x 68
Casos COVID	32 x 3 232002 x 25

Tabla 3.7: Archivos.

Para que sea posible su reconocimiento mediante la librería Pandas, el archivo Excel que contiene el conjunto de datos necesita estar guardado bajo el formato CSV (Valores Separados por Comas), generando una tabla donde existen diversas filas y una columnas queda definida por cada punto y coma [51, 52].

Posteriormente, se importa la librería Pandas con el comando que se muestra en la Figura 3.1, además de otras librerías que sirven para generar elementos visuales con estilo.

```
In [17]: import pandas as pd
         import seaborn as sns
         import statsmodels.formula.api as smf
         import matplotlib.pyplot as plt
         sns.set_style('darkgrid')
```

Figura 3.1: Bibliotecas necesarias para la manipulación de datos.

El archivo CSV se leerá de la manera que se muestra en la Figura 3.2, donde se indica el nombre del mismo entre comillas con el formato adecuado.

```
In [2]: # Cargar datos
        df = pd.read_csv("IDEFC_NM_jun23.csv")
        df
```

Figura 3.2: Línea de código para leer datos en formato CSV.

Se muestran los registros de cada homicidio según el año, la entidad y la modalidad, incluyendo los homicidios efectuados en cada mes como se aprecia en la Figura 3.3. En total, el conjunto de datos lo conforman 28, 224 registros incluyendo los datos nulos tal como se observa en la Figura 3.4.

Out[2]:

	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre
0	2015	1	Aguascalientes	La vida y la Integridad corporal	Homicidio	Homicidio doloso	Con arma de fuego	3	0	2	1	1	1	2.0	1.0	2.0
1	2015	1	Aguascalientes	La vida y la Integridad corporal	Homicidio	Homicidio doloso	Con arma blanca	1	1	0	0	0	1	0.0	1.0	0.0
2	2015	1	Aguascalientes	La vida y la Integridad corporal	Homicidio	Homicidio doloso	Con otro elemento	0	0	2	2	3	2	0.0	1.0	2.0
3	2015	1	Aguascalientes	La vida y la Integridad corporal	Homicidio	Homicidio doloso	No especificado	2	0	0	1	0	0	0.0	0.0	0.0
4	2015	1	Aguascalientes	La vida y la Integridad corporal	Homicidio	Homicidio culposo	Con arma de fuego	0	0	0	0	1	0	0.0	0.0	0.0

Figura 3.3: Columnas del dataframe homicidios.

25083	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Falsificación	Falsificación	Falsificación	5	14	7	10	9	19	6.0	5.0	10.0
25084	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Contra el medio ambiente	Contra el medio ambiente	Contra el medio ambiente	1	0	0	0	0	0	0.0	0.0	0.0
25085	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Delitos cometidos por servidores públicos	Delitos cometidos por servidores públicos	Delitos cometidos por servidores públicos	12	26	45	84	26	25	22.0	31.0	24.0
25086	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Electorales	Electorales	Electorales	1	0	3	0	1	0	0.0	0.0	1.0
25087	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Otros delitos del Fuero Común	Otros delitos del Fuero Común	Otros delitos del Fuero Común	146	122	174	150	133	170	141.0	182.0	158.0

25088 rows × 19 columns

Figura 3.4: Total de registros y datos nulos.

Preparación de los datos

En la etapa de preparación engloba las actividades enfocadas en construir el conjunto de datos apropiado para pasar a la siguiente etapa, una de las actividades más importante es la limpieza de datos, esto se refiere a tratar con valores faltantes o que no son válidos, en resumen, su objetivo es eliminar duplicados y dar un formato adecuado.

El proceso de limpieza de datos consiste en una secuencia de pasos con el objetivo de identificar posibles problemas con los datos y posteriormente corregirlos. Entre los principales tipos de errores sobre los datos se encuentran:

- Datos faltantes
- Datos irrelevantes
- Duplicación de datos
- Valores atípicos: Intervienen directamente en el rendimiento del modelo.
- Errores estructurales: Errores referentes a la tipografía y otras incoherencias.

El principal objetivo de la limpieza de datos es mejorar la calidad de ellos, los procesos que emplea esta etapa sirven para corregir o en otro caso eliminar, registros incompletos, inexactos, irrelevantes o corruptos.

Sin aplicar correctamente la limpieza podría repercutir directamente en la calidad de los datos, es decir, ocasionar que los resultados obtenidos del análisis estén distorsionados. De la misma manera, un modelo que involucre IA como el aprendizaje automático, entrenado con datos no aptos puede estar sujeto a presentar un bajo rendimiento [56].

La línea de código de la Figura 3.5 se deshace de datos faltantes dentro del dataframe, lo que disminuirá la cantidad de registros como se evidencia en la Figura 3.6, pero es necesario para un análisis concreto. El proceso de limpieza es un paso importante para comenzar a preparar los datos.

```
In [3]: # Eliminar datos nulos
        df.dropna(inplace=True)
        df
```

Figura 3.5: Línea de código para eliminar datos faltantes.

25085	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Delitos cometidos por servidores públicos	Delitos cometidos por servidores públicos	Delitos cometidos por servidores públicos	12	26	45	84	26	25	22.0	31.0	24.0
25086	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Electorales	Electorales	Electorales	1	0	3	0	1	0	0.0	0.0	1.0
25087	2022	32	Zacatecas	Otros bienes jurídicos afectados (del fuero co...	Otros delitos del Fuero Común	Otros delitos del Fuero Común	Otros delitos del Fuero Común	146	122	174	150	133	170	141.0	182.0	158.0

25088 rows × 19 columns

Figura 3.6: Registros respecto a homicidios excluyendo datos faltantes.

Existen columnas que cuentan con contenido no apto para el análisis, de esta manera se le va dando la estructura deseada al dataset. La Figura 3.7 muestra la manera de ejecutarlo, con la finalidad de no incluir información poco relevante. En este caso, se tomó en cuenta la columna, Año, Entidad, Tipo de Delito y la de cada mes, como se evidencia en la Figura 3.8.

```
In [4]: # Seleccionar columnas a utilizar
        df = df[['Año', 'Entidad', 'Tipo de delito', 'Enero', 'Febrero', 'Marzo', 'Abril', 'Mayo', 'Junio', 'Julio', 'Agosto', 'Septiemb
        df
```

Figura 3.7: Línea de código para seleccionar columnas específicas.

Out[4]:

	Año	Entidad	Tipo de delito	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	2015	Aguascalientes	Homicidio	3	0	2	1	1	1	2.0	1.0	2.0	2.0	2.0	1.0
1	2015	Aguascalientes	Homicidio	1	1	0	0	0	1	0.0	1.0	0.0	0.0	0.0	1.0
2	2015	Aguascalientes	Homicidio	0	0	2	2	3	2	0.0	1.0	2.0	0.0	0.0	0.0
3	2015	Aguascalientes	Homicidio	2	0	0	1	0	0	0.0	0.0	0.0	0.0	0.0	0.0
4	2015	Aguascalientes	Homicidio	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.0	0.0

Figura 3.8: Dataframe con las columnas específicas.

Con las columnas excluidas se generó un nuevo dataframe con registros repetidos, tal como se observa en la Figura 3.8, porque no se tomó en cuenta la columna Modalidad. En tal columna se describe cómo se llevó a cabo el homicidio. Entonces, para evitar la redundancia de datos se agruparon las filas, con la finalidad de juntar esos registros, indicando, con la línea de código presentada en la Figura 3.9, que se agrupe con base al año, entidad y tipo de delito.

```
In [5]: # Agrupar filas por:
        df_agrupado = df.groupby(['Año', 'Entidad', 'Tipo de delito']).sum().reset_index()
        df_agrupado
```

Figura 3.9: Línea de código para agrupar filas.

Las filas disminuyeron a 10240 registros, debido a que ahora solo se visualiza un solo tipo de delito por año y entidad federativa, como se observa en la Figura 3.10.

	Año	Entidad	Tipo de delito	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	2015	Aguascalientes	Aborto	0	3	1	0	0	0	0.0	0.0	0.0	0.0	1.0	0.0
1	2015	Aguascalientes	Abuso de confianza	41	33	31	22	36	43	30.0	40.0	40.0	34.0	43.0	26.0
2	2015	Aguascalientes	Abuso sexual	6	4	0	2	1	1	1.0	1.0	0.0	0.0	1.0	1.0
3	2015	Aguascalientes	Acoso sexual	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0
4	2015	Aguascalientes	Allanamiento de morada	13	11	18	16	16	19	18.0	19.0	15.0	11.0	4.0	9.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10235	2022	Zacatecas	Tráfico de menores	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0
10236	2022	Zacatecas	Violación equiparada	5	13	15	5	13	7	14.0	14.0	10.0	8.0	4.0	9.0
10237	2022	Zacatecas	Violación simple	6	14	27	20	12	23	14.0	9.0	14.0	18.0	12.0	8.0
10238	2022	Zacatecas	Violencia de género en todas sus modalidades d...	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0
10239	2022	Zacatecas	Violencia familiar	194	235	272	288	388	347	322.0	315.0	292.0	264.0	255.0	236.0

10240 rows × 15 columns

Figura 3.10: Dataframe con agrupaciones.

Para este análisis sólo se incluyeron los registros de la columna tipo de delito que se refiere a homicidios ejemplificado en la Figura 3.11, por lo que se genera un dataframe que solo incluya el tipo de delito seleccionado por año y entidad, como se observa en la Figura 3.12.

```
In [6]: # Seleccionar el tipo de delito
        homicidios = df_agrupado.loc[:, 'Tipo de delito'] == 'Homicidio'
        df_hom = df_agrupado.loc[homicidios]
        df_hom
```

Figura 3.11: Línea de código para seleccionar el tipo de delito.

	Año	Entidad	Tipo de delito	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
18	2015	Aguascalientes	Homicidio	22	20	14	21	18	16	16.0	14.0	27.0	16.0	9.0	18.0
58	2015	Baja California	Homicidio	75	83	87	94	93	106	113.0	114.0	120.0	110.0	118.0	106.0
98	2015	Baja California Sur	Homicidio	20	11	11	22	18	17	17.0	28.0	32.0	9.0	10.0	6.0
138	2015	Campeche	Homicidio	13	7	18	10	10	13	2.0	9.0	8.0	10.0	6.0	12.0
178	2015	Chiapas	Homicidio	124	149	145	162	134	116	122.0	128.0	115.0	132.0	107.0	118.0

Figura 3.12: Dataframe de homicidios.

Se incluyó una nueva columna llamada Sumatoria. La Figura 3.13 muestra la manera en la que se ejecutó esta acción, con el objetivo de tener un valor que involucre la suma de homicidios por cada mes según el año y la entidad federativa, como se muestra en la Figura 3.14

```
In [7]: # Agregar una nueva columna
        df_hom ['Sumatoria'] = df_hom ['Enero'] + df_hom ['Febrero'] + df_hom ['Marzo'] + df_hom ['Abril'] + df_hom ['Mayo'] +
        df_hom
```

Figura 3.13: Línea de código para agregar una nueva columna.

	Año	Entidad	Tipo de delito	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Sumatoria
18	2015	Aguascalientes	Homicidio	22	20	14	21	18	16	16.0	14.0	27.0	16.0	9.0	18.0	211.0
58	2015	Baja California	Homicidio	75	83	87	94	93	106	113.0	114.0	120.0	110.0	118.0	106.0	1219.0
98	2015	Baja California Sur	Homicidio	20	11	11	22	18	17	17.0	28.0	32.0	9.0	10.0	6.0	201.0
138	2015	Campeche	Homicidio	13	7	18	10	10	13	2.0	9.0	8.0	10.0	6.0	12.0	118.0
178	2015	Chiapas	Homicidio	124	149	145	162	134	116	122.0	128.0	115.0	132.0	107.0	118.0	1552.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10058	2022	Tamaulipas	Homicidio	99	77	101	99	113	112	116.0	133.0	90.0	84.0	80.0	86.0	1190.0
10098	2022	Tlaxcala	Homicidio	7	19	12	21	13	14	11.0	16.0	10.0	22.0	7.0	13.0	165.0
10138	2022	Veracruz de Ignacio de la Llave	Homicidio	155	132	198	168	171	141	138.0	164.0	139.0	146.0	128.0	165.0	1845.0
10178	2022	Yucatán	Homicidio	17	18	20	15	24	18	20.0	13.0	14.0	13.0	11.0	15.0	198.0
10218	2022	Zacatecas	Homicidio	136	78	120	121	95	114	109.0	96.0	98.0	132.0	131.0	130.0	1360.0

256 rows × 16 columns

Figura 3.14: Dataframe con la columna Sumatoria.

Teniendo la columna sumatoria es posible tener un dato general sobre la cantidad de homicidios ocurridos en cada entidad durante 2015 y 2022, con base al año en el que se registraron, y con ello poder emplear este tipo de función, tal como se observa en la Figura 3.15, la cual ordena los estados con mayor número de homicidios con base en la sumatoria. Los resultados muestran

que por tres años consecutivos (2018, 2019 y 2020), Guanajuato fue el estado más violento del país, como se evidencia en la Figura 3.16.

```
In [8]: # Visualizar los estados más violentos
        by_sum = df_hom.sort_values('Sumatoria', ascending=False)
        by_sum.head(10)
```

Figura 3.15: Línea de código para ordenar los estados más violentos.

	Año	Entidad	Tipo de delito	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Sumatoria
6818	2020	Guanajuato	Homicidio	443	400	444	399	403	392	418.0	392.0	431.0	430.0	408.0	380.0	4940.0
5538	2019	Guanajuato	Homicidio	366	382	407	395	403	352	319.0	341.0	362.0	366.0	385.0	416.0	4494.0
4258	2018	Guanajuato	Homicidio	323	283	382	361	392	335	349.0	369.0	414.0	376.0	328.0	386.0	4298.0
9618	2022	México	Homicidio	320	274	354	295	332	298	312.0	303.0	361.0	338.0	305.0	330.0	3822.0
8338	2021	México	Homicidio	286	303	310	311	336	265	295.0	295.0	307.0	341.0	313.0	356.0	3718.0
8098	2021	Guanajuato	Homicidio	350	283	340	335	319	269	325.0	278.0	281.0	307.0	281.0	305.0	3673.0
7058	2020	México	Homicidio	259	259	333	315	273	288	286.0	306.0	295.0	329.0	330.0	302.0	3575.0
9378	2022	Guanajuato	Homicidio	271	241	317	261	320	294	297.0	270.0	299.0	335.0	298.0	310.0	3513.0
5778	2019	México	Homicidio	292	278	328	298	301	279	294.0	279.0	271.0	257.0	290.0	305.0	3472.0
2978	2017	Guanajuato	Homicidio	260	228	292	266	286	274	271.0	282.0	253.0	323.0	305.0	326.0	3366.0

Figura 3.16: Dataframe con 10 registros respecto al mayor número de homicidios.

3.4. Análisis y visualización de datos

Una vez finalizado el proceso de preparación de datos, se comienza con el análisis de los mismos. Con ayuda de la librería Pandas es posible la manipulación de los datos y se le da la estructura deseada para facilitar su procesamiento. Cabe mencionar que el equipo de cómputo utilizado tiene las siguientes características, sistema operativo Windows 11 Home, memoria RAM de 8 GB, almacenamiento de 256 GB SSD, procesador AMD Ryzen 3, y pantalla FHD de 15.6 pulgadas. Respecto al análisis, existen diferentes tipos, para estos ejemplos se aplicó el análisis descriptivo, el cual se basa principalmente en la visualización de datos de manera gráfica, se aplican gráficos de barras, circulares, lineales, tablas, entre otros, con la finalidad de comprender mejor el comportamiento de la información.

Análisis diagnóstico:

Se especializa en un análisis profundo y detallado del conjunto de datos para comprender por qué genera tales resultados. Emplea técnicas que involucran

la minería de datos o las correlaciones. Estas técnicas utilizan múltiples transformaciones para procesar datos brutos.

Análisis predictivo:
Para este tipo de análisis es necesario usar datos históricos con el objetivo de realizar previsiones precisas acerca de las tendencias futuras. Las técnicas basadas en el análisis predictivo son el machine leraning, el modelado predictivo, la coincidencia de patrones y la previsión.

Análisis prescriptivo:
Parte del análisis predictivo, lo que permite que no solo realice predicciones de futuros acontecimientos, también sugiere una respuesta óptima para tales predicciones. Con este tipo de análisis es posible analizar las diferentes implicaciones y recomendar el mejor curso de acción. Se destaca por emplear la simulación, el análisis de gráficos, el procesamiento de eventos complejos, los motores de recomendación y las redes neuronales.

Para poder mostrar el resultado del análisis descriptivo, se emplearon las líneas de código mostradas en la Figura 3.17, para obtener la gráfica "Homicidios por entidad federativa" la cual se representa en la Figura 3.18. Se aplicó el mismo procedimiento para las representaciones gráficas de cada conjunto de datos.

```
In [9]: # Visualizar graficamente los estados más violentos
grouped_data = df_hom.groupby('Entidad')['Sumatoria'].sum().reset_index()
plt.figure(figsize=(10, 6))
plt.bar(grouped_data['Entidad'], grouped_data['Sumatoria'], color='skyblue')
plt.xlabel('Entidad')
plt.ylabel('Homicidios')
plt.title('Homicidios por Entidad Federativa')
plt.xticks(rotation=45, ha='right')

plt.show()
```

Figura 3.17: Código para generar la gráfica "Homicidios por entidad federativa".

Como se puede obsevar en la Figura 3.18, los datos muestran que Guanajuato es el estado de la república con el mayor número de homicidios, registrando un total de 29,254 casos. Le siguen el Estado de México con 27,170 casos y Baja California con 20,671 casos. En contraste, los estados del sur entre los que se encuentran Yucatán (1,102), Campeche (1,213) y Tlaxcala (1,838) tienen los índices más bajos de homicidios, de acuerdo con los registros comprendidos entre los años 2015 y 2022.

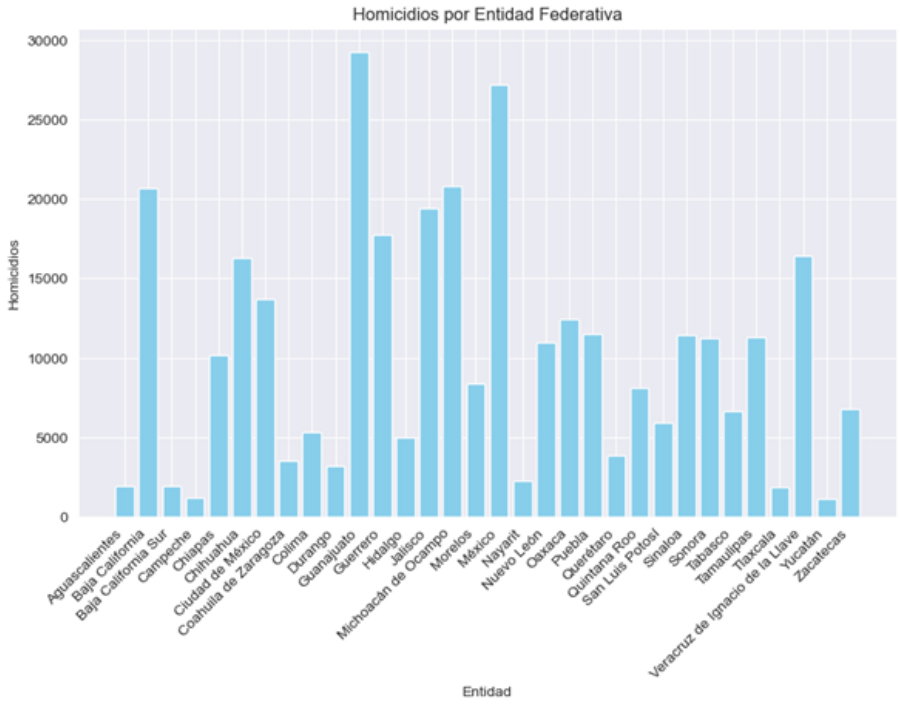


Figura 3.18: Homicidios por entidad federativa.

Como se puede observar en la Figura 3.19, el estado de Tamaulipas se posiciona como la entidad con el mayor número de personas desaparecidas, registrando un total de 5,990 casos. Le siguen el Estado de México con 3,890 casos, Jalisco con 3,362 y Sinaloa con 3,027. Por el contrario, los estados de la república con menos desapariciones son Tlaxcala con 24 casos, Campeche con 35, Baja California Sur con 39 y Quintana Roo con 61 casos. Cabe mencionar que Tlaxcala es el estado que tiene los índices más bajos de personas desaparecidas en la República Mexicana, según los datos registrados entre los años 2006 y 2018.

