TECNOLÓGICO NACIONAL DE MÉXICO



Opción:

TITULACIÓN INTEGRAL

TESIS PROFESIONAL

Caracterización geográfica de temas delictivos en México con algoritmos de clustering y regresión

ALUMNO(A):

Abril Aime Nuñez Rodríguez

CARRERA:

Ingeniería en Tecnologías de la Información y Comunicación

No. DE CONTROL:

13170789

Culiacán, Sinaloa febrero 2020.

DIVISIÓN DE ESTUDIOS PROFESIONALES
ASUNTO**: Autorización de Impresión**

Culiacán, Sinaloa, 13/02/2020 13:20:36

C.- ABRIL AIME NUÑEZ RODRIGUEZ
PASANTE(S) DE LA CARRERA DE INGENIERÍA EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIONES
P R E S E N T E.

De acuerdo con fallo emitido por la Comisión Revisora Integrada por:

		Cedula Profesional
Presidente:	Licenciado en Informática **Nora Esmeralda Cancela García**	3373654
Secretario:	Maestría en Comunicación y Tecnologías Educativas **Mirna del Rosario Quevedo Camacho**	1899586
Vocal:	Maestría en Ciencias en Ciencias de la Computación **Gloria Ekaterine Peralta Peñuñuri**	6038275

Y considerando que reúne los requisitos establecidos en la Opción: **TITULACIÓN INTEGRAL: TESIS PROFESIONAL "CARACTERIZACION GEOGRAFICA DE TEMAS DELICTIVOS EN MEXICO CON ALGORITMOS DE CLUSTERING Y REGRESION"** damos a usted nuestra autorización para que proceda a imprimir su trabajo profesional.

Sin otro particular, quedo de usted.

A T E N T A M E N T E
Excelencia en Educación Tecnológica.®

M.C. CARLOS RAFAEL LIZÁRRAGA ARREOLA
JEFE DE LA DIVISIÓN DE ESTUDIOS PROFESIONALES

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PUBLICA
TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE CULIACÁN

DEPARTAMENTO DE DIVISIÓN DE ESTUDIOS PROFESIONALES

Caracterización geográfica de temas delictivos en México con algoritmos de clustering y regresión.

RESUMEN

Actualmente en México, el índice de delincuencia va en aumento, existen diferentes factores y comportamientos por cada municipio, pero no es posible detectarlos a simple vista, para ello, la ciencia de datos proporciona técnicas de análisis, con el fin de obtener dicha información.

El proyecto consta de diferentes etapas, empezando por la extracción de datos, los cuales se obtienen de las bases de datos de INEGI. Una vez obtenida, se prosigue a limpiar, esto con el uso de tecnologías del lenguaje Python, ya que posee librerías que facilitan el trabajo. Después de limpiar los datos, se implementan algoritmos de machine learning tanto para el descubrimiento de grupos en los datos, así como la predicción de variables destacadas, todo esto a su vez, plasmado en un sistema de información geográfica.