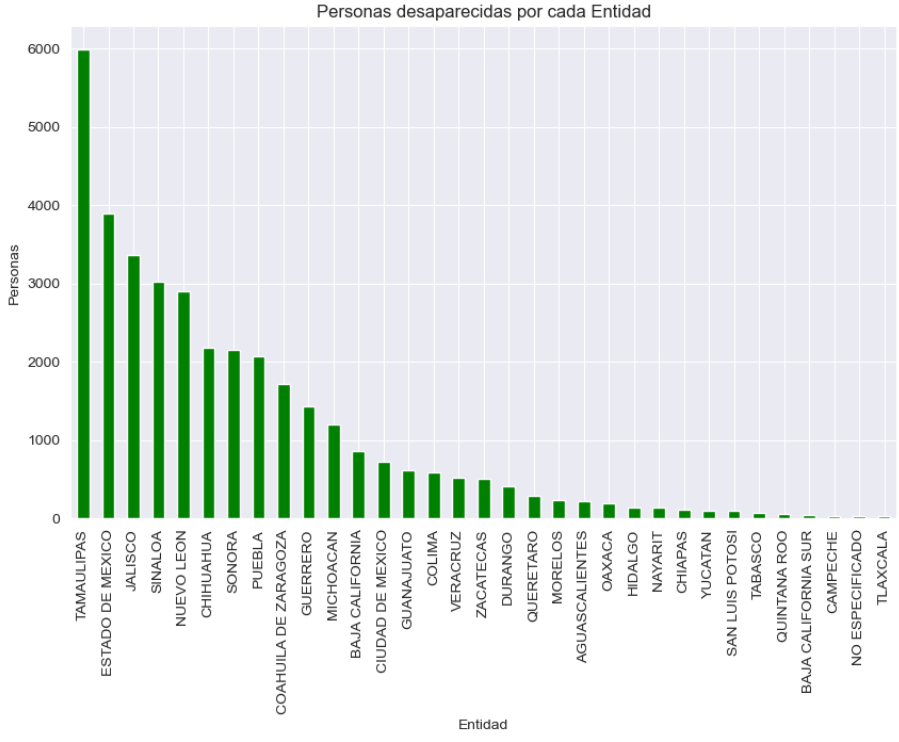


Figura 3.19: Personas desaparecidas por entidad federativa.

Por otra parte, como se puede observar en la Figura 3.20, entre los años 2006 y 2018 se reportaron un total de 26,938 hombres desaparecidos, lo que representa el 74.3 % del total, y un total de 9,327 mujeres desaparecidas, que corresponden al 25.7 %. Estos datos evidencian una marcada diferencia entre ambos géneros en cuestión de desapariciones en México.

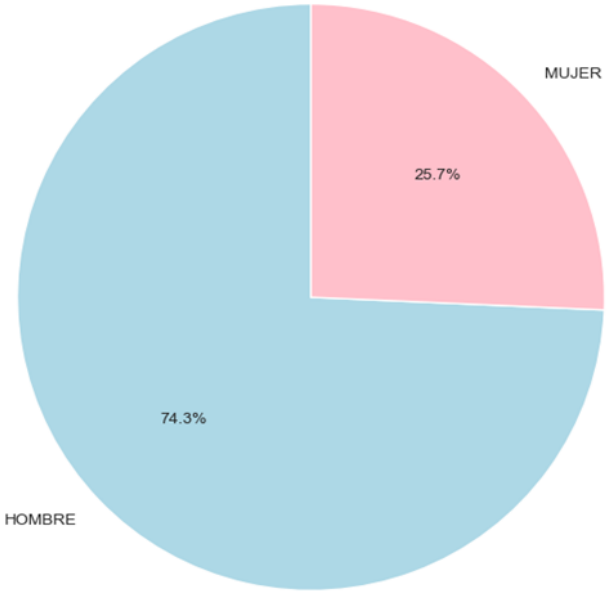


Figura 3.20: Porcentaje de personas desaparecidas por género.

De igual forma, la Figura 3.21 muestra los estados de la república con el mayor número de robos registrados entre los años 1997 y 2017. Cabe mencionar que la Ciudad de México es el estado que lidera la posición con un total de 2,080,881 incidentes, seguida por el Estado de México con 2,064,767, Baja California con 1,214,908 y Jalisco con 761,245 incidentes. Por otra parte, los estados con las cifras más bajas de robos son Campeche con 20,148 incidentes, Nayarit con 55,121, Tlaxcala con 57,547 y Colima con 70,979 incidentes. Es importante mencionar que los datos incluyen tres tipos diferentes de robos: comunes, en carreteras y en instituciones bancarias.

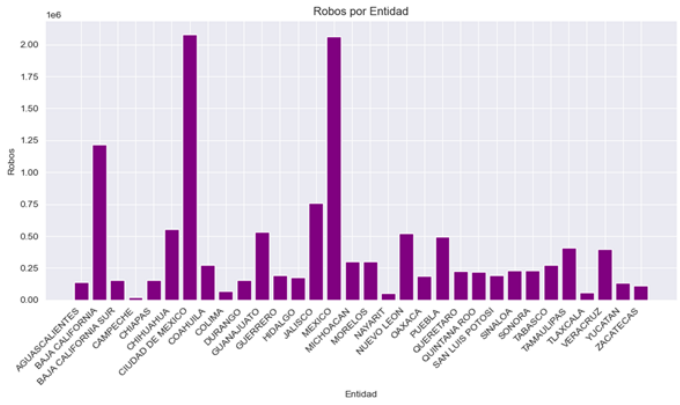


Figura 3.21: Robos por entidad federativa.

A nivel nacional, en la Figura 3.22 se puede observar que el Estado de México registró 1,347 secuestros entre 2015 y 2022. Le siguen las entidades de Veracruz con 1,083 casos, la Ciudad de México con 744 y Tamaulipas con 727. En contraste, los estados de Baja California Sur (21), Campeche (35), Durango (41) y Nayarit (42) presentaron cifras significativamente más bajas.

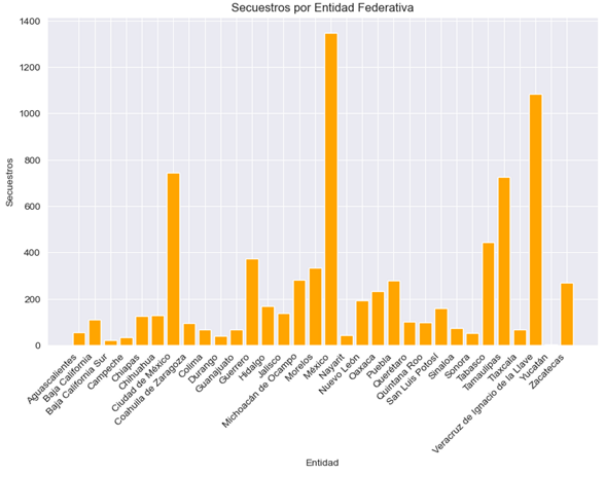


Figura 3.22: Secuestros por entidad federativa.

La gráfica respecto al desempleo en México se muestra en la Figura 3.23, destacan los eventos económicos globales, como la crisis financiera de 2008 y la pandemia de COVID-19, han influido significativamente en el desempleo. Entre los años 2009-2012 la tasa de desempleo aumentó considerablemente durante y después de la crisis financiera, alcanzando un pico de 5.3 % en 2011 antes de comenzar a disminuir. Durante 2020-2021 la tasa de desempleo aumentó nuevamente en 2020 a 4.5 %, probablemente debido a la pandemia de COVID-19 y sus efectos en el mercado laboral. Sin embargo, la tasa comenzó a disminuir en 2021 a 3.8 %. Durante 2022-2024, la tasa de desempleo continuó disminuyendo, alcanzando los niveles más bajos de la serie de datos, con un mínimo de 2.7 % en 2024.

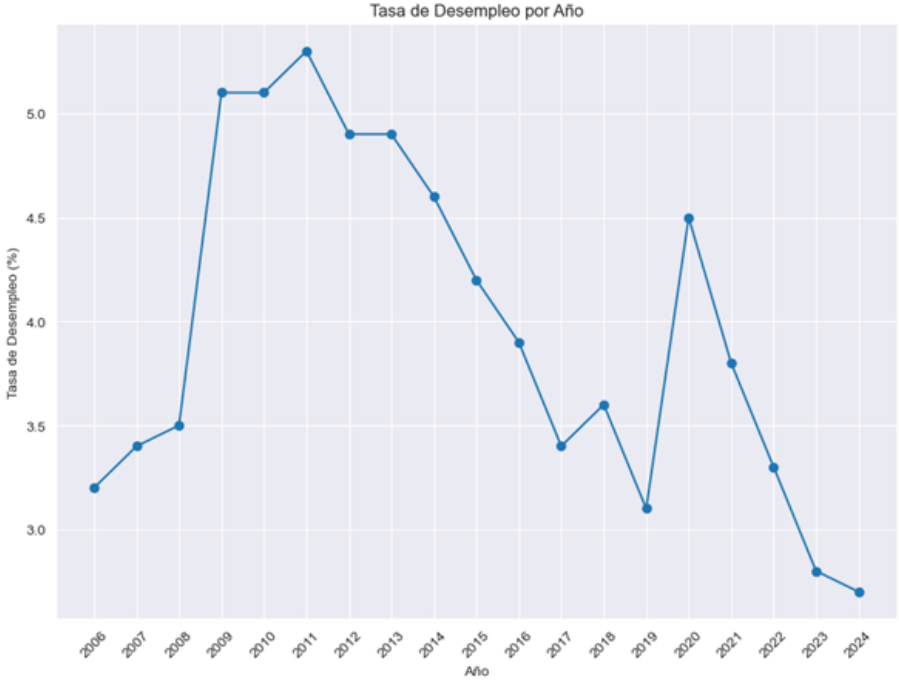


Figura 3.23: Tasa de desempleo por año.

Los datos mostrados en la Figura 3.24 indican una notable variación en el número de fallecidos entre diferentes estados durante el 2022, con una concentración más alta en entidades con mayor densidad poblacional y centros

urbanos grandes como el Estado de México con 55,786 fallecidos y la Ciudad de México con 51,545 fallecidos. Por otro lado, estados con menor densidad poblacional como Chiapas con 1,918 fallecidos y Baja California Sur con 2,424 fallecidos, reportan cifras significativamente menores de defunciones.

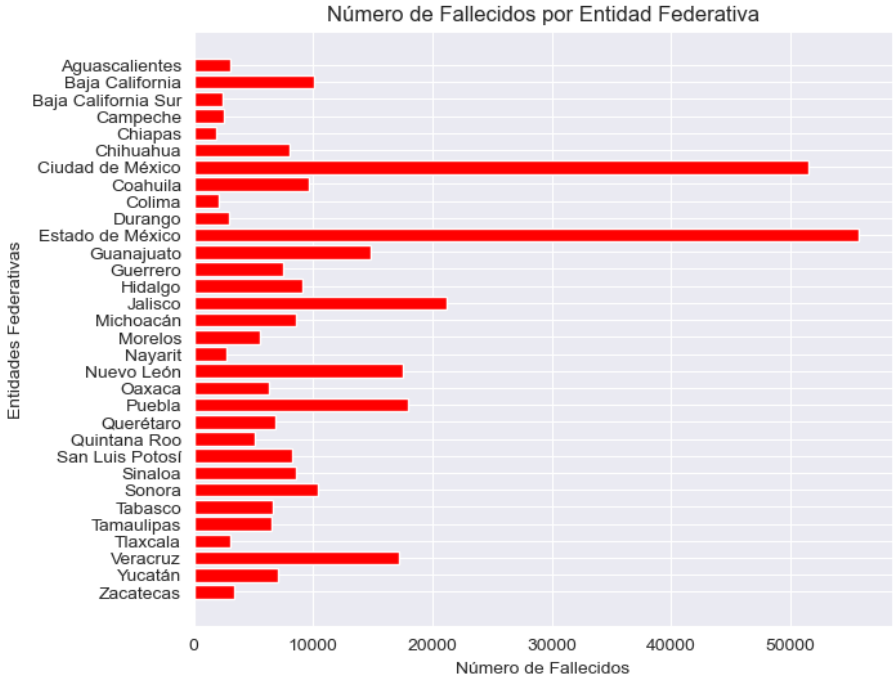


Figura 3.24: Número de fallecidos por entidad federativa.

La Ciudad de México destaca significativamente con la mayor contribución, siendo el principal motor económico del país con una contribución de 14.22 %, tal como se muestra en la Figura 3.25, al igual que el Estado de México (9.03 %), Nuevo León (8.01 %), Jalisco (7.56 %). La brecha entre las entidades con mayores y menores contribuciones es considerable, indicando una concentración de la actividad económica en ciertos estados. Los estados con menor contribución tienden a ser menos industrializados y con menor densidad poblacional, tales como Tlaxcala (0.60 %), Colima (0.61 %), Nayarit (0.68 %) y Baja California Sur (0.74 %), lo que puede influir en su menor participación en el PIB nacional.

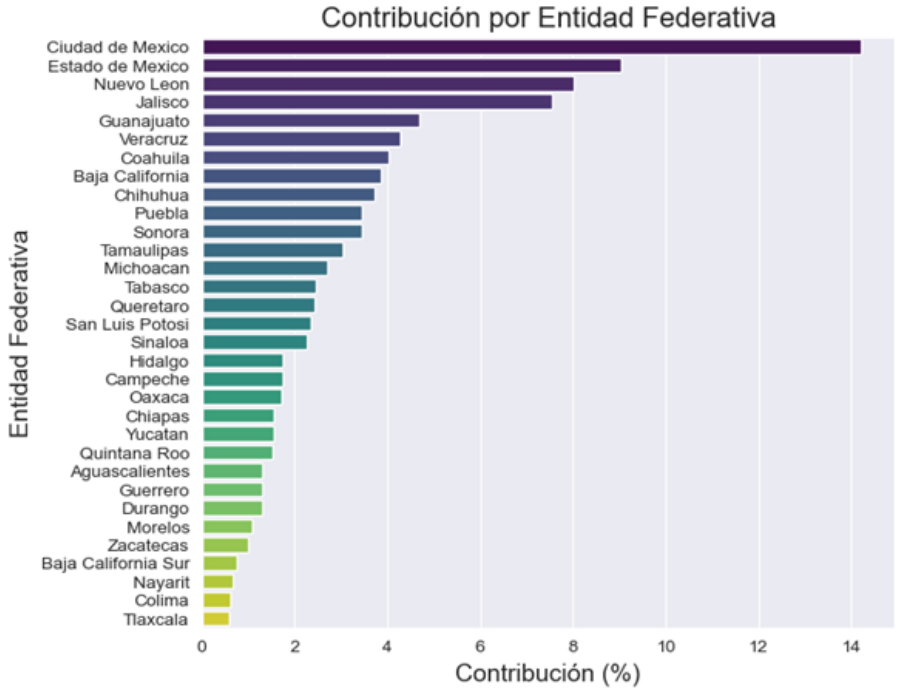


Figura 3.25: Contribución por entidad federativa.

La gráfica con respecto a la escolaridad del país se muestra en la Figura 3.26, donde se observa que la Ciudad de México con un grado de escolaridad de 11.5, lidera el ranking. Nuevo León, conocido por su desarrollo industrial y económico, tiene un grado de escolaridad de 10.7. Querétaro con un grado de escolaridad de 10.5 se destaca por su rápido crecimiento y desarrollo en los sectores industrial y de servicios. También, se observa que Chiapas tiene el menor grado de escolaridad con 7.8. Chiapas enfrenta desafíos significativos en términos de infraestructura educativa y acceso a la educación, especialmente en áreas rurales. Oaxaca, con un grado de escolaridad de 8.1 también enfrenta obstáculos similares a Chiapas. Similarmente, Guerrero es un estado con un grado de escolaridad de 8.4.

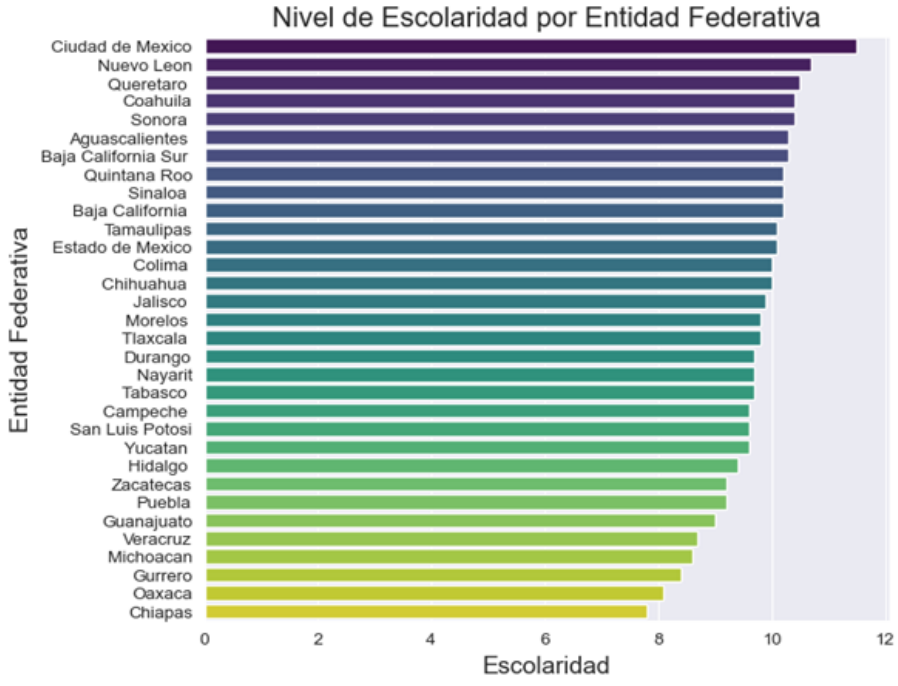


Figura 3.26: Nivel de escolaridad por entidad federativa.

3.5. Desarrollo de la plataforma DATAMEX

El diseño de la plataforma DATAMEX se realizó en Figma. Para la interfaz visual del sitio se estableció la distribución de los elementos, tales como botones, menú de navegación, iconos, imágenes, las secciones para cada encabezado y párrafo. De igual manera, el editor gráfico Figma ayudó a definir la tipografía y colorimetría deseada. El diseño se observa en la Figura 3.27.

Figura 3.27: Maquetación de DATAMEX en Figma.

En la parte lateral izquierda se desglosan los elementos utilizados para la interfaz principal, tal como se observa en la Figura 3.28.

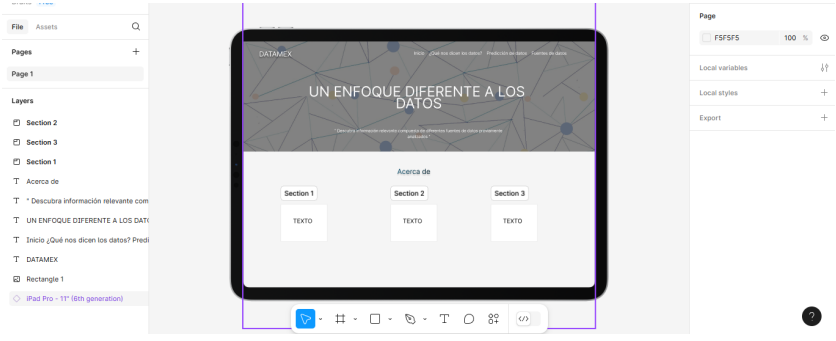


Figura 3.28: Herramientas de Figma

Se utilizó HTML (HyperText Markup Language) para dar estructura al contenido de la página web. Cabe mencionar que se utilizó la versión HTML 5. En la Figura 3.29 se muestra una parte de código, aplicando el lenguaje de etiquetas, perteneciente a la página principal de DATAMEX, al igual que las demás secciones que conforman el proyecto en la parte izquierda.

Figura 3.29: Fragmento de código HTML en Visual Studio Code.

Se utilizó CSS (Hojas de Estilo en Cascada) que es considerado un lenguaje de estilos. Cabe mencionar que se utilizó la versión 3. La función principal de CSS es dividir la presentación de la estructura: se encarga de la apariencia y el estilizado de los elementos, permitiendo gestionar colores respecto al fondo, tipografías, diseños y otros aspectos visuales. Es como adornar la estructura HTML con un atuendo sofisticado para que la página luzca tal como deseas [57].

En la Figura 3.30 se muestra la hoja de estilo en cascada aplicada a la página principal que se mostró en la Figura 3.27. Con ello, cada sección del sitio web se representa de una manera ordenada y estética gracias a HTML y CSS.