CONTENIDO

ÍNDICE DE FIGURAS Y TABLAS 5

1 INTRODUCCIÓN 7

1.1 Antecedentes 7

1.2 Descripción del problema 8

1.3 Hipótesis o preguntas de investigación 8

1.4 Objetivo general 9

1.5 Objetivos específicos 9

1.6 Justificación 10

1.7 Alcance 10

2 FUNDAMENTO TEÓRICO **11**

3 PROCEDIMIENTO **17**

3.1 Proceso ETL. 17

3.1.1 Extracción. 18

3.1.2 Transformación. 20

3.1.3 Carga. 22

3.2 Machine Learning. 23

3.2.1 Aprendizaje no supervisado. 23

3.2.2 Clustering. 23

3.2.3 Exploración de algoritmos de clustering. 25

3.2.4 Algoritmo KMeans. 25

3.2.5 Aplicación de PCA. 27

3.2.6 Adición de columna Cluster. 28

3.2.7 Generación de diccionario. 29

3.3 Tabla de incidencias. 29

3.4 Aprendizaje supervisado. 32

3.4.1 Regresión. 32

3.4.1.1 División de datos. 32

3.4.1.2 Selección de algoritmo de regresión. 33

3.4.1.3 Algoritmo Gaussian Naive Bayes. 33

3.5 Data Visualization. 34

3.5.1 Geocoding. 34

4 RESULTADOS **34**

6 CONCLUSIONES **35**

7 FUENTES DE INFORMACIÓN 37

ÍNDICE DE FIGURAS Y TABLAS

Figura 1. Proceso ETL. 16

Figura 2. Información encuesta ENDUTIH. 18

Figura 3. Base de datos ENDUTIH. 19

Figura 4. Tabla viviendas 2015. 20

Figura 5. Archivo csv. 21

Figura 6. Método de Elbow. 23

Figura 7. Proceso para generar clusters. 24

Figura 8. Clusters obtenidos. 25

Figura 9. Clusters con PCA. 25

Figura 10. Columna Cluster. 26

Figura 11. Proceso de obtención de moda. 26

Figura 12. Diccionario de viviendas. 27

Figura 13. Tabla de incidencias. 27

Figura 14. Adición de columna Cluster. 28

Figura 15. Resultado de iteración. 28

Figura 16. Proceso de separación de datos. 30

Figura 17. Entrenamiento de algoritmo de regresión. 30

Figura 18. Predicción de algoritmo de regresión. 31

Figura 19. Performance de algoritmo de regresión. 32

Figura 20. Mapa geocodificado de incidencias delictivas. 33

Figura 21. Clusters con MeanShift. 33

1 INTRODUCCIÓN

Esta investigación centra su atención en el análisis de información de delitos cometidos en los últimos años, la cual ha sido recopilada por un organismo mexicano a partir de encuestas o censos aplicados alrededor del territorio.

Recientemente, gracias a las iniciativas de los llamados datos abiertos, es decir, datos a los cuales cualquier persona puede acceder, usar o compartir, es posible efectuar esta labor de investigación sin necesidad de invertir tiempo en recopilar la información. Las iniciativas de Datos Abiertos se han fijado como políticas en múltiples países, ejemplo de ellos son, Estados Unidos, Reino Unido, Canadá y Nueva Zelanda, regiones que desarrollaron acciones concretas para la apertura de su información pública, y México no es la excepción.

Las herramientas que ofrece la ciencia de datos para el procesamiento y análisis de los datos permiten identificar patrones, encontrar agrupamiento de objetos de acuerdo con sus características, predecir comportamientos futuros, entre otras tareas que ayudan a identificar los factores que dan pie a distintos problemas y a partir de ello poder tomar decisiones acertadas que contribuyan a disminuir o erradicar el problema en cuestión.

Para este proyecto se seleccionaron las bases de datos de INEGI que almacenan información referente a la disponibilidad y uso de las tecnologías en los hogares, así como las bases de datos que contienen la caracterización de los delitos cometidos en México. De los datos recopilados se hicieron agrupamientos y se realizaron predicciones para determinar el número de delitos ejecutados por cada estado de la República Mexicana.

1.1 Antecedentes

La delincuencia es una conducta humana, es una actitud, que se produce en todos los estamentos de una sociedad, que afecta a todos los sectores sociales, incluyendo el poder político, religioso y empresarial. Pero donde más agitación tiene es en los sectores marginados, por la manera en que se presenta, por la violencia con que se da, por el método que usa. La búsqueda de una solución a este problema debe ser organizada por los organismos del Gobierno responsables de aplicar las políticas.

La incidencia delictiva en el país mexicano se ha incrementado progresivamente a través de los años, sin embargo, no todas las modalidades de delitos aumentan en la misma proporción, entre los ilícitos que han sido registrados por el organismo mexicano INEGI en el país mexicano están homicidio doloso, homicidio culposo, feminicidio, secuestro, extorsión, trata de personas y robo en distintas categorías.

A causa de la necesidad de disponer de información más clara y específica sobre la incidencia delictiva, el Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública del país mexicano, en conjunto con distintas áreas rectoras y técnicas especializadas, tanto gubernamentales como de la sociedad civil, se dio a la tarea de elaborar e implementar una nueva metodología para registrar y clasificar los delitos y las víctimas a fin de obtener estadísticas. La información registrada a partir de enero de 2018 con base en la nueva metodología constituye la estadística oficial de incidencia delictiva en el país. Se dispone de la información de incidencia delictiva para los ámbitos estatal y municipal de ilícitos como homicidio, secuestro, extorsión, lesiones, corrupción de menores, tráfico de menores, rapto y trata de personas; así como para unidades robadas de vehículos y tractores.