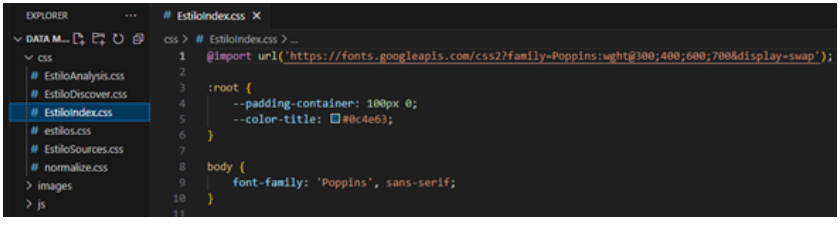


Figura 3.30: Fragmento de código CSS en Visual Studio Code.

JavaScript, por su lado, es el único lenguaje de programación entre los tres mencionados anteriormente. En este caso se encarga de añadir interactividad y dinamismo a las páginas web. Facilita la implementación de funciones como animaciones, validaciones de formularios, controlar multimedia y actualizaciones de contenido en tiempo real [58]. Además de estructura y diseño el sitio web debe contar con interactividad y dinamismo, para ello se programaron varias funciones encargadas de ello con el lenguaje JavaScript como se visualiza

en la Figura 3.31.

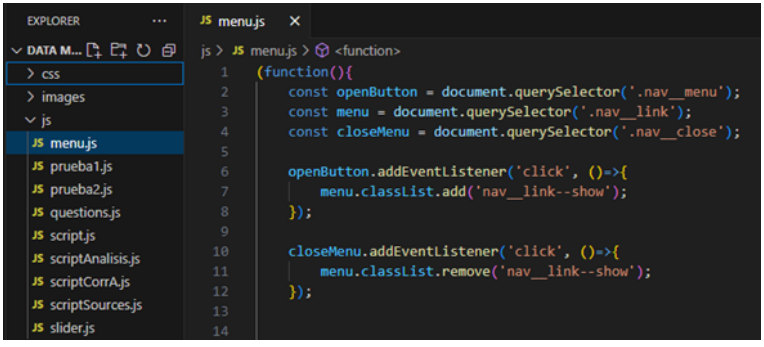


Figura 3.31: Fragmento de código JavaScript en Visual Studio Code.

JavaScript es un lenguaje de programación de alto nivel que se ejecuta de forma interpretada. De acuerdo con la Encuesta de Desarrolladores de StackOverflow de 2022, es el más utilizado a nivel mundial [59]. Esto se debe en gran medida a que JavaScript es el lenguaje estándar interpretado por los navegadores, y junto con HTML (Lenguaje de Marcado de Hipertexto) y CSS (Hojas de Estilo en Cascada), conforman la base fundamental de toda la Web [60].

El entorno de desarrollo utilizado para el sitio web DATAMEX fue Visual Studio Code en su versión 1.95 debido a las ventajas que este posee, tales como la amplia biblioteca de extensiones que permiten personalizar y expandir sus funcionalidades según nuestras necesidades específicas. Esto incluye soporte para una variedad de lenguajes de programación y herramientas de desarrollo web como HTML, CSS, JavaScript, TypeScript, frameworks como React, Angular y Vue.js, entre otros.

Para desarrollar cada sección del proyecto se requirió de un entorno de desarrollo capaz de brindar las herramientas necesarias para trabajar en conjunto con HTML, CSS y JavaScript. En la Figura 3.32 se observa la estructura de los archivos en Visual Studio Code.

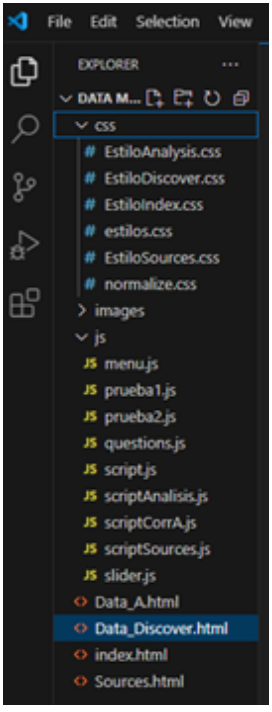


Figura 3.32: Entorno de Desarrollo Integrado (IDE) utilizado en el proyecto DATAMEX.

Como se puede observar, en este proyecto se utilizaron archivos CSS, JavaScript y HTML. Los resultados del proyecto se pueden consultar en la página web desarrollada que se encuentra disponible en www.datamex.com.mx.

Capítulo 4

Resultados

4.1. Predicciones

Se trabajó con diferentes algoritmos enfocados al análisis de datos, con el objetivo de conocer el funcionamiento, revisar la precisión e identificar la resistencia al sobreajuste de cada modelo con base en los conjuntos de datos seleccionados, para finalmente verificar cuál algoritmo se adapta mejor al entrenamiento y ofrece una mejor precisión. La Figura 4.1 muestra un ejemplo de las técnicas utilizadas.

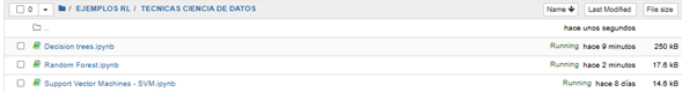


Figura 4.1: Técnicas de ciencia de datos.

El procedimiento llevado a cabo para predecir el salario con base al nivel de escolaridad aplicando árboles de decisión se describe a continuación. Primeramente, se importan las librerías correspondientes para utilizar el algoritmo enfocado a regresión, tal como se observa en la Figura 4.2.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import DecisionTreeRegressor
from sklearn import tree
import matplotlib.pyplot as plt
```

Figura 4.2: Biblioteca para árboles de decisión.

Posteriormente, se carga el conjunto de datos seleccionado con el que se entrenará el modelo, tal como se observa en la Figura 4.3.

```
# Datos proporcionados (Escolaridad y Salario)
data = {
    'Escolaridad': [10.3, 10.2, 10.3, 9.6, 10.4, 10, 7.8, 10, 11.5, 9.7, 9, 8.4, 9.4, 9.9, 10.1, 8.6, 9.8, 9.7, 10.7, 8.1,
                    9.2, 10.5, 10.2, 9.6, 10.2, 10.4, 9.7, 10.1, 9.8, 8.7, 9.6, 9.2],
    'Salario': [14.065, 18.868, 18.596, 13.475, 15.4, 14.446, 11.185, 17.522, 19.723, 11.984, 13.237, 10.25, 13.077,
                13.887, 12.612, 13.007, 11.193, 12.689, 19.216, 11.438, 11.602, 15.556, 14.249, 15.685, 14.312, 15.104,
                13.399, 14.178, 10.844, 11.154, 14.525, 12.632]
}
```

Figura 4.3: Conjunto de datos prueba.

En la Figura 4.4 se muestra cómo se divide el conjunto de datos en dos partes, entrenamiento y pruebas, se crea el modelo de árbol de decisión basado en regresión, el modelo se entrena con la parte del conjunto correspondiente, el porcentaje de pruebas se utiliza para realizar las predicciones y finalmente se crea una función que muestra el diagrama representativo a árbol de decisión. Las hojas finales del árbol se muestran en la Figura 4.5.

```
# Convertir el conjunto de datos en un DataFrame
dfx = pd.DataFrame(data)

# Dividir los datos en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(dfx[['Escolaridad']], dfx['Salario'], test_size=0.2, random_state=42)

# Crear el modelo de Árbol de Decisión para regresión
model = DecisionTreeRegressor(random_state=42)

# Entrenar el modelo con los datos de entrenamiento
model.fit(X_train, y_train)
```

```
DecisionTreeRegressor
DecisionTreeRegressor(random_state=42)
```

```
# Realizar predicciones con el conjunto de prueba
y_pred = model.predict(X_test)

# Visualizar el Árbol de Decisión
plt.figure(figsize=(12, 8))
tree.plot_tree(model, feature_names=['Escolaridad'], filled=True)
plt.title("Árbol de Decisión - Escolaridad vs Salario")
plt.show()
```

Figura 4.4: Procesamiento de datos con árboles de decisión.

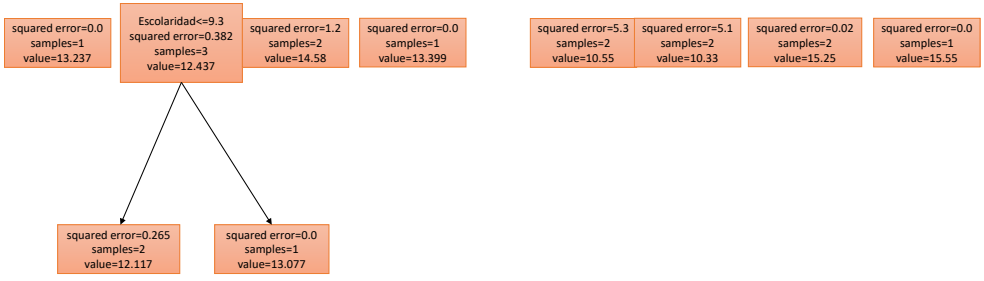


Figura 4.5: Las hojas finales del árbol de decisión sobre las variables de escolaridad y salario.

Con la función ejemplificada en la Figura 4.6 es posible obtener la precisión del modelo, la cual fue del 66 %, lo que indica que las predicciones no serán tan exactas.

```
# Calcular la precisión del modelo
accuracy = model.score(X_test, y_test)
accuracy
```

```
0.6614360311388788
```

Figura 4.6: Precisión del modelo.

Posteriormente, se importan las librerías necesarias para emplear el algoritmo de bosques aleatorios, tal como se observa en la Figura 4.7, a efectos de comparación, se carga el mismo conjunto de datos utilizado en árboles de decisión.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_squared_error
```

Figura 4.7: Biblioteca para bosques aleatorios.

Una vez realizado los dos pasos anteriores se continúa con la creación del

modelo, tal como se visualiza en la Figura 4.8, para ello se identifican las variables predictoras (x) y objetivo (y), es decir, se busca predecir el salario (y) con base en el nivel de escolaridad (x), se dividen los datos proporcionados en dos conjuntos, datos de entrenamiento y de prueba. El modelo se crea y se entrena con el conjunto de datos reservado para ello, las predicciones se efectúan con los datos de prueba. Finalmente, se evalúa el modelo y se calcula la precisión.

Figura 4.8: Procesamiento de datos con bosques aleatorios.

En los resultados se observa que el modelo tiene una precisión del 33 %, un porcentaje menor que el obtenido en el algoritmo anterior, sus predicciones se muestran con la última función, tal como se esperaba, el valor arrojado no coincide con lo esperado debido al porcentaje de precisión tal como se observa en la Figura 4.9.

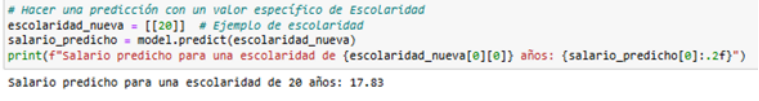


Figura 4.9: Predicción aplicando bosques aleatorios.

Posteriormente, se trabajó con el algoritmo de la Regresión de Vectores de Soporte (SVR). Cabe mencionar que mientras que el SVM busca encontrar un hiperplano que separe clases en problemas de clasificación, SVR busca encontrar una función que prediga un valor continuo, por ejemplo, el salario con base en la escolaridad. En la Figura 4.10 se observan las bibliotecas necesarias para trabajar distintas técnicas de máquinas de vectores de soporte.

```python
# Importar las bibliotecas necesarias
import numpy as np
import pandas as pd
from sklearn.svm import SVR
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score
```

Figura 4.10: Biblioteca para máquinas de vectores.

El modelo se basó en el conjunto de datos mostrado anteriormente, el cual contiene información sobre el salario y el nivel de escolaridad. Una vez creado el dataframe se continua con la separación de las variables, en este caso la variable a predecir corresponde a salario (y), por el contrario, escolaridad se define como la variable independiente (x). Antes de entrenar el modelo, se escalan tanto las variables independientes como las dependientes, para el entrenamiento se utiliza un kernel RBF (radial basis function), que es el más utilizado para SVR. Finalmente, se realiza una predicción de los salarios basados en la escolaridad escalada y se desescala la predicción para interpretarla en términos de los valores originales. Este proceso se observa en la Figura 4.11.

```python
# Convertir a DataFrame
df = pd.DataFrame(data)
```

```python
# Separar variables independientes (X) y dependientes (y)
X = df['Escolaridad'].values.reshape(-1, 1)
y = df['Salario'].values
```

```python
# Escalar las características
scaler_X = StandardScaler()
scaler_y = StandardScaler()

X_scaled = scaler_X.fit_transform(X)
y_scaled = scaler_y.fit_transform(y.reshape(-1, 1)).flatten()
```

```python
# Entrenar el modelo SVR
svr_model = SVR(kernel='rbf')
svr_model.fit(X_scaled, y_scaled)
```

```
▾ SVR
SVR()
```

Figura 4.11: Procesamiento de datos con máquinas de vectores de soporte.

El modelo SVR intenta ajustar una curva lo más cerca posible de los puntos de datos reales, manteniendo un margen para permitir ciertas variaciones en los datos, tal como lo muestra la Figura 4.12.

```python
# Visualizar resultados
plt.scatter(X, y, color='red', label="Datos Reales")
plt.plot(X, y_pred, color='blue', label="Predicción SVR")
plt.title("SVR - Predicción de Salario basado en Escolaridad")
plt.xlabel("Escolaridad")
plt.ylabel("Salario")
plt.legend()
plt.show()
```

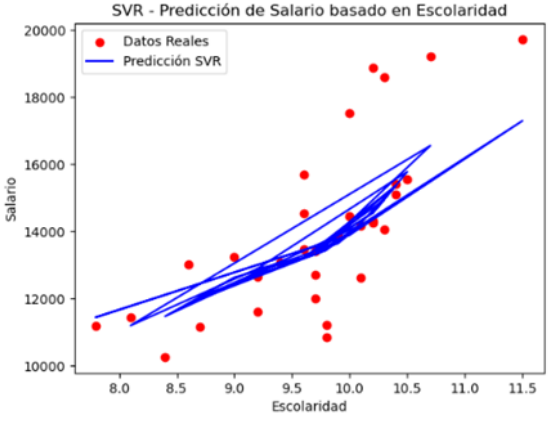


Figura 4.12: Gráfica resultante al aplicar SVR.

Para medir la precisión de un modelo de regresión SVR, se utilizan métricas específicas para problemas de regresión. Estas métricas son:

- Coeficiente de determinación (R^2): Indica que tan bien las predicciones del modelo se ajustan a los valores reales. Un valor de R^2 cerca de 1 indica un buen ajuste del modelo.
- Error cuadrático medio (MSE): Mide el promedio de los errores al cuadrado entre los valores reales y las predicciones.
- Raíz del error cuadrático medio (RMSE): Similar al MSE pero en la misma escala que los datos originales. La Figura 4.13 muestra el resultado de la evaluación del modelo. Se obtuvo un coeficiente de determinación del 0.54, lo que significa que las predicciones realizadas no tendrán buena exactitud.

```
# Calcular el error cuadrático medio (MSE)
mse = mean_squared_error(y, y_pred)

# Calcular la raíz del error cuadrático medio (RMSE)
rmse = np.sqrt(mse)

# Calcular el coeficiente de determinación R²
r2 = r2_score(y, y_pred)

# Mostrar los resultados
print(f"Error Cuadrático Medio (MSE): {mse}")
print(f"Raíz del Error Cuadrático Medio (RMSE): {rmse}")
print(f"Coeficiente de Determinación (R²): {r2}")
```

```
Error Cuadrático Medio (MSE): 2812773.1391864354
Raíz del Error Cuadrático Medio (RMSE): 1677.1324155195484
Coeficiente de Determinación (R²): 0.5480262316046391
```

Figura 4.13: Evaluación del modelo utilizando máquinas de vectores de soporte.

Posteriormente, para el algoritmo bayesiano se utilizaron las siguientes librerías, para aplicar el teorema de bayes con el conjunto de datos seleccionado, las cuales se ejemplifican en la Figura 4.14.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
```

Figura 4.14: Bibliotecas para teorema de bayes.

Como se mostró en los casos prácticos anteriores, se aplican una serie de pasos sobre el conjunto de datos, que van desde la preparación de los mismos hasta la creación del modelo y finalmente la obtención de los resultados, como se observa en la Figura 4.15.

```
# Convertir los datos en un DataFrame
df = pd.DataFrame(data)

# Definir las variables dependiente (y) e independiente (X)
X = df['Escolaridad']
y = df['Salario']

# Añadir una constante a X para la regresión
X = sm.add_constant(X)

# Ajustar el modelo de regresión lineal
model = sm.OLS(y, X).fit()

# Resumen del modelo
summary = model.summary()
summary
```

Figura 4.15: Procesamiento de datos con teorema de bayes.

El modelo de regresión lineal ajustado entre la escolaridad y el salario muestra los siguientes resultados.

- Coeficiente de determinación (R-squared): 0.531, lo que indica que aproximadamente el 53.1 % de la variación en los salarios puede explicarse por la escolaridad.
- Constante (intercepto): -8823.9995, que representa el salario esperado cuando la escolaridad es cero, aunque en este caso no tiene mucho sentido interpretar un valor negativo dado el contexto.
- p-valor para el coeficiente de escolaridad: 0.000, lo que indica que la relación entre escolaridad y salario es estadísticamente significativa.

Estos resultados se muestran en la Figura 4.16.

OLS Regression Results

Dep. Variable:	Salario	**R-squared:**	0.531
Model:	OLS	**Adj. R-squared:**	0.515
Method:	Least Squares	**F-statistic:**	33.92
Date:	Mon, 30 Sep 2024	**Prob (F-statistic):**	2.28e-06
Time:	22:32:51	**Log-Likelihood:**	-283.60
No. Observations:	32	**AIC:**	571.2
Df Residuals:	30	**BIC:**	574.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8823.9995	3937.081	-2.241	0.033	-1.69e+04	-783.407
Escolaridad	2354.2903	404.217	5.824	0.000	1528.768	3179.812

Omnibus:	0.665	**Durbin-Watson:**	1.840
Prob(Omnibus):	0.717	**Jarque-Bera (JB):**	0.658
Skew:	0.307	**Prob(JB):**	0.720
Kurtosis:	2.659	**Cond. No.**	124.

Figura 4.16: Resultados aplicando teorema de bayes.

Finalmente, se trabajó el mismo caso aplicando regresión lineal. El modelo mostrado a continuación define la relación entre una variable dependiente y una variable independiente. Se comenzó importando las librerías que corresponden a este algoritmo, tal como se observa en la Figura 4.17.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
```

Figura 4.17: Bibliotecas para la aplicación de regresión lineal.

Se carga el conjunto de datos que contiene registros sobre escolaridad y salario a nivel estado, su estructura se observa en la Figura 4.18.

```
df = pd.read_csv("Escolaridad-Salario Corr.csv")
df
```

	Entidad	Clave	Escolaridad	Salario
0	AGUASCALIENTES	1	10.3	14.065
1	BAJA CALIFORNIA	2	10.2	18.868
2	BAJA CALIFORNIA SUR	3	10.3	18.596
3	CAMPECHE	4	9.6	13.475
4	COAHUILA DE ZARAGOZA	5	10.4	15.400
5	COLIMA	6	10.0	14.446
6	CHIAPAS	7	7.8	11.185
7	CHIHUAHUA	8	10.0	17.522
8	CIUDAD DE MEXICO	9	11.5	19.723
9	DURANGO	10	9.7	11.984
10	GUANAJUATO	11	9.0	13.237
11	GUERRERO	12	8.4	10.250
12	HIDALGO	13	9.4	13.077
13	JALISCO	14	9.9	13.887
14	MEXICO	15	10.1	12.612
15	MICHOACAN DE OCAMPO	16	8.6	13.007
16	MORELOS	17	9.8	11.193
17	NAYARIT	18	9.7	12.689
18	NUEVO LEON	19	10.7	19.216
19	OAXACA	20	8.1	11.438
20	PUEBLA	21	9.2	11.602

Figura 4.18: Conjunto de datos escolaridad y salario.

Una vez indicando con que variables se desea trabajar se procede a dividir el conjunto de datos en dos partes, entrenamiento y prueba. Los parámetros de entrada que recibe son dos, escolaridad siendo la variable independiente y salario, la cual es la variable que se busca predecir, es decir, la dependiente. Se indica que el 20 % de los datos originales se reserva para pruebas, y el 80 % restante se usa para entrenar el modelo. Se crea el modelo con los valores definidos previamente y al mismo tiempo se realiza su entrenamiento, para realizar las predicciones con los datos de prueba, tal como se muestra en la Figura 4.19.

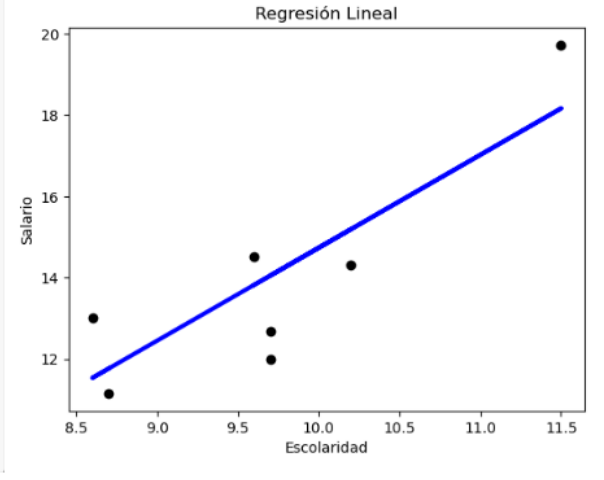
```
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(dfx[['Escolaridad']], dfx['Salario'], test_size=0.2, random_state=42)

# Crear el modelo de regresión lineal
model = LinearRegression()

# Entrenar el modelo con los datos de entrenamiento
model.fit(X_train, y_train)
```

LinearRegression
LinearRegression()

```
# Hacer predicciones con los datos de prueba
y_pred = model.predict(X_test)
```

Figura 4.19: Modelo de regresión lineal.

Gráficamente, en la parte inferior se coloca la función con la cual se obtienen las predicciones del salario con base en el grado de escolaridad. Por ejemplo, en la Figura 4.20 se muestra que para una escolaridad de 16 (equivalente al último año de nivel superior), el salario mensual sería de 28,454 pesos.




```
esc = np.array([[16]])
predicted_sal = model.predict(esc)
print(f"Para Escolaridad {esc[0][0]} {predicted_sal[0]} Salario")
```

```
Para Escolaridad 16 28.454916666666662 Salario
```

Figura 4.20: Datos predichos respecto al salario con base en la escolaridad.

La precisión del modelo se obtiene mediante las funciones que se observan en la Figura 4.21, especialmente con el coeficiente de determinación (R^2), el cual en este caso indica que aplicando la técnica de regresión lineal al conjunto

de datos seleccionado da lugar a un modelo con 74 % de la variabilidad de los datos, lo que significa que el 74 % de las variaciones en la variable dependiente pueden ser explicadas por el modelo a partir de las variables independientes utilizadas en la regresión, mostrando una diferencia significativa respecto a los algoritmos explicados previamente.

```python
# Calcular el R²
r2 = r2_score(y_test, y_pred)

# Calcular el MSE y el RMSE
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

# Calcular el MAE
mae = mean_absolute_error(y_test, y_pred)

# Imprimir resultados
print(f"R²: {r2:.2f}")
print(f"MSE: {mse:.2f}")
print(f"RMSE: {rmse:.2f}")
print(f"MAE: {mae:.2f}")
```

```
R²: 0.74
MSE: 1.77
RMSE: 1.33
MAE: 1.24
```

Figura 4.21: Métricas calculadas.

Las variables consideradas para realizar el análisis de regresión pertenecen a las siguientes categorías.

Categorías	**Variables**
SEGURIDAD	Homicidios Robos Secuestros Tráfico de personas

EDUCACIÓN	Escolaridad Becas Abandono Escolar
SALUD	Casos Covid Adicciones
EMPLEO	Asegurados Desempleo Salario
ECONOMÍA	Producto Interno Bruto Apoyo a adultos mayores Inflación

Tabla 4.1: Variables poblacionales.

Para visualizar la manera en la que se relacionan los valores de las variables, se puede considerar el uso de una matriz de correlación, en la cual se muestra el coeficiente de correlación para cada par de variables, esto es posible con la función que se muestra en la Figura 4.22.

```
In [8]: correlation_matrix = dfm.corr(method='pearson')
        print(correlation_matrix)
```

Figura 4.22: Línea de código para generar matriz de correlación.

Una vez aplicada la función, resulta una cantidad de valores que abarcan del -1 hasta el 1, refiriendose al nivel de correlación que existe tal como se observa en la Figura 4.23. Entre más cerca esté del -1 y del 1 indica que la correlación es mucho mayor, la diferencia es que para -1 es correlación negativa y para 1 es correlación positiva, mientras que para valores cercanos al 0 la correlación es baja o incluso hasta nula [62, 63].

	Homicidios	Robos	Trafico_personas	Secuestros	\
Homicidios	1.000000	0.619027	0.423594	0.434804	
Robos	0.619027	1.000000	0.390335	0.663190	
Trafico_personas	0.423594	0.390335	1.000000	0.407707	
Secuestros	0.434804	0.663190	0.407707	1.000000	
PIB	0.510719	0.796318	0.420857	0.533902	
Inflacion	0.229586	0.070935	-0.085383	0.000573	
Escolaridad	-0.156836	0.286272	0.243043	-0.012866	
Abandono_escolar	-0.069293	-0.386003	-0.147659	-0.007971	
Becas	0.247356	0.572750	0.068855	0.430142	
Desempleo	0.384035	0.464991	0.168075	0.303305	
Salario	0.039253	0.175754	0.135188	-0.142817	
Asegurados	0.711422	0.869644	0.499123	0.741076	
Apoyo_adultos_mayores	0.609994	0.550906	0.220037	0.671593	
Adicciones	0.143065	0.046331	0.115393	0.016931	
Casos_Covid	0.282214	0.697697	0.171086	0.505859	

Figura 4.23: Matriz de correlación.

En este caso se utilizó la función heatmap de la librería Seaborn y el código para generarlo se muestra en la Figura 4.24.

```
In [11]: # Generar el mapa de calor
         plt.figure(figsize=(10, 8))
         sns.heatmap(correlation_matrix, annot=True, cmap='viridis')
         plt.title('Mapa de Calor')
         plt.show()
```

Figura 4.24: Código para generar mapa de calor.

En Seaborn se puede usar la función heatmap() para representar un mapa de calor a partir de los datos de una matriz de correlación. Para obtener la matriz a partir del conjunto de datos se empleó el método corr() como se mostró en la Figura 4.22.

El mapa de calor parte de la matriz de correlación mostrada en la Figura 4.23, aplicando dicha herramienta se puede observar de una mejor manera los valores de la matriz. En estos gráficos, cada valor de la matriz se representa en una celda con un color distinto. Las celdas que muestran valores de correlación cercanos a 1, lo que indica una correlación positiva fuerte, se representan con tonos cálidos (amarillos). Por otro lado, las celdas con valores de correlación cercanos a -1, indicando una fuerte correlación negativa, se representan con tonos fríos (azules). Los valores cercanos a 0, que indican una ausencia de correlación, se muestran en tonos intermedios (color verde turquesa). El mapa de calor se muestra en la Figura 4.25.

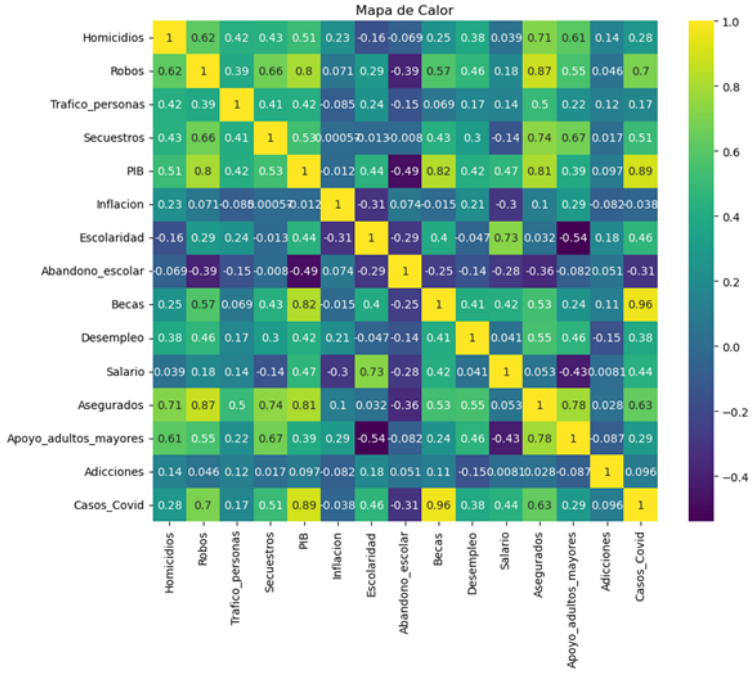


Figura 4.25: Mapa de calor creado con Seaborn.

Para este caso práctico, de acuerdo al mapa de calor, las variables que cuentan con una correlación alta son: Escolaridad – Salario, PIB – Robos, PIB – Becas, PIB – Asegurados, Asegurados – Secuestros y Asegurados – Robos. Es decir, alguna de las dos variables se ve influenciada por la otra. Para tal situación la técnica de ciencia de datos que se aplicó fue regresión lineal simple.

En el siguiente ejemplo se aplica sobre el nivel de correlación que existe entre producto interno bruto y robos.

Para ello es necesario la importación de algunas librerías especializadas en regresión lineal, tal como se indica en la Figura 4.26.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from sklearn.linear_model import LinearRegression
        from sklearn.model_selection import train_test_split
```

Figura 4.26: Librerías utilizadas.

Posteriormente, se carga el conjunto de datos como se aprecia en la Figura 4.27. Respecto a las variables seleccionadas, el dataframe generado se muestra en la Figura 4.28 para su posterior entrenamiento.

```
In [2]: df = pd.read_csv("PIB - Robos Corr.csv")
        df
```

Figura 4.27: Línea de código para leer datos en formato CSV.

Out[2]:

	Entidad	Clave	PIB	Robos
0	AGUASCALIENTES	1	1.31	14943
1	BAJA CALIFORNIA	2	3.86	50242
2	BAJA CALIFORNIA SUR	3	0.74	10725
3	CAMPECHE	4	1.74	839
4	COAHUILA DE ZARAGOZA	5	4.03	10240
5	COLIMA	6	0.61	6603
6	CHIAPAS	7	1.56	8805
7	CHIHUAHUA	8	3.72	16945
8	CIUDAD DE MEXICO	9	14.22	102678
9	DURANGO	10	1.29	11019
10	GUANAJUATO	11	4.69	39426
11	GUERRERO	12	1.30	9820
12	HIDALGO	13	1.75	14146

Figura 4.28: Dataframe respecto a PIB – robos.

En la Figura 4.29 se muestran los pasos aplicados para la técnica de regresión lineal simple. El primer paso fue la separación del conjunto de datos, mediante la función train_test_split de sklearn.model_selection para dividir los

datos en conjuntos de entrenamiento y prueba. La variable dfx es un DataFrame que contiene las columnas 'PIB' y 'Robos'. Por otra parte, test_size=0.2 indica que el 20 % de los datos se usaron para pruebas, y random_state=42 asegura que la división sea reproducible.

Posteriormente se creó una instancia del modelo de regresión lineal utilizando la clase LinearRegression de sklearn.linear_model.

La tercera línea de código se usó para entrenar el modelo de regresión lineal utilizando los datos de entrenamiento. X_train contiene los valores del PIB, mientras que y_train contiene los valores correspondientes de robos.

Una vez entrenado el modelo, se utilizaron los datos de prueba (X_test) para hacer predicciones de los valores de robos (y_pred).

Finalmente, estas líneas se encargaron de graficar los resultados:

- plt.scatter(X_test, y_test, color='black') crea un diagrama de dispersión de los datos reales de prueba.
- plt.plot(X_test, y_pred, color='blue', linewidth=3) dibuja la línea de regresión utilizando las predicciones del modelo.
- plt.title('Regresión Lineal') establece el título del gráfico.
- plt.xlabel('PIB') y plt.ylabel('Robos') etiquetan los ejes X e Y respectivamente.
- plt.show() muestra el gráfico

```
In [4]: # Dividir los datos en conjuntos de entrenamiento y prueba
        X_train, X_test, y_train, y_test = train_test_split(dfx[['PIB']], dfx['Robos'], test_size=0.2, random_state=42)

In [5]: # Crear el modelo de regresión lineal
        model = LinearRegression()

In [6]: # Entrenar el modelo con los datos de entrenamiento
        model.fit(X_train, y_train)

Out[6]: LinearRegression
        LinearRegression()

In [7]: # Hacer predicciones con los datos de prueba
        y_pred = model.predict(X_test)

In [8]: # Graficar los resultados
        plt.scatter(X_test, y_test, color='black')
        plt.plot(X_test, y_pred, color='blue', linewidth=3)
        plt.title('Regresión Lineal')
        plt.xlabel('PIB')
        plt.ylabel('Robos')
        plt.show()
```

Figura 4.29: Sección de código referente a regresión lineal simple.

División de los datos: Los datos se dividieron en conjuntos de entrenamiento y prueba, utilizando el 20 % de los datos para pruebas.

Creación y entrenamiento del modelo: Se creó y entrenó un modelo de regresión lineal con los datos de entrenamiento.

Predicciones: Se realizaron predicciones utilizando los datos de prueba.

Visualización: Se graficaron los resultados para mostrar cómo el modelo predice el número de robos en función del PIB.

Escolaridad y Salario: En Figura 4.30 se muestra el resultado de la relación entre escolaridad y salario, donde los puntos negros representan los datos reales de prueba y la línea azul muestra las predicciones del modelo. Este diagrama de dispersión en particular cuenta con algunos valores atípicos, que podrían haber presentado problemas con la predicción, pero no fue así de acuerdo con la tendencia de los datos.

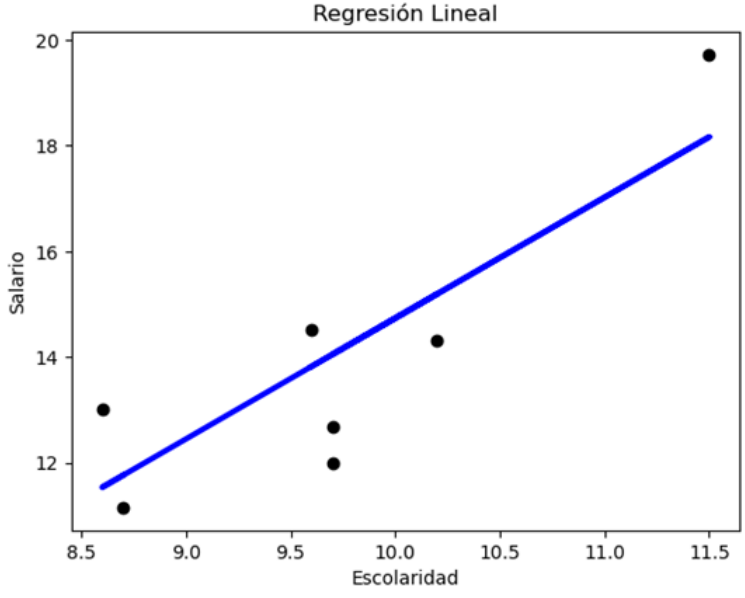


Figura 4.30: Diagrama de dispersión Escolaridad – salario.

Interpretación

- Valores Observados: Los registros (8 a 11) representan datos reales del salario asociado a ciertos niveles de escolaridad.

- Valores Predichos: Los anteriores (4 a 6) y siguientes registros de escolaridad (13 a 16) son valores predichos por un modelo de regresión lineal simple que predice el salario basado en la escolaridad, como se aprecia en la Figura 4.31.

Tendencia: La relación entre escolaridad y salario muestra una tendencia ascendente, lo que indica que a mayor nivel de escolaridad, mayor es el salario.

Los valores predichos son consistentes con la tendencia observada en los datos y sugieren que el modelo de regresión está capturando de forma adecuada la relación entre escolaridad y salario tal como se observa en la Figura 4.31.

Con un grado de escolaridad 4 (equivalente a cuarto año de primaria) se estima que el sueldo para una persona va a ser menor que 5,000. En otro caso, se estima que una persona con un grado de escolaridad de 16 (equivalente al último año de estudios universitarios) su ingreso sería cerca de 30,000 de acuerdo con las predicciones realizadas en el modelo de regresión lineal. Esto se observa en la Figura 4.31.

Figura 4.31: Diagrama de predicción Escolaridad – salario.

PIB y Robos

La gráfica resultante se aprecia en la Figura 4.32 que muestra la relación entre el PIB y el número de robos, donde los puntos negros representan los datos reales de prueba y la línea azul muestra las predicciones del modelo.

El punto que se encuentra muy alejado de los demás es un outlier o valor atípico. Los outliers son datos que se encuentran a una distancia considerable

del resto de los puntos y pueden tener un impacto significativo en el análisis y los resultados de los modelos de regresión. En este caso, el outlier es el punto con un PIB de 14.22 y 102,678 robos.

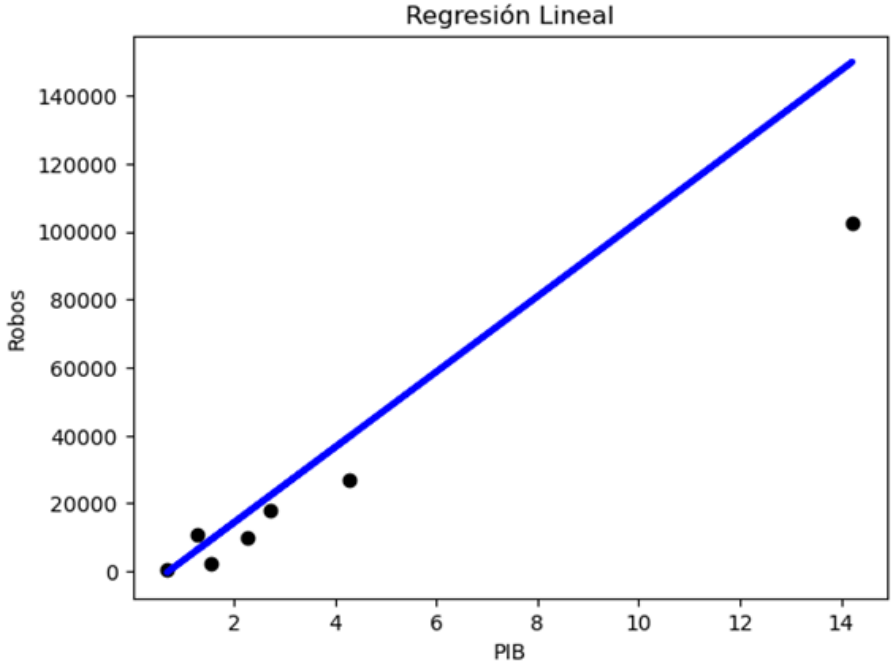


Figura 4.32: Diagrama de dispersión PIB – robos.

Interpretación

- Valores observados: Los primeros cuatro registros representan datos reales del número de robos asociados a ciertos valores del PIB.
- Valores predichos: Los siguientes cuatro registros (de PIB 15 a 18) son valores predichos por el modelo de regresión lineal simple que se entrenó anteriormente, como se muestra en la Figura 4.33.

Tendencia: Los valores predichos muestran una tendencia ascendente en el número de robos a medida que aumenta el PIB, siguiendo la línea de regresión ajustada.

Para estados de la república mexicana con una contribución cercana al 5 por ciento (como es el caso de Coahuila, Guanajuato y Veracruz), el número de robos efectuados es menor de 5,000 según los datos reales. En cambio, para estados con una contribución cercana a 15 por ciento (un claro ejemplo es la ciudad de México que está cerca), se estima que el número de robos incremente a un poco más de 15,000 tal como se muestra en la Figura 4.33.

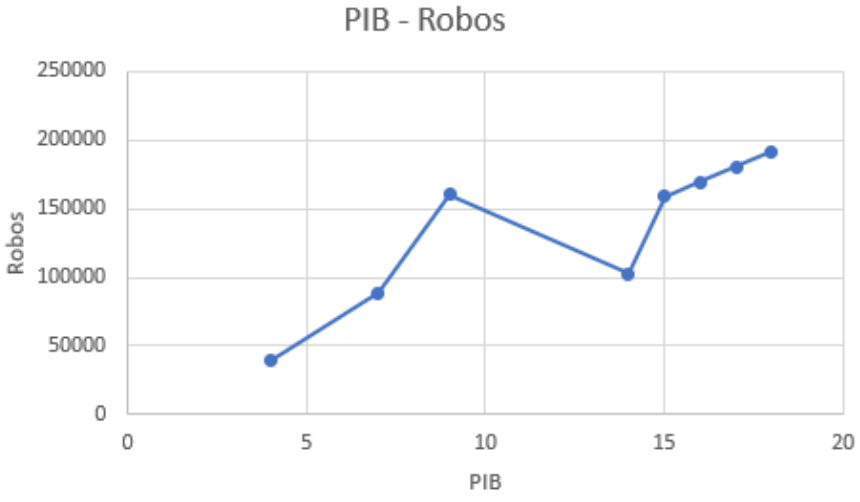


Figura 4.33: Diagrama de predicción PIB – robos.

PIB y Becas

En la Figura 4.34 se muestra el diagrama de dispersión respecto a la predicción en relación de PIB y becas, la mayor parte de los datos reales de prueba se encuentran cerca de la línea azul que representa las predicciones del modelo, lo que indica que se cuenta con valores aptos para la predicción.

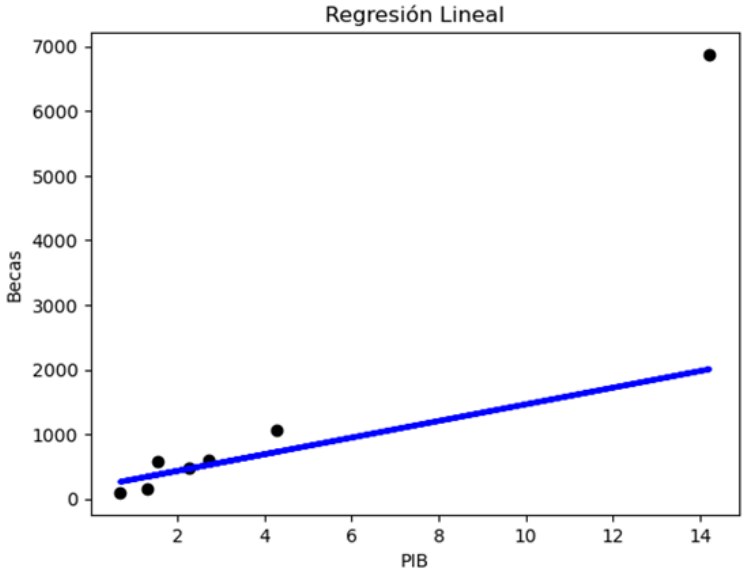


Figura 4.34: Diagrama de dispersión PIB – becas.

Interpretación

- Valores observados: Los primeros cuatro registros representan datos reales del número de becas asociados a ciertos valores del PIB.
- Valores predichos: Los siguientes cuatro registros (de PIB 15 a 18) son valores predichos por el modelo de regresión lineal simple que se entrenó anteriormente, como se muestra en la Figura 4.35.

Tendencia: Los valores predichos muestran una tendencia ascendente en el número de robos a medida que aumenta el PIB, este patrón sugiere que, a medida que la economía crece, la capacidad para ofrecer becas también aumenta de manera considerable. El crecimiento en el número de becas no es lineal en los valores más bajos de PIB (específicamente entre 3 y 7), pero se vuelve más consistente a partir de un PIB de 8 tal como se evidencia en la Figura 4.35.

Para el estado que pueda obtener una contribución cerca del 17 por ciento, el número de becas otorgadas aumentaría cerca de 2,500 como se observa en la Figura 4.35.

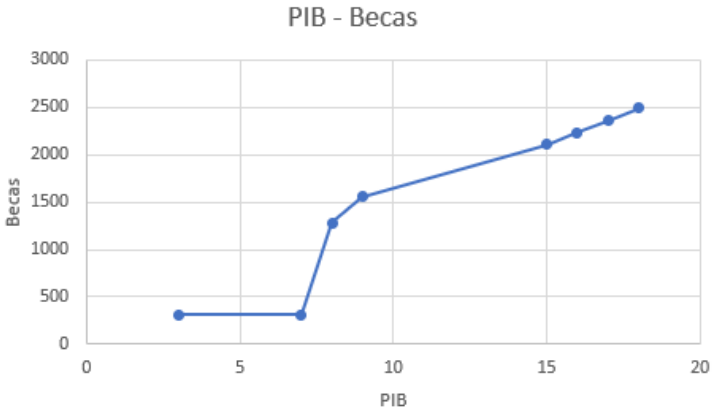


Figura 4.35: Diagrama de predicción PIB – becas.

PIB y Asegurados. El diagrama de dispersión que se visualiza en la Figura 4.36, pertenece a los datos predichos sobre asegurados con base en el producto interno bruto. Los datos reales de prueba se encuentran dentro del rango de las predicciones.

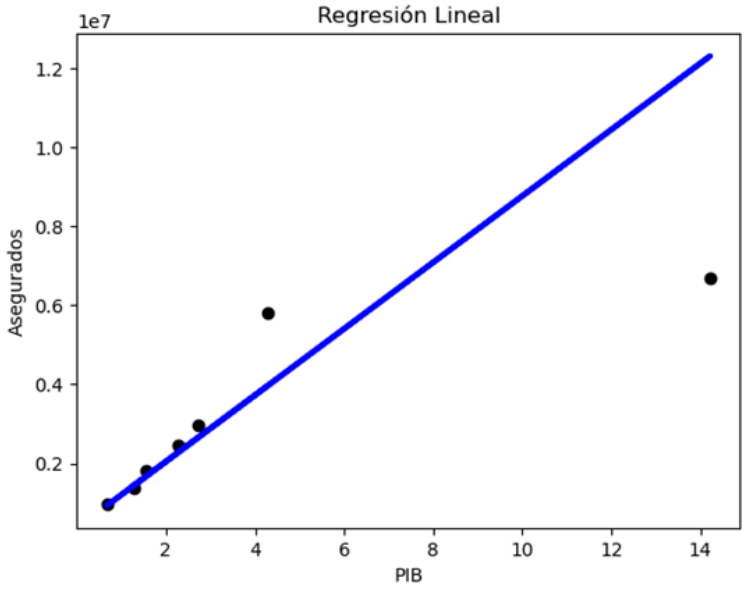


Figura 4.36: Diagrama de dispersión PIB – asegurados.

Interpretación

- Valores observados: Los primeros cuatro registros (PIB 1 a 14) representan datos reales del número de personas aseguradas asociados a ciertos valores del PIB.
- Valores predichos: Los siguientes cuatro registros (PIB 16 a 19) son valores predichos por el modelo de regresión lineal simple que se entrenó anteriormente, como se observa en el diagrama de la Figura 4.37.

Tendencia: Se observa un aumento en el número de asegurados conforme aumenta el PIB, es decir, la mejora en el PIB tiene un impacto considerable en la cantidad de personas aseguradas, reflejando una mejora en la capacidad económica para acceder a seguros.

El crecimiento en el número de asegurados no es lineal, hay ciertos puntos donde el aumento es más pronunciado, especialmente entre los PIB de 1 a 2 y de 14 a 16. A partir de un PIB de 16, el crecimiento en el número de asegurados sigue siendo significativo, pero más estable. Para entidades federativas que

logren una contribución de 16 por ciento se estima que el número de asegurados sería alrededor de 14 millones, tal como se observa en la Figura 4.37.

Figura 4.37: Diagrama de predicción PIB – asegurados.

Asegurados y Secuestros

El siguiente diagrama de dispersión consta de datos predichos respecto a secuestros con base en el número de asegurados, a comparación de los diagramas anteriores el que se muestra en la Figura 4.38, los datos reales de prueba están ligeramente dispersos de la predicción del modelo.

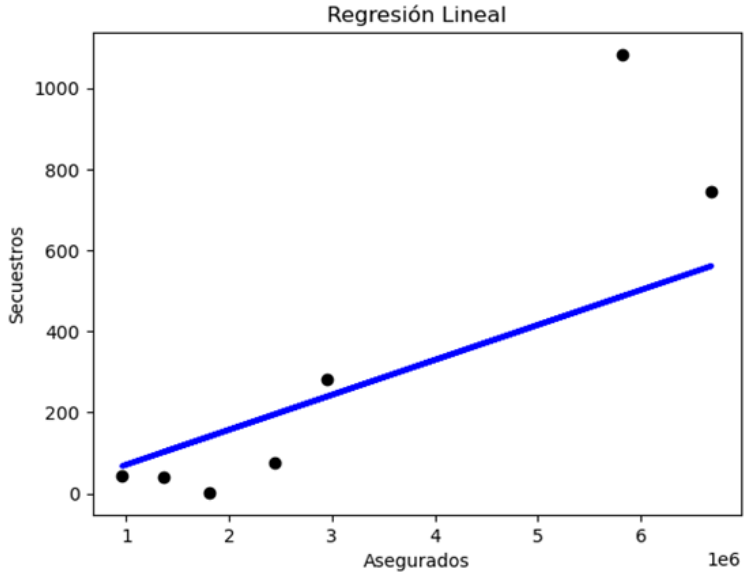


Figura 4.38: Diagrama de dispersión Asegurados – secuestros.

Interpretación

- Valores observados: Los primeros cuatro registros (2 millones a 11 millones) representan datos reales del número de secuestros asociados a ciertos valores del número de asegurados.
- Valores predichos: Los siguientes cuatro registros (13 millones a 16 millones) son valores predichos por el modelo de regresión lineal simple que se entrenó anteriormente, el comportamiento de los valores predichos se observa en el diagrama de la Figura 4.39.

El Estado de México cuenta con el mayor número de personas aseguradas, por lo que, si continúa creciendo ese valor alrededor de 13 millones de personas con seguro, se estima que el número de secuestros aumentan a un poco más de 1,100 como se evidencia en la Figura 4.39. Por otro lado, los datos reales indican que para 2 millones de asegurados el número de secuestros apenas es de 100, tal es el caso de estados como Baja California y Coahuila.

Figura 4.39: Diagrama de predicción Asegurados – secuestros

Asegurados y Robos

El último diagrama de dispersión consta de datos predichos respecto a robos con base en el número de asegurados, tal diagrama se representa en la Figura 4.40, que consta de un par de valores atípicos, pero a pesar de ello los valores predichos se mantienen lineal a los datos reales.

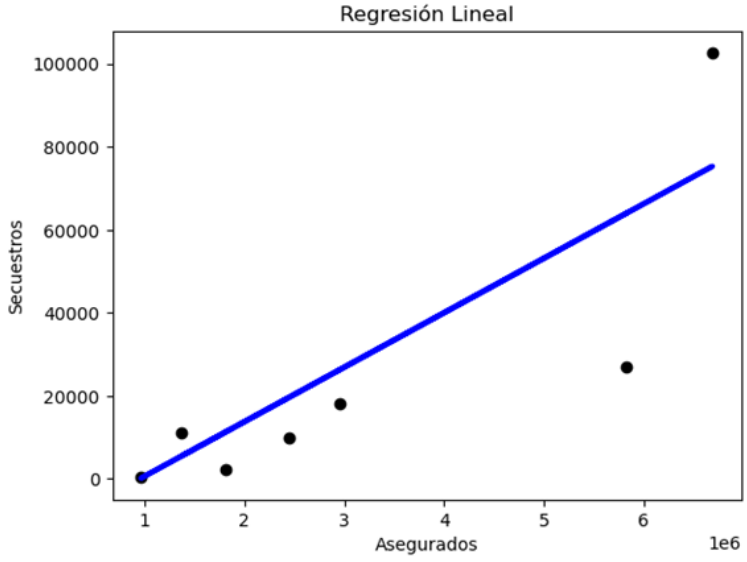


Figura 4.40: Diagrama de dispersión Asegurados – robos.

Interpretación

- Valores observados: Los primeros cuatro registros (2 millones a 11 millones) representan datos reales del número de robos asociados a ciertos valores del número de asegurados.
- Valores predichos: Los siguientes cuatro registros (13 millones a 16 millones) son valores predichos por el modelo de regresión lineal simple que se entrenó anteriormente, el comportamiento de los datos reales y los predichos se muestra en el diagrama de la Figura 4.41.

Otro estado que cuenta con un alto número de personas aseguradas es la Ciudad de México, en algún punto, si la capital logra llegar a 14 millones de asegurados, se estima que el número de robos efectuados sobrepasaron los 150,000, dicha predicción se aprecia en la Figura 4.41. Para alrededor de 3 millones de asegurados, en estados como San Luis Potosí, Sinaloa y Sonora, los robos apenas llegan a 5,000 según los datos reales.

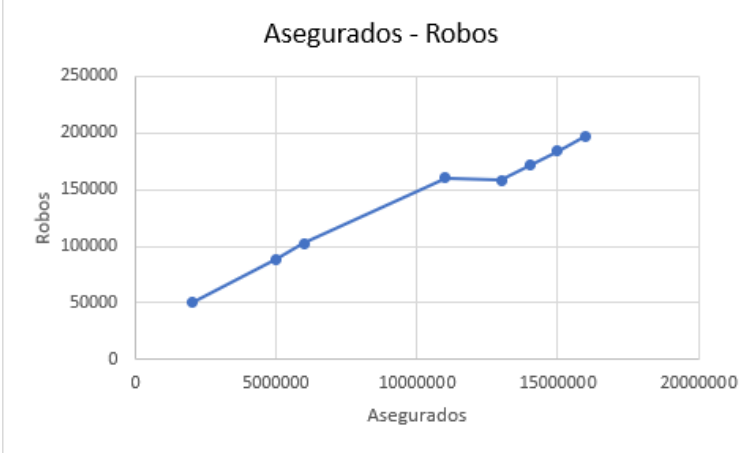


Figura 4.41: Diagrama de predicción Asegurados – robos.

Como se puede observar en la Figura 4.41, los robos aumentan conforme al número de personas aseguradas.

4.2. Interfaces de la plataforma DATAMEX

En esta sección se presentan las interfaces desarrolladas. Cabe mencionar que el diseño web responsivo es un enfoque que propone que el diseño y el desarrollo se adapten al comportamiento y al entorno del usuario, teniendo en cuenta el tamaño de la pantalla, la plataforma y la orientación del dispositivo.

Esta práctica combina cuadrículas y diseños flexibles, imágenes adaptables y un uso eficiente de consultas de medios CSS. Así, cuando un usuario cambia de su computadora portátil a una tableta, el sitio web debe ajustarse automáticamente para adaptarse a la resolución, tamaño de imagen y capacidades de scripting. También, es importante considerar la configuración de los dispositivos del usuario; por ejemplo, si tienen una VPN en su iPad, el sitio web no debería bloquear el acceso a la página. En resumen, el sitio web debe ser capaz de responder automáticamente a las preferencias del usuario, eliminando la necesidad de un diseño y desarrollo específicos para cada nuevo dispositivo en el mercado [66].

El sitio web desarrollado tiene como objetivo brindar información relevante a la población sobre temas de interés común, lo que acoplarlo a diferentes dispositivos es primordial para que se cumpla. En la Figura 4.42 se presenta

la página principal de DATAMEX adaptada al entorno del usuario, en este caso se tomó como ejemplo las dimensiones de un iPhone SE, el diseño y el desarrollo del sitio web sigue manteniéndose a pesar del cambio de dispositivo.

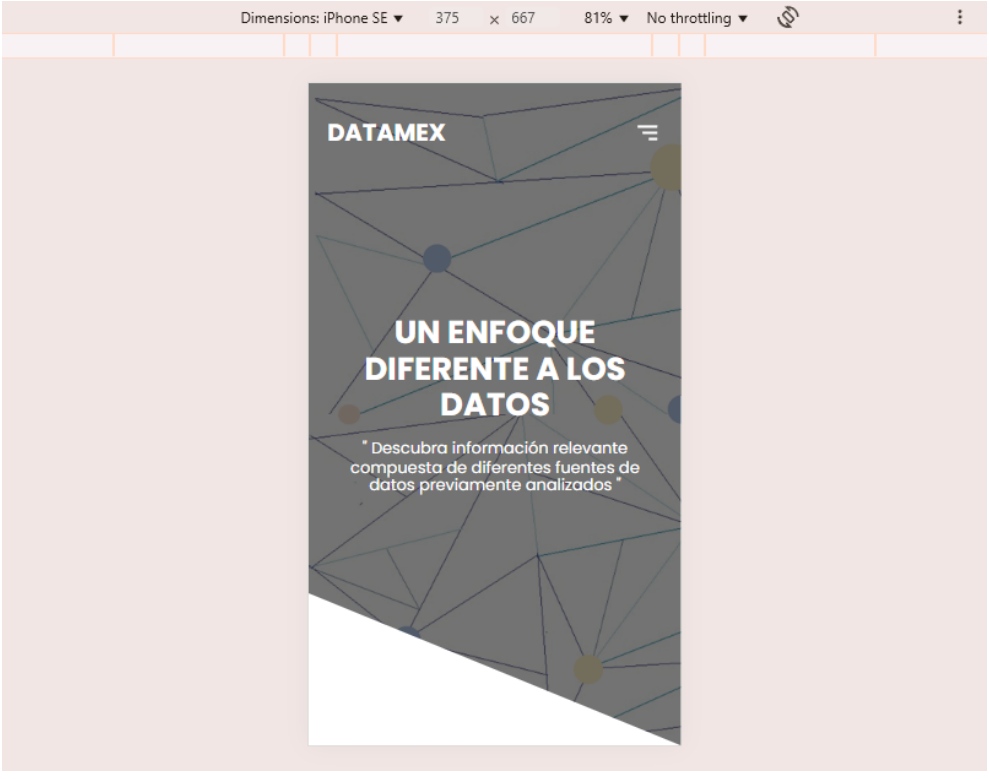


Figura 4.42: Página principal de manera responsive.

Como se mencionó anteriormente, el diseño y el desarrollo en la sección "Predicción de datos" se adaptó al comportamiento y al entorno del usuario, sin presentar problemas en la navegabilidad tal como se observa en la Figura 4.43.

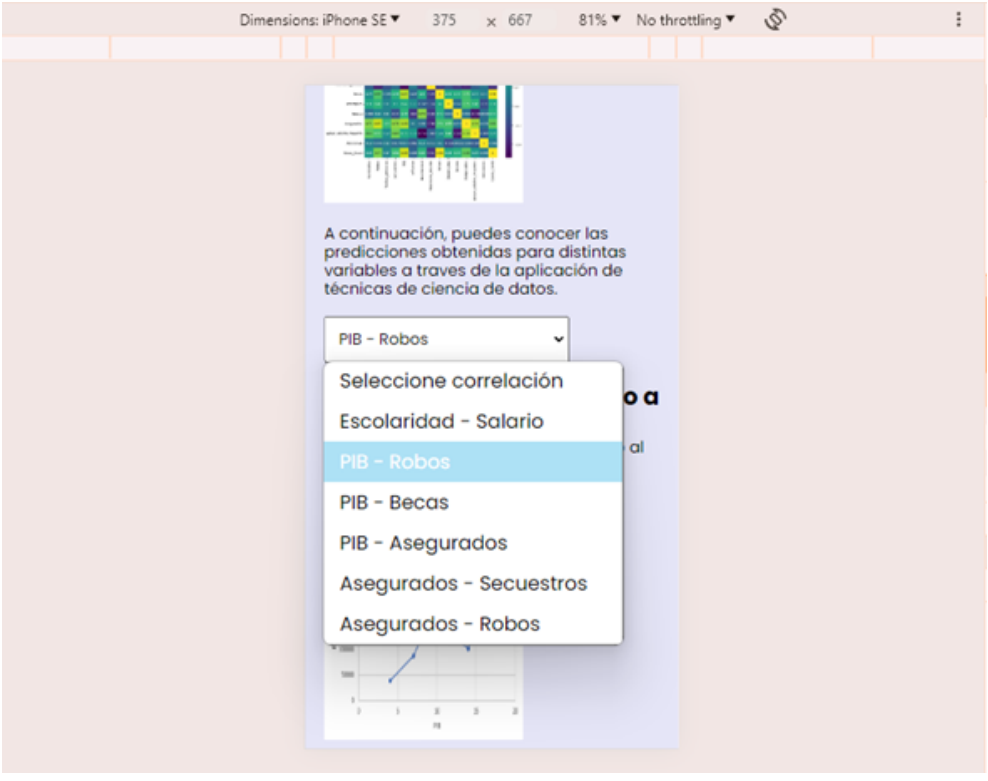


Figura 4.43: Sección "Predicción de datos" de manera responsive.

Los elementos que conforman la estructura del HTML tales como imágenes, párrafos, iconos, entre otros cambian su definición al momento que el usuario navega en algún dispositivo móvil gracias al diseño responsivo como se evidencia en la Figura 4.44.

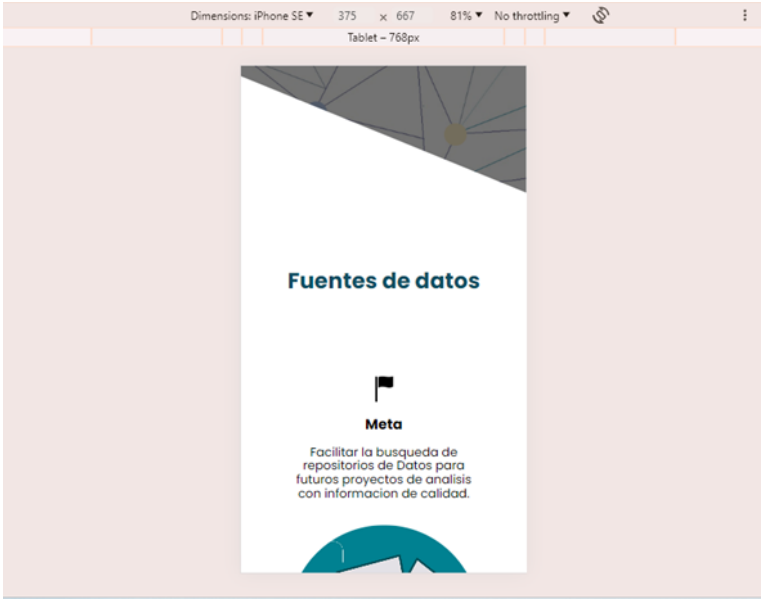


Figura 4.44: Sección “Fuentes de consulta” de manera responsive.

Cabe mencionar que el diseño de la experiencia del usuario (UX) es un proceso empleado por los equipos de diseño para desarrollar productos que ofrezcan experiencias significativas y relevantes a los usuarios. Este enfoque abarca todo el proceso de adquisición e integración del producto, incluyendo elementos de branding, diseño, usabilidad y funcionalidad.

La Organización Internacional de Normalización (ISO) define la experiencia del usuario como:

”Las percepciones y respuestas de una persona que resultan del uso o uso anticipado de un producto, sistema o servicio”.

— ISO 9241-210, Ergonomía de la interacción hombre-sistema—Parte 210: Diseño centrado en el ser humano para sistemas interactivos [67].

En las siguientes figuras se representan las interfaces de DATAMEX, la Figura 4.45, la Figura 4.46 y la Figura 4.47 son ejemplos de la página principal, en la Figura 4.45 se observa en la parte superior el menú de navegación que consta de 3 secciones más, y el nombre del sitio web, al igual que una frase referente al proyecto desarrollado ubicada en el encabezado de cada sección que conforman DATAMEX.

Figura 4.45: Encabezado de DATAMEX.

En la Figura 4.46 se muestran 3 secciones donde se explica la meta, la metodología y las herramientas empleadas para el desarrollo del proyecto de análisis de datos.

Figura 4.46: Acerca del proyecto.

Por último, en la parte inferior de la página principal se da una introducción a cada opción del menú de navegabilidad, en cada recuadro se encuentra un icono azul en forma de flecha ubicado a la derecha para que al seleccionarlo se visualice la información correspondiente como se evidencia en la Figura 4.47.

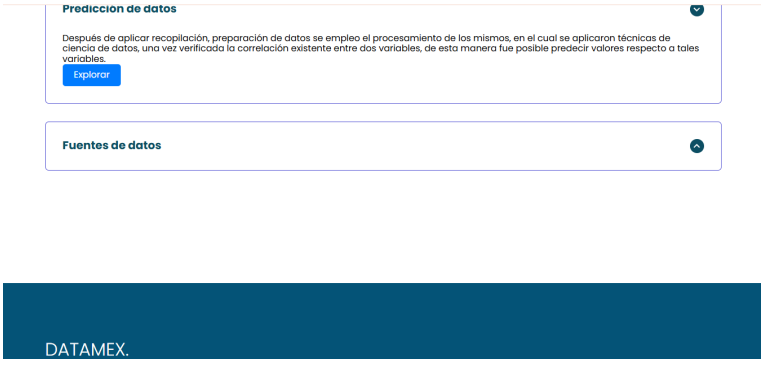


Figura 4.47: Introducción a las secciones de DATAMEX.

En la siguiente sección "¿Qué nos dice los datos?" se muestra información relevante de manera gráfica sobre cada tema de interés para la población, seguridad, educación, economía, salud y empleo. Cada gráfica presentada cuenta con su respectiva explicación a un costado tal como se evidencia a partir de la Figura 4.49, de tal manera será posible comparar fácilmente los resultados obtenidos a nivel estatal o de manera anual. En la Figura 4.48 se muestra el primer vistazo de esta sección con su respectiva meta.

Figura 4.48: ¿Qué nos dicen los datos?

El contenido de esta sección se va mostrado según la relevancia del tema según la encuesta aplicada anteriormente, por lo que las representaciones gráficas respecto a seguridad se presentan a continuación, homicidios, desapa-

riciones, robos y secuestros. Los datos respecto a homicidios se aprecian en la Figura 4.49, de igual manera se observa una explicación con la finalidad de interpretar mejor la gráfica de barras color azul en este caso. Para continuar visualizando el contenido de los siguientes temas, se empleó una función que facilita la navegabilidad, representada con el icono de flecha color negro, tal icono se ubica en los extremos.

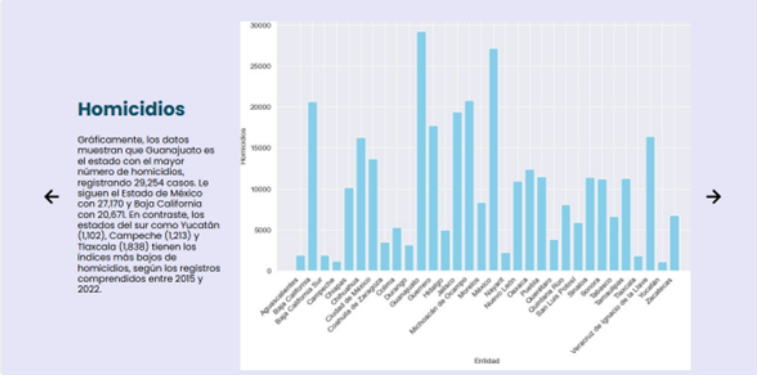


Figura 4.49: Datos respecto a homicidios.

Los datos respecto a Desaparecidos se dividen en dos casos, el número de desaparecidos a nivel estatal conforme se visualiza en la Figura 4.50 mediante una gráfica de barras de color verde con la descripción en donde se resaltan los estados con más incidentes. Por otro lado, en la Figura 4.51 se presenta una gráfica diferente, en la cual se observa el valor de los datos en porcentajes según el género.

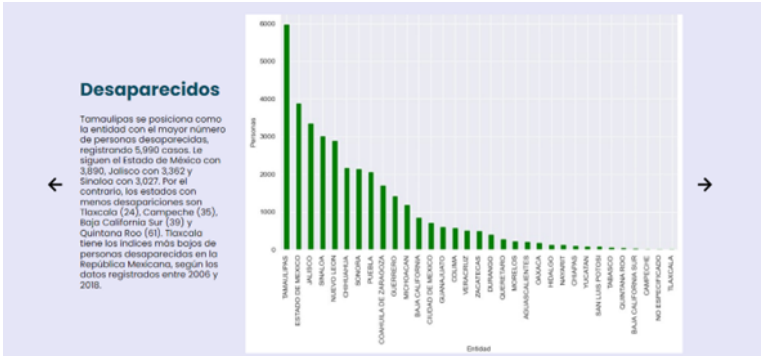


Figura 4.50: Datos respecto a desaparecidos caso 1.

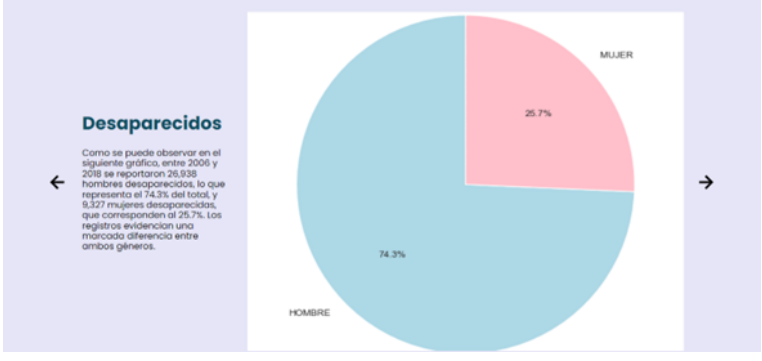


Figura 4.51: Datos respecto a desaparecidos interfaz caso 2.

La siguiente interfaz representa la Figura 4.52, la cual muestra mediante una gráfica de barras color morado los datos relacionados con los robos efectuados en el país, además de su explicación correspondiente.

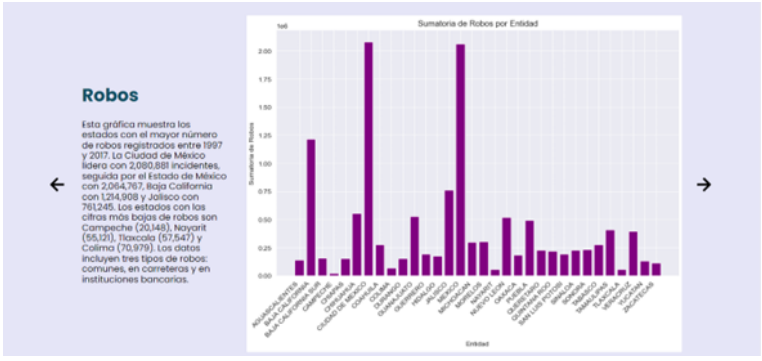


Figura 4.52: Datos respecto a desaparecidos interfaz caso 2.

La última gráfica acerca de seguridad se observa en la Figura 4.53, los datos descritos se basan en el número de secuestros registrados y representados en una gráfica de barras color naranja junto a su descripción.

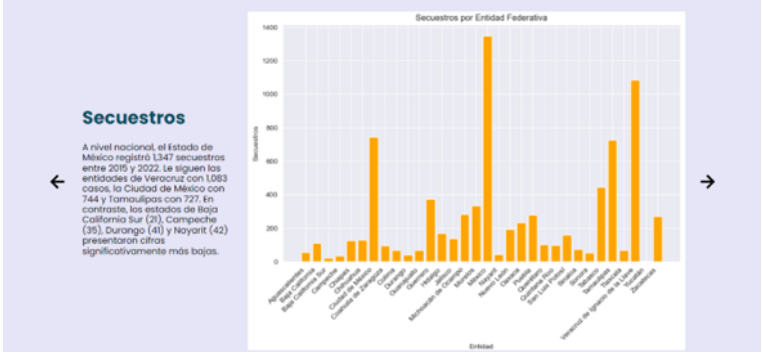


Figura 4.53: Datos respecto a secuestros.

El siguiente tema de interés es Educación, por lo cual se generó una gráfica que representa el nivel de escolaridad por entidad federativa tal como se observa en la Figura 4.54, destacando los estados con niveles superiores de escolaridad.

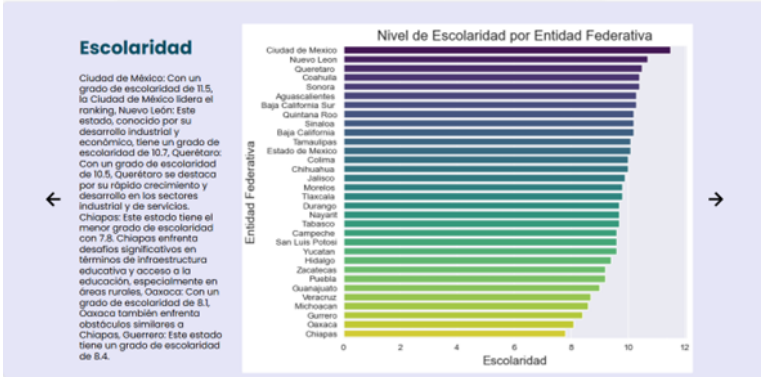


Figura 4.54: Datos respecto a la escolaridad.

El tercer tema es Economía, la gráfica de barras representada en la Figura 4.55 se basa en los porcentajes de contribución que brinda cada estado a la economía del país.

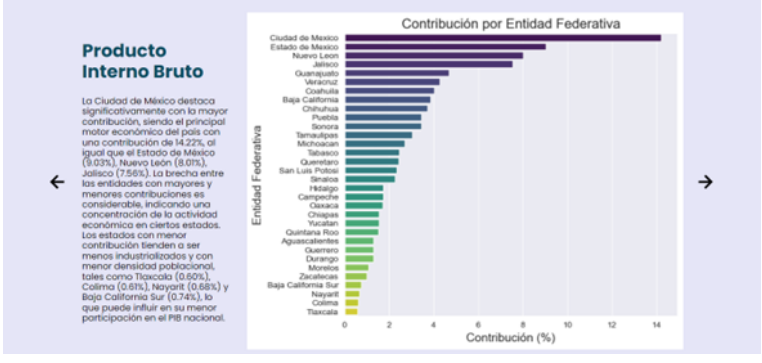


Figura 4.55: Datos respecto al producto interno bruto.

Los datos sobre la salud en México son muy relevantes entre la población, en situación alarmante fue la pandemia de Covid – 19, por lo que las defunciones a causa del virus durante la pandemia en el país se aprecian en la Figura 4.56.

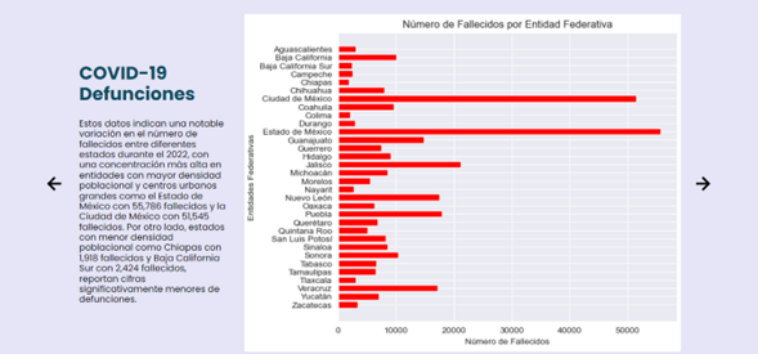


Figura 4.56: Datos respecto a defunciones por covid-19.

Un punto abordado sobre el último tema fue el desempleo, los niveles del mismo basados en el año, se muestran en la Figura 4.57, esta sección consta de la descripción del tema junto a su respectiva gráfica.

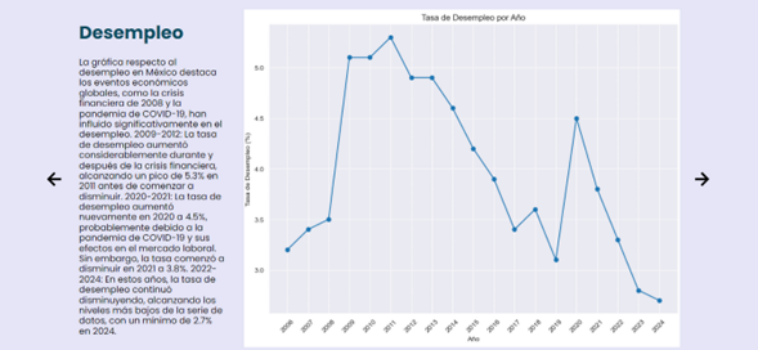


Figura 4.57: Datos respecto al desempleo.

La sección “Predicción de datos” tiene como objetivo ofrecer resultados precisos y relevantes, ayudando a los usuarios a comprender las tendencias y relaciones entre diversos factores que impactan la seguridad, educación, economía, salud y el empleo en México. Una herramienta clave para comprender de dónde parte el análisis de datos es un mapa de calor, tal herramienta se visualiza en la Figura 4.58 y se explica mejor en la parte de resultados previamente mostrada.

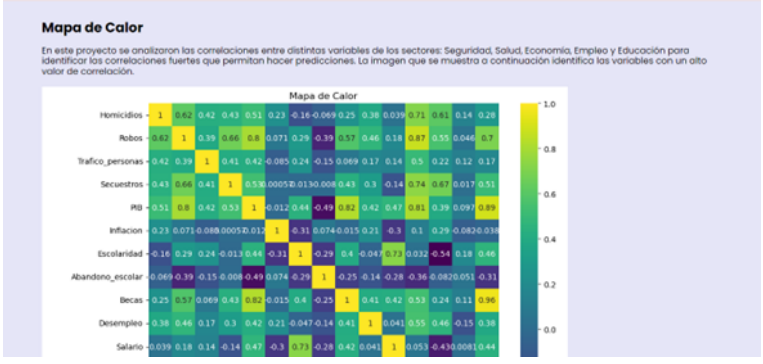


Figura 4.58: Interfaz mapa de calor.

Una vez seleccionado cada par de variables con base en el mapa de calor para calcular su correlación, es posible proceder con la predicción de datos. En la Figura 4.59 se observa una lista desplegable que cuenta con diferentes opciones, tales opciones corresponden al par de variables aptas para aplicar regresión lineal simple.

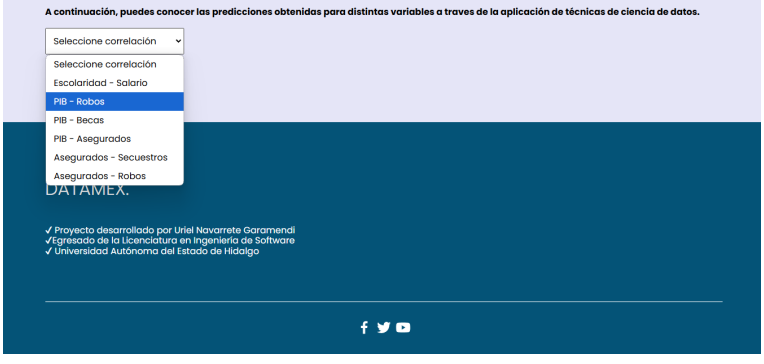


Figura 4.59: Lista desplegable para seleccionar correlación.

En el primer par de variables, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al nivel de escolaridad, que van de la escolaridad 13 a la escolaridad 16, para posteriormente visualizar en pantalla el valor que corresponde a el salario con base en la escolaridad seleccionada. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos

reales registrados y los datos predichos como se evidencia en la Figura 4.60.

Figura 4.60: Interfaz predicción Escolaridad – Salario.

En el siguiente par de variables, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al porcentaje del producto interno bruto, que van de 15 a 18, para posteriormente visualizar en pantalla el valor que corresponde a el número de robos posiblemente efectuados con base en el valor del PIB seleccionado. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos reales registrados y los datos predichos como se evidencia en la Figura 4.61.

Figura 4.61: . Interfaz predicción PIB – Robos.

Para las variables PIB y becas, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al porcentaje del producto interno bruto, que van de 15 a 18, para posteriormente visualizar en pantalla el valor que corresponde a el número de becas posiblemente otorgadas con base en el valor del PIB seleccionado. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos reales registrados y los datos predichos como se evidencia en la Figura 4.62.

Figura 4.62: . Interfaz predicción PIB – Becas.

Para las variables PIB y asegurados, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al porcentaje del producto interno bruto, que van de 16 a 19, para posteriormente visualizar en pantalla el valor que corresponde a el número de personas posiblemente aseguradas con base en el valor del PIB seleccionado. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos reales registrados y los datos predichos como se evidencia en la Figura 4.63.

Figura 4.63: . Interfaz predicción PIB – Asegurados.

Para las variables asegurados y secuestros, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al número de personas aseguradas, que van de 13 millones a 16 millones, para posteriormente visualizar en pantalla el valor que corresponde a el número de secuestros posiblemente efectuados con base en el valor de asegurados seleccionado. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos reales registrados y los datos predichos como se evidencia en la Figura 4.64.

Figura 4.64: Interfaz predicción Asegurados – Secuestros.

Por último, para las variables asegurados y robos, el usuario podrá seleccionar de una lista desplegable un rango de valores respecto al número de personas

aseguradas, que van de 13 millones a 16 millones, para posteriormente visualizar en pantalla el valor que corresponde a el número de robos posiblemente efectuados con base en el valor de asegurados seleccionado. La gráfica que aparece en la parte inferior muestra la tendencia entre los datos reales registrados y los datos predichos como se evidencia en la Figura 4.65.

Figura 4.65: Interfaz predicción asegurados – robos.

La última sección de DATAMEX es "Fuente de datos y tiene como objetivo facilitar la búsqueda de repositorios de datos para futuros proyectos de ciencia de datos con información de calidad. De acuerdo con la categoría seleccionada en la lista desplegable, tal como se observa en la Figura 4.66, se enlistan algunas fuentes de consulta de donde se recopiló información como se aprecia en la Figura 4.67.

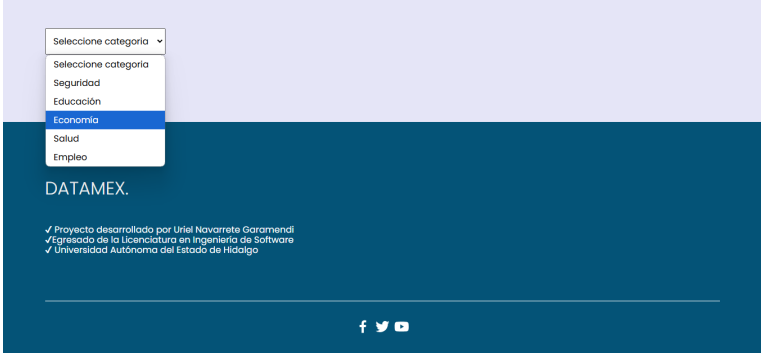


Figura 4.66: Lista desplegable sobre fuentes de datos.

Figura 4.67: Fuentes de consulta de empleo.

4.3. Pruebas de aceptación de usuarios

Se plantearon tres tareas, cada una de ellas con el objetivo de evaluar diferentes procesos que se enfocan en la experiencia del usuario al momento de que se navega por el sitio web. Participaron 7 personas en las pruebas de aceptación de usuarios. Los datos demográficos de estas personas se muestran en la Tabla 4.2

Entidad Federativa	Rango de edad	Sexo
Las personas que participaron son de 2 estados de la República Mexicana. La mayoría son del Estado de Hidalgo (71 %), seguido por la Ciudad de México (19 %)	La población esta entre 19 y 55 años.	Mujer 71 % Hombre 19 %

Tabla 4.2: Datos demográficos de las personas.

Tarea 1: Buscar información y tema especifico

Objetivo: Evaluar la facilidad de búsqueda.

Instrucciones: Busca información de tu interés en la sección "¿Qué nos dicen los datos?".

¿Cuánto tiempo te tomó encontrar la información? (segundos)

Resultado: A excepción de un usuario, que le tomo más tiempo, los demás completaron la actividad en un tiempo aproximado de 2 segundos.

¿Tuviste algún problema? (Sí/No) ¿Cuál?

Resultado: Los usuarios opinaron que no se les presentó ningún problema al momento de navegar a la sección descrita.

Tarea 2: Consulta de datos predichos

Objetivo: Evaluar el proceso referente a la consulta de datos predichos obtenidos para distintas variables.

Instrucciones: Ingresa y consulta en la sección "Predicción de datos" los datos predichos respecto al salario con base en los siguientes niveles de escolaridad:

Para una escolaridad de 16 (equivalente al último año de nivel superior) ¿de cuánto sería el salario mensual en pesos?.

Para una escolaridad de 18 (equivalente a dos años de nivel maestría) ¿de cuánto sería el salario mensual en pesos?.

Resultado: Las respuestas por parte de los usuarios en este ejercicio fueron concretas, el 100 % entendió este proceso y obtuvo los valores predichos del salario con base en el nivel de escolaridad.

¿Encontraste la sección de predicciones fácilmente? (Sí o No)

Resultado: Todos los usuarios señalaron que les resultó sencillo encontrar y navegar en la sección de predicción de datos.

¿Cuánto tiempo te tomó completar el proceso? (...segundos) Resultado: Debido a la interacción que tuvo el usuario en esta tarea, el tiempo total efectuado por los usuarios fue en promedio de 5 segundos.

Tarea 3: Navegar por el sitio web

Objetivo: Evaluar la navegación general y la estructura del sitio.

Instrucciones: Explora cada sección del sitio y encuentra las fuentes de consulta que se consideraron para cada tema mostrado.

¿Te resultó fácil entender la navegación del sitio? (Sí o No)

Resultado: En términos generales, los usuarios argumentaron que la navegación les resultó sencilla, conforme más exploraban el sitio web. Es decir, la curva de aprendizaje fue bastante buena.

Entre las diferentes secciones ¿Cuál de todas te resulto más interesante? y ¿Por qué?:

Resultado: Predicción de datos fue el área que a los usuarios les pareció más interesante, argumentaron que el análisis aplicado les resultaría útil para anticipar comportamientos y a su vez emplear la toma de decisiones con base en información de diferente índole.

De igual forma, los temas que resultaron más atractivos para los usuarios en la sección ¿Qué nos dicen los datos? fueron seguridad, educación y salud, tal como se muestra en la Figura 4.68, debido a que la información expuesta corresponde al área de especialidad de cada uno.

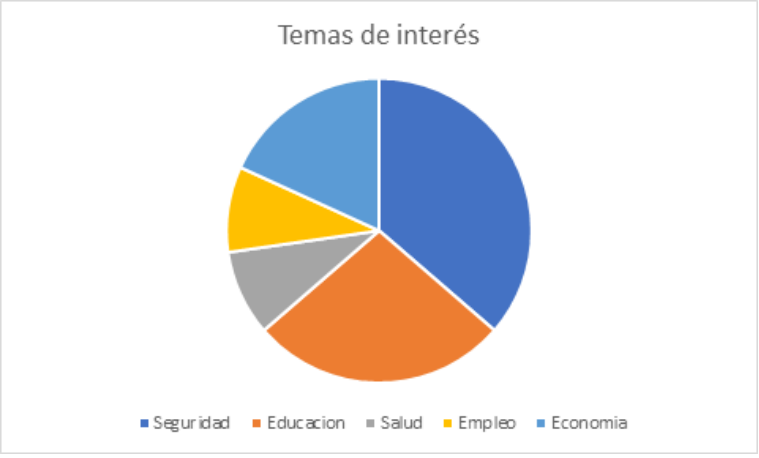


Figura 4.68: Temas de interés en los usuarios.

El tiempo empleado por un usuario en completar cada actividad estuvo entre 1 y 7 segundos. Esto se debe a que no se presentaron complicaciones al navegar, consultar y entender la información solicitada en cada tarea. En resumen, el diseño de la interfaz es amigable e intuitiva, cada sección está bien definida y el usuario identifica fácilmente que clase de información puede encontrar en cada una de ellas.

Escala de satisfacción

En la segunda parte se evaluó la satisfacción del usuario en una escala del 1 al 5. El número 1 representa el puntaje más bajo mientras que el número 5 demuestra el grado más alto de satisfacción, con base en los nueve puntos clave para medir la usabilidad que se enumeran a continuación.

1. Facilidad de navegación
2. Claridad de la información
3. Velocidad del sitio
4. Estética y diseño
5. Facilidad para consultar la información
6. Adaptabilidad en dispositivos móviles

7. Interactividad del sitio
8. Información útil
9. Satisfacción general

En la Figura 4.69 se observan los valores presentados de manera gráfica. En cada caso, el nivel de satisfacción del usuario se observa cercano a 5 que es el valor de satisfacción máximo, lo que muestra que DATAMEX es un sitio web que cumple con las características deseadas para un alto nivel de satisfacción entre los usuarios.

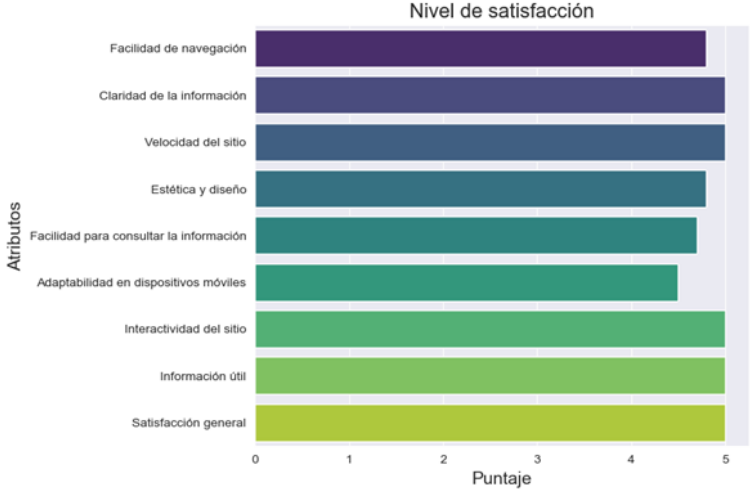


Figura 4.69: Niveles de satisfacción.

Por último, se llevó a cabo la parte de preguntas abiertas, con lo cual se busca obtener una retroalimentación directa por parte de los usuarios. Las preguntas fueron las siguientes.

1. ¿Qué te gustó más del sitio web?

   Respuesta:

   Como respuesta más común por parte de los usuarios, se mencionó que les agrado mucho el tipo de información mostrada y la facilidad para acceder a ella. Además de presentar los datos de una manera clara, les

resulto interesante poder emplear la predicción de los mismos con base a sus intereses. La nube de palabras generada se muestra en la Figura 4.70.

Figura 4.70: Nube de palabras ¿Qué te gustó más del sitio web?

2. Con base en tu experiencia al navegar por el sitio web y el contenido visualizado, ¿cuál es la importancia de los datos para ti?

   Respuesta:

   Los usuarios argumentaron que están conscientes de la importancia del procesamiento de datos para la toma de decisiones, de esta forma se crea conciencia de problemáticas u oportunidades que existen en cada entorno. La nube de palabras generada se muestra en la Figura 4.71.

Figura 4.71: Nube de palabras ¿cuál es la importancia de los datos para ti?

3. ¿Hubo algo que te confundiera respecto al sitio? (Si o No)

   Respuesta:

   Respecto a los problemas presentados, solamente uno de los usuarios manifestó tener confusión en la página de inicio, debido a que en su dispositivo móvil se le complicó navegar a la siguiente sección, pero después entendió la nevegación y continúo explorando.

4. ¿Hay algo que cambiarías o mejorarías?

   Respuesta:

   Todos los usuarios comentaron que actualmente el sitio web cumple con los objetivos planteados en cada sección consultada.

5. ¿Recomendarías este sitio web a un amigo? ¿Por qué?

   Respuesta:

   El 100 % de los usuarios expresaron que si recomendarían el sitio web DATAMEX a sus amigos y conocidos, por la importancia de obtener datos precisos que puede brindar un panorama más amplio a las personas sobre la situación en el país respecto a varias temáticas que sobresalen. La nube de palabras generada se muestra en la Figura 4.72.

Figura 4.72: Nube de palabras ¿Recomendarías este sitio web a un amigo? ¿Por qué?

Capítulo 5

Conclusiones

El presente trabajo de investigación se centró en el análisis de datos en México en áreas de seguridad, educación, economía, salud y empleo, utilizando datos estructurados en formato CSV y datos semiestructurados. Para tal análisis se utilizaron técnicas de ciencia de datos, para posteriormente presentar los resultados en un sistema web desarrollado. Este sistema tiene como finalidad facilitar la consulta de datos por parte de la población y proporcionar una visión más clara de los datos obtenidos, contribuyendo a la toma de decisiones. En un contexto de constante innovación tecnológica y aumento en la generación de datos, el análisis de los mismos se ha convertido en un recurso crucial. La gran cantidad de datos disponibles permite identificar tendencias y resolver problemas en áreas como la banca, finanzas, marketing, salud, empleo, entre otros. La creciente preocupación de la población por factores económicos y sociales refuerza la necesidad de una herramienta que facilite la consulta de información relevante y permita realizar predicciones útiles. Como parte del trabajo, se aplicó una encuesta que mostró el interés de los usuarios en contar con un análisis de datos sobre temas de mayor preocupación, tales como seguridad, salud, educación y economía. Los resultados mostraron una demanda significativa por una herramienta que brinde un análisis detallado y predicciones confiables, lo que respalda la relevancia y necesidad del proyecto. El objetivo fue desarrollar un sistema web que muestre los resultados del análisis de datos poblacionales en México, facilitando su consulta y proporcionando una visión clara y comprensible de los resultados. Para ello fue necesario la identificación y recopilación de datos, limpieza de datos, el desarrollo de un modelo predictivo, la evaluación y validación del modelo y la creación del sistema web para presentar los resultados obtenidos del análisis de datos. La investigación partió de la hipótesis que existen relaciones significativas entre

diversos factores sociales y económicos de interés para la población mexicana. Estas relaciones, una vez identificadas y analizadas, facilitaron la comprensión de los datos y pueden contribuir a una toma de decisiones más informada. Se propuso desarrollar un sistema web con la capacidad de presentar, de una manera clara y accesible el análisis de datos en México mediante técnicas de ciencia de datos que fueron previamente recopilados de diferentes fuentes, preparados y evaluados. La metodología utilizada se basó en el enfoque de IBM para la ciencia de datos. De igual forma, se realizó una revisión exhaustiva de diversas investigaciones centradas en el análisis de datos poblacionales en México. Estos trabajos fueron comparados con los resultados presentados en DATAMEX. Aplicando las fases de identificación y recopilación de datos, limpieza de los mismos, desarrollo de un modelo predictivo, evaluación y validación del modelo abordado, el análisis de regresión lineal mostró una clara relación entre el nivel de escolaridad y el salario, lo que subraya la importancia de la educación en el incremento del ingreso personal, así como una fuerte correlación entre el PIB y factores como el número de becas otorgadas y el número de personas aseguradas. Este análisis no solo valida la hipótesis de que una mayor educación conlleva a mejores salarios, sino que también destaca la utilidad de las técnicas de regresión y visualización para interpretar y comunicar relaciones estadísticas en datos reales. Estos hallazgos pueden ser utilizados para mejorar políticas educativas y laborales, enfatizando la inversión en educación como medio para mejorar los ingresos y potencialmente la calidad de vida. De igual manera, el análisis mostró que el crecimiento económico tiene un impacto positivo en el acceso a la educación y los seguros, mejorando las condiciones de vida en la población. Por otro lado, se demostró que en cuestiones de seguridad, la cantidad de robos se ve influenciada por el crecimiento económico, es decir, entre mayor sea la contribución del estado, el número de incidentes delictivos, en este caso robos, aumenta notablemente. También, se desarrolló el sitio web DATAMEX, cumpliendo con el objetivo principal planteado, que se basa en presentar los resultados del análisis obtenido de una manera clara y accesible para cualquier usuario. Se optó por utilizar Visual Studio Code, un entorno de desarrollo integrado con las características necesarias para trabajar en conjunto con HTML, CSS y JavaScript. El principal lenguaje de programación implementado fue JavaScript, el cual se encargó de incluir interactividad y dinamismo al momento de realizar la consulta de información. Además, el lenguaje de etiquetas HTML, sirvió para dar estructura a todo el contenido como encabezados, párrafos, imágenes, enlaces, entre otros. De esta forma, se definió la jerarquía entre ellos, asegurando que cada elemento estuviera ordenado y en el lugar correspondiente. Finalmente, con las hojas de estilo en cascada CSS, fue posible dar el diseño apropiado a los elementos que conforman DATAMEX. CSS permitió gestionar colores de fondo, tipografía y otros aspectos visuales

necesarios para brindar una buena experiencia de usuario. La usabilidad también es un punto que se tomó en cuenta, con las hojas de estilo se integró un diseño web responsivo, con este enfoque aplicado al sitio web es posible que el diseño y el desarrollo se adapten al comportamiento y entorno del usuario, es decir, no importa el dispositivo por el cual el usuario este navegando, cada elemento que conforma DATAMEX se adapta automáticamente a la resolución del entorno. Esta herramienta facilita el acceso a información crítica de manera comprensible, apoyando la toma de decisiones. Este trabajo de investigación ha demostrado la importancia del análisis de datos para mejorar la toma de decisiones en México. La implementación de una plataforma que presente de manera clara predicciones basadas en datos, previamente recopilados, preparados y evaluados podría tener un impacto significativo en diversos sectores, contribuyendo al bienestar y desarrollo del país.

El trabajo a futuro incluye la integración de datos en tiempo real, así como ampliar el alcance de las variables analizadas para proporcionar una visión más completa de la situación que se vive en México.

Bibliografía

[1] N. Aldana Martínez and M. Avellaneda Hortua, "Regresión lineal: Aplicaciones financieras," 2022.

[2] N. Roustaei, "Application and interpretation of linear-regression analysis," *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, vol. 13, no. 3, p. 151, 2024.

[3] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of applied science and technology trends*, vol. 2, no. 01, pp. 20–28, 2021.

[4] J. S. Kushwah, A. Kumar, S. Patel, R. Soni, A. Gawande, and S. Gupta, "Comparative study of regressor and classifier with decision tree using modern tools," *Materials Today: Proceedings*, vol. 56, pp. 3571–3576, 2022.

[5] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of support vector machine algorithm in big data background," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 5594899, 2021.

[6] C. Avcı, M. Budak, N. Yağmur, and F. Balçık, "Comparison between random forest and support vector machine algorithms for lulc classification," *International Journal of Engineering and Geosciences*, vol. 8, no. 1, pp. 1–10, 2023.

[7] G. A. Martinez Gonzalez *et al.*, "Aplicación de ciencia de datos para el análisis de datos de mortalidad por covid-19 de méxico," 2022.

[8] F. Zapata, "Variables estadísticas." `https://www.lifeder.com/variables-estadisticas/`, 2022 [Fecha de último acceso: 04.02.2025].

[9] A. Kuz and R. Morales, "Ciencia de datos educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en

méxico," *Education in the Knowledge Society (EKS)*, vol. 24, pp. e30080–e30080, 2023.

[10] G.-S. Vicente and J. A. Leyva Moreno, "La probabilidad del crimen y su relación con el crecimiento económico en méxico: un análisis regional," 2018.

[11] G. Cerda-Guillén, S. Cruz-Aké, and M. T. V. Martínez-Palacios, "Efectos del endeudamiento de los hogares mexicanos en su ahorro y consumo: Un enfoque de ciencia de datos," *Revista mexicana de economía y finanzas*, vol. 18, no. 2, 2023.

[12] R. M. Campos Vázquez and S. E. López-Araiza, "Grandes datos, google y desempleo," *Estudios Económicos (México, DF)*, vol. 35, no. 1, pp. 125–151, 2020.

[13] M. De Donno, K. Tange, and N. Dragoni, "Foundations and evolution of modern computing paradigms: Cloud, iot, edge, and fog," *IEEE access*, vol. 7, pp. 150936–150948, 2019.

[14] J. Garcıa, J. Molina, A. Berlanga, M. Patricio, A. Bustamante, and W. Padilla, "Ciencia de datos: Técnicas analıticas y aprendizaje estadıstico," 2018.

[15] IPSOS, "Lo que preocupa al mundo." `https://www.ipsos.com/es-mx/lo-que-preocupa-el-mundo`, 2024 [Fecha de último acceso: 05.12.2024].

[16] F. Crespo, T. Alves, M. Soto, *et al.*, "Ciencia de datos, inteligencia artificial, y sus impactos sobre la sociedad," *Observatorio Económico*, no. 169, pp. 9–11, 2022.

[17] V. Grossi, F. Giannotti, D. Pedreschi, P. Manghi, P. Pagano, and M. Assante, "Data science: a game changer for science and innovation," *International Journal of Data Science and Analytics*, vol. 11, no. 4, pp. 263–278, 2021.

[18] J. M. Rosa and E. L. Frutos, "Ciencia de datos en salud: desafíos y oportunidades en américa latina," *Revista Médica Clínica Las Condes*, vol. 33, no. 6, pp. 591–597, 2022.

[19] F. C. Reillo, "Estratificación social y uso del transporte público en el contexto de la crisis de seguridad en méxico," *Revista de Ciencias Sociales*, no. 172, pp. 157–175, 2021.

[20] O. Quintero Ávila, "El análisis y mapeo delictivo para el desarrollo de políticas públicas de seguridad en méxico," *Constructos Criminológicos*, vol. 4, no. 7, pp. 159–170, 2024.

[21] J. R. N. Ríos, "Factores determinantes de la percepción de seguridad en méxico: análisis estadístico a nivel regional," *Cimexus*, vol. 17, no. 2, pp. 228–243, 2022.

[22] F. V. Cortés and D. S. C. Islas, "La brecha digital como una nueva capa de vulnerabilidad que afecta el acceso a la educación en méxico," *Revista Academia y Virtualidad*, vol. 14, no. 1, pp. 169–187, 2021.

[23] F. S. Gutiérrez Cruz, J. C. Moreno Brid, and J. Sánchez Gómez, "Inversión pública y privada en méxico:¿ motores complementarios del crecimiento económico?," *El trimestre económico*, vol. 88, no. 352, pp. 1043–1071, 2021.

[24] P. M. Reyes, M. R. R. Hernández, and R. V. González, "La pandemia de covid-19 en la economía mexicana: condiciones iniciales, estrategias de política y efectos productivos," *Paradigma económico. Revista de economía regional y sectorial*, vol. 14, no. 2, pp. 55–83, 2022.

[25] C. Chiatchoua, C. Lozano, and J. Macías-Durán, "Análisis de los efectos del covid-19 en la economía mexicana," *Revista del Centro de Investigación de la Universidad la Salle*, vol. 14, no. 53, pp. 265–290, 2020.

[26] H. R. Ramírez, "El empleo en méxico durante el covid-19," *Observatorio de la Economía Latinoamericana*, no. 11, pp. 1–25, 2020.

[27] O. E. Ceballos Mina and A. De Anda Casas, "Estructura productiva laboral y pobreza en méxico: análisis municipal en tres regiones," *Desarrollo y Sociedad*, no. 88, pp. 129–168, 2021.

[28] J. E. Mendoza Cota, "¿ son la desigualdad y la pobreza un freno al crecimiento económico en méxico? correlación y causalidad desde una perspectiva regional," *El trimestre económico*, vol. 89, no. 356, pp. 1121–1151, 2022.

[29] J. J. L. Rivero, "Ciencia de datos e inteligencia artificial como apoyo para investigaciones cualitativas," *Revista EDUCARE-UPEL-IPB-Segunda Nueva Etapa 2.0*, vol. 26, no. 2, pp. 186–209, 2022.

[30] A. S. Hernández Téllez, "Desigualdad distributiva en méxico: comparación del índice de palma con respecto al índice de gini, 2000-2021.," 2024.

[31] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive bayes classification algorithm for traffic risk management," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 30, 2021.

[32] C. A. Cervera, P. R. Pérez, and J. M. de la Torre Hernández, "Una pequeña mirada a la estadística bayesiana en el análisis de datos cardiológicos," *REC: Interventional Cardiology*, vol. 4, no. 3, pp. 207–215, 2022.

[33] F. G. F. Guillen and J. B. Chaparro, "Análisis bayesiano del compromiso académico en estudiantes de psicología: diferencias según sexo y edad," *Revista de investigación en psicología*, vol. 24, no. 1, pp. 5–18, 2021.

[34] A. Donkers, B. de Vries, D. Yang, P. Pauwels, M. Poveda-Villalón, and W. Terkaj, "Knowledge discovery approach to understand occupant experience in cross-domain semantic digital twins.," in *LDAC@ ESWC*, pp. 77–86, 2022.

[35] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of big data knowledge discovery process," *Procedia computer science*, vol. 148, pp. 30–36, 2019.

[36] K. S. Hlaing and Y. Thaw, "Applications, techniques and trends of data mining and knowledge discovery database," *Int. J. Trend Sci. Res. Dev*, vol. 3, no. 5, pp. 1604–1606, 2019.

[37] J. Pérez-Ortega, N. N. Almanza-Ortega, K. Torres-Poveda, G. Martínez-González, J. C. Zavala-Díaz, and R. Pazos-Rangel, "Application of data science for cluster analysis of covid-19 mortality according to sociodemographic factors at municipal level in mexico," *Mathematics*, vol. 10, no. 13, p. 2167, 2022.

[38] IBM, "Metodología fundamental para la ciencia de datos." `https://www.ibm.com/downloads/cas/6RZMKDN8`, 2015 [Fecha de último acceso: 05.08.2024].

[39] M. E. G. Graus, "Estadística aplicada a la investigación educativa," *Dilemas Contemporáneos: Educación, Política y Valores*, 2018.

[40] M. Chavira Ibarvo, V. Ibarvo Urista, and G. A. Quijano Vega, "Índice de factores que inciden en el desarrollo de las zonas metropolitanas de méxico," 2020.

[41] A. Méndez, "El análisis factorial: una introducción conceptual para la enseñanza y aprendizaje," *Enseñanza e Investigación en Psicología*, vol. 6, no. 1, pp. 1–13, 2024.

[42] C. Li, Y. Chen, and Y. Shang, "A review of industrial big data for decision making in intelligent manufacturing," *Engineering Science and Technology, an International Journal*, vol. 29, p. 101021, 2022.

[43] M. B. van Egmond, G. Spini, O. van der Galien, A. IJpma, T. Veugen, W. Kraaij, A. Sangers, T. Rooijakkers, P. Langenkamp, B. Kamphorst, *et al.*, "Privacy-preserving dataset combination and lasso regression for healthcare predictions," *BMC medical informatics and decision making*, vol. 21, pp. 1–16, 2021.

[44] S. Wang, Y. Chen, Z. Cui, L. Lin, and Y. Zong, "Diabetes risk analysis based on machine learning lasso regression model," *Journal of Theory and Practice of Engineering Science*, vol. 4, no. 01, pp. 58–64, 2024.

[45] S. Peniche-Camps and M. Cortez-Huerta, "La costumbre al envenenamiento: El caso de los contaminantes atmosféricos de la ciudad de guadalajara, méxico," *Revista de Ciencias Ambientales*, vol. 54, no. 2, pp. 1–19, 2020.

[46] H. Cervantes and R. Kazman, *Designing software architectures: a practical approach*. Addison-Wesley Professional, 2024.

[47] M. Valez, M. Yen, M. Le, Z. Su, and M. A. Alipour, "Student adoption and perceptions of a web integrated development environment: An experience report," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pp. 1172–1178, 2020.

[48] F. Bedoya De la Cruz, "Analizador de código como extensión de visual studio code," 2023.

[49] S. G. Pérez Ibarra, J. R. Quispe, F. F. Mullicundo, and D. A. Lamas, "Herramientas y tecnologías para el desarrollo web desde el frontend al backend," in *XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja)*, 2021.

[50] T. Jalolov, "Frontend and backend developer difference and advantages," *Multidisciplinary Journal of Science and Technology*, vol. 4, no. 2, pp. 178–179, 2024.

[51] J. Vázquez Moreno *et al.*, "Herramienta web para la extracción y procesamiento de información a partir de ficheros csv. parte 1," 2020.

[52] R. B. Marqas, S. M. Almufti, and R. R. Asaad, "Firebase efficiency in csv data exchange through php-based websites," *Academic Journal of Nawroz University (AJNU)*, vol. 10, no. 3, 2022.

[53] A. F. Hidalgo Romero, "Comenzando con numpy y pandas," 2024.

[54] S. Saabith, T. Vinothraj, and M. Fareez, "A review on python libraries and ides for data science," *Int. J. Res. Eng. Sci*, vol. 9, no. 11, pp. 36–53, 2021.

[55] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, 2020.

[56] H. Hassani and E. S. Silva, "The role of chatgpt in data science: how ai-assisted conversational interfaces are revolutionizing the field," *Big data and cognitive computing*, vol. 7, no. 2, p. 62, 2023.

[57] I. D. Varela Español, S. Contreras Manjarres, and Y. S. Pinzon Caicedo, "Html al alcance de todos: una visión educativa," 2023.

[58] A. Wirfs-Brock and B. Eich, "Javascript: the first 20 years," *Proceedings of the ACM on Programming Languages*, vol. 4, no. HOPL, pp. 1–189, 2020.

[59] K. Mahmood, G. Rasool, F. Sabir, and A. Athar, "An empirical study of web services topics in web developer discussions on stack overflow," *IEEE Access*, vol. 11, pp. 9627–9655, 2023.

[60] M. Wadholm, "Green and sustainable javascript: a study into the impact of framework usage," 2023.

[61] E. Osorio Amaya, S. Inzunza Cáceres, S. Evelyn Ward, *et al.*, "Modelización estadística para el aprendizaje de la correlación y regresión lineal," 2023.

[62] M. A. Nova Martínez, E. G. Sorza Álvarez, and L. M. Zabala Arango, "Adecuación de modelos de regresión lineal simple en r-studio," 2023.

[63] J. O. Pinilla and A. F. O. Rico, "¿pearson y spearman, coeficientes intercambiables?," *Comunicaciones en Estadística*, vol. 14, no. 1, pp. 53–63, 2021.

[64] M. L. Waskom, "Seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[65] A. H. Sial, S. Y. S. Rashdi, and A. H. Khan, "Comparative analysis of data visualization libraries matplotlib and seaborn in python," *International Journal*, vol. 10, no. 1, pp. 277–281, 2021.

[66] B. L. V. Márquez, L. A. I. Hanampa, and M. G. M. Portilla, "Design thinking aplicado al diseño de experiencia de usuario," *Innovación y software*, vol. 2, no. 1, pp. 6–19, 2021.

[67] N. Segovia-García, "Propuesta de mejora en el diseño de interfaz y experiencia de usuario (ux) en moodle: valoración del alumnado," *Edutec, Revista Electrónica de Tecnología Educativa*, no. 82, pp. 199–216, 2022.

Apéndice A

Manual de Usuario

Para un funcionamiento óptimo del sitio, se recomienda:
Navegadores compatibles: Google Chrome, Microsoft Edge, Safari.
Conexión a internet estable.
Compatibilidad con dispositivos móviles.

A.1. Navegación

Figura A.1: Interfaz principal.

El sitio web DATAMEX cuenta con un menú principal, en la parte superior, con las siguientes secciones:

Inicio: Información general sobre el sitio. En la parte inferior se encuentran tres recuadros que brindan información de apertura para cada uno de los módulos. Seleccione el icono de la fecha para acceder a ella y haga clic en el botón “Explorar” para ingresar a la sección correspondiente. Esto se observa en la Figura A.2.

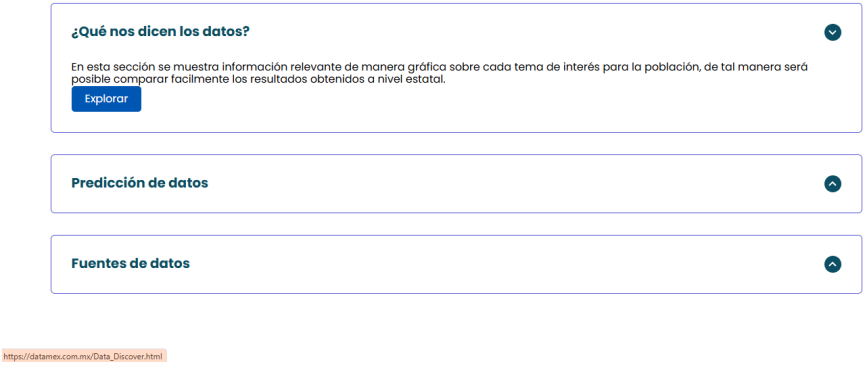


Figura A.2: Interfaz explorar.

¿Qué nos dicen los datos?: Información relevante de manera gráfica sobre cada tema de interés para la población. Seleccione las flechas de los laterales para obtener más información sobre algunos temas relevantes, en esta parte se implementa texto para una explicación breve y una imagen ilustrativa, de esta manera es posible entender cómo se comportan los datos al paso del tiempo.

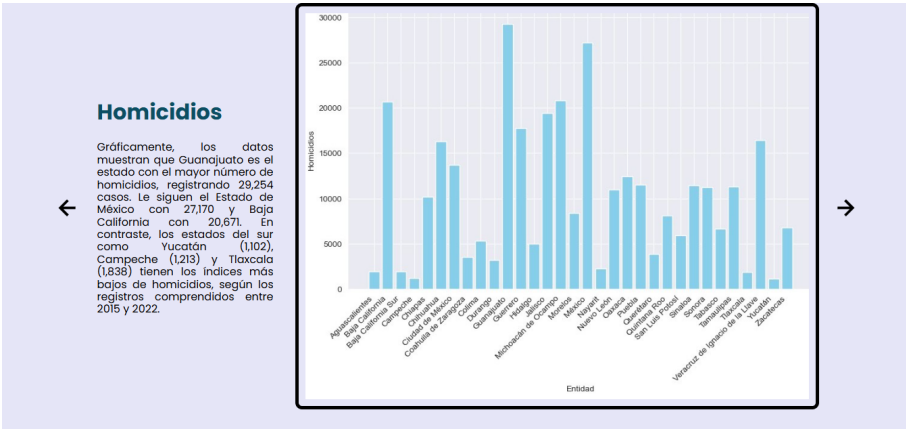


Figura A.3: Interfaz ¿qué nos dicen los datos?.

Predicción de datos: Como primer contenido se muestra una herramienta llamada “Mapa de calor” la cual permite visualizar el nivel de correlación que existe entre cada variable analizada.

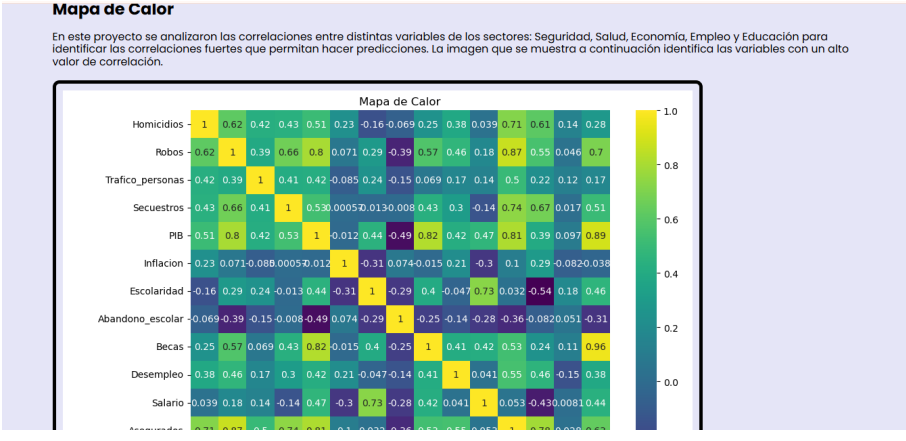


Figura A.4: Interfaz Mapa de calor.

Posterior a ello, se muestran las predicciones obtenidas con cada par de variable. Haga clic en el recuadro “Seleccione correlación” e ingrese a la misma.

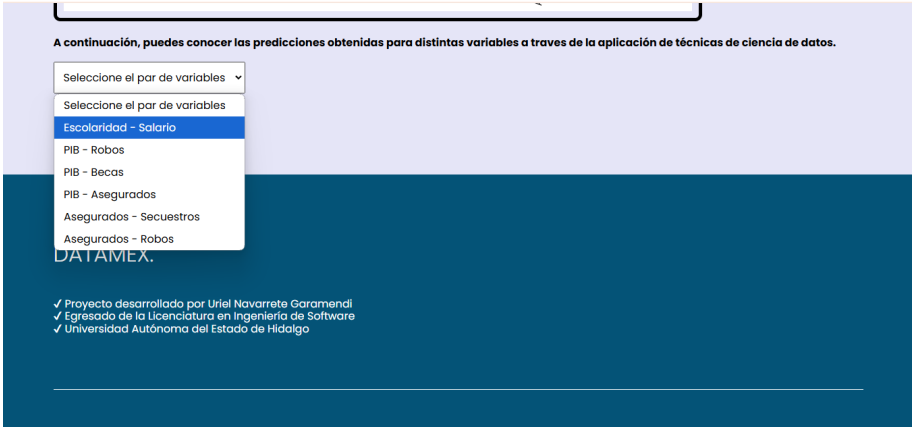


Figura A.5: Interfaz seleccione par de variables.

Para obtener el valor predicho respecto al Salario con base en la Escolaridad, como se muestra en este ejemplo, haga clic en el recuadro "Escolaridad" selección el nivel de escolaridad.

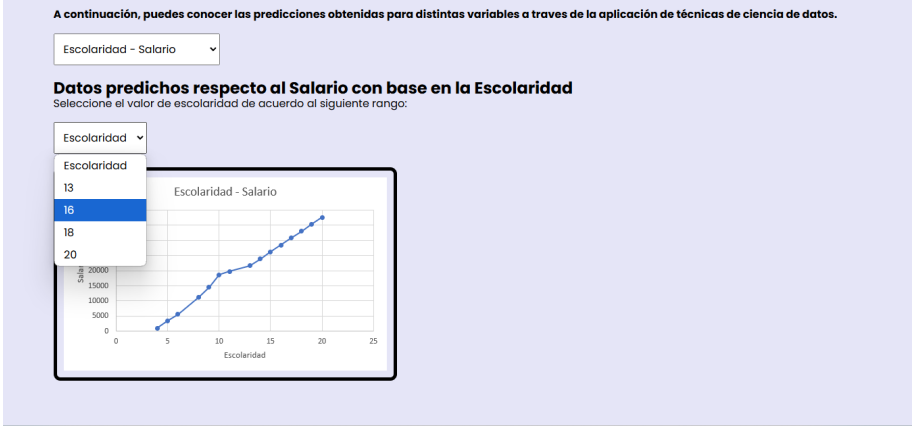


Figura A.6: Interfaz nivel de escolaridad.

Y de manera inmediata podrá conocer el dato predicho del salario mensual con base al nivel de escolaridad, tal como se muestra en la Figura A.7.

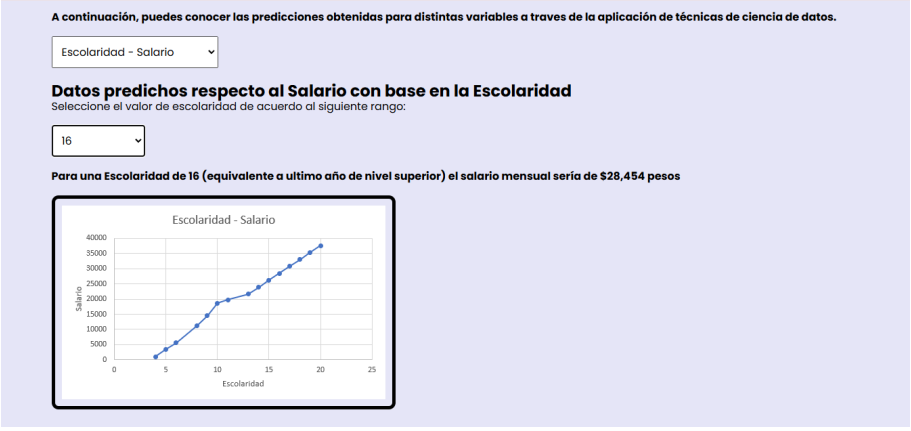


Figura A.7: Valores predichos.

Fuentes de datos: Repositorios de Datos para futuros proyectos de análisis.

En el recuadro “Seleccione categoría” haga clic en cualquier opción para conocer las fuentes que se utilizaron en su respectivo análisis.

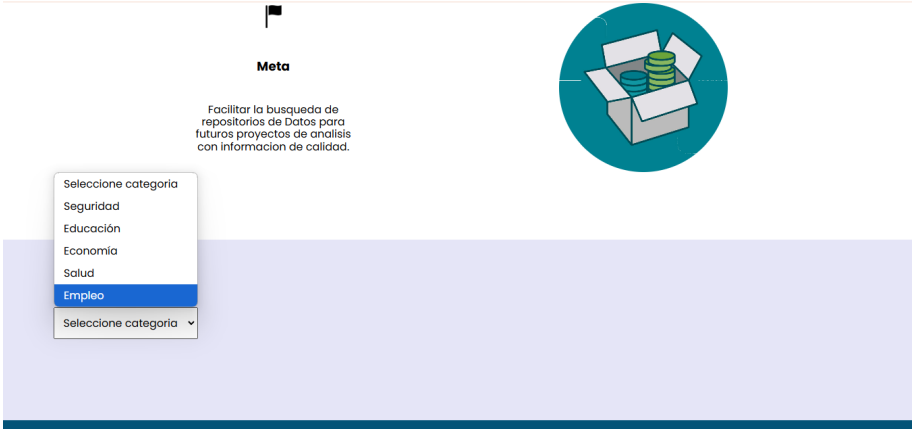


Figura A.8: Seleccionar categoria.

En forma de lista podrá obtener cada repositorio considerado, para posteriormente consultarlo y visualizar las bases del análisis realizado.

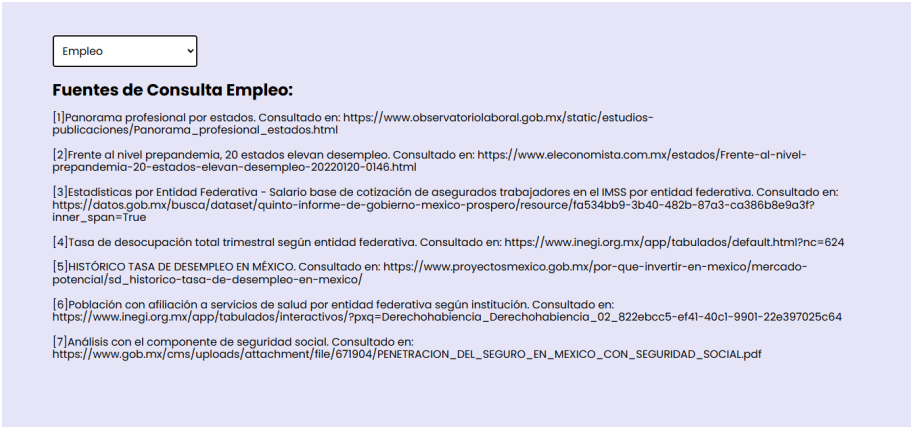


Figura A.9: Fuentes de consulta.