Entre las numerosas iniciativas tomadas para la disminución del crimen se encuentra la propuesta por la empresa Wynuard Group junto con Motorola. Desarrollaron un sistema que ayuda a identificar patrones de conducta delictiva y descubrir conexiones ocultas en distintas fuentes de datos, como bases de datos de delincuentes y expedientes en apenas minutos. El modelo, aunque parezca estar inspirado en alguna serie policiaca estadounidense, se maneja con gigantescas bases de datos que combinan varios datos para obtener diagnósticos cercanos.

1.2 Descripción del problema

La crisis de inseguridad en el país de México durante los últimos años se ha generado por diversos ilícitos, el Observatorio Nacional Ciudadano de Seguridad y Justicia y Legalidad (ONC) anunció en su estudio “Incidencia de los delitos de alto impacto en México” del año 2017 las tasas de incidencias de distintos delitos tales como homicidio doloso, homicidio culposo, feminicidio, secuestro, extorsión, trata de personas, robo en diferentes categorías, violación y narcomenudeo. Se encontró que para el año 2018 la violencia podría incrementar en un rango de entre 6.6 y 15.8%. Esto es un problema que afecta a la sociedad y atenta contra la calidad de vida, por ello es necesario buscar alternativas que contribuyan a su disminución o erradicación. Es aquí donde interviene la ciencia de datos proporcionando herramientas que hacen posible la búsqueda de información de valor contenida en distintas fuentes de datos, y en base a ella tomar decisiones y acciones a favor de la seguridad.

Existe mucha información de delincuencia en múltiples bases de datos que pueden ser explotadas para un bien común, ejemplo de ello son las generadas por el Instituto Nacional de Estadística y Geografía (INEGI) a partir de censos y encuestas aplicados a lo largo del país. Tomando esto como motivación, la investigación se enfoca en analizar datos con información de delitos y exponer los resultados en un mapa del país de México geo codificado.

1.3 Hipótesis o preguntas de investigación

¿Es posible conocer la relación de los municipios mexicanos en base a la información de los índices delictivos y uso de tecnologías en los hogares proporcionada por INEGI, a través de la Ciencia de datos?

1.4 Objetivo general

Conocer a través de un mapa geo codificado, las agrupaciones conformadas por los municipios de la República Mexicana, de acuerdo con los delitos cometidos y el uso de tecnologías de información y comunicación en los hogares, aplicando tecnologías de minería de datos y Machine Learning.

1.5 Objetivos específicos

- Analizar distintas fuentes de información que proporcionan datos sobre índices de actos delictivos en los municipios de la república y el uso de tecnologías en los hogares.
- Encontrar las agrupaciones de los datos que comparten características similares.
- Predecir el número de delitos cometidos en los municipios de los estados de la República Mexicana.
- Mostrar en un mapa geo codificado los grupos formados por los municipios a partir de la información relativa a los índices de delincuencia.

1.6 Justificación

Esta investigación se enfoca en el análisis de información de los índices de criminalidad dentro de la República Mexicana, debido al incremento de ilícitos en distintos municipios del país. Esto con ayuda del análisis de datos ya que su principal fundamento es la búsqueda de conocimiento en los datos y sin el análisis no es posible detectar comportamientos o características que influyan en los crímenes cometidos en el país. Así, será posible conocer los municipios que comparten características similares en relación con el tema de delincuencia.

1.7 Alcance

La presente investigación explora los datos recopilados por el organismo público mexicano INEGI a partir de encuestas o censos aplicados alrededor del territorio, sobre los cuales pretende generar agrupaciones de acuerdo con las similitudes de estos y mostrarlas en un mapa de la República Mexicana. Se seleccionaron únicamente aquellas bases de datos que proporcionan información acerca del uso de tecnologías en los hogares (ENDUTIH) en algunos de los municipios mexicanos, igualmente las bases de datos que almacenan las incidencias delictivas con sus características, limitada también a ciertos municipios y al año 2014 específicamente.

2 FUNDAMENTO TEÓRICO

2.1 Data

La definición formal de data sería una recopilación o un conjunto de variables cualitativas y / o cuantitativas que contienen valores basados en observaciones. Por lo general, los datos se miden y se recopilan a partir de varias observaciones (Sarkar & Bali & Sharma, 2018).

2.1.1 Variable

Las variables son propiedades o características de algún evento, objeto o persona que pueden tomar diferentes valores o cantidades (Ziemer, 2007).

2.1.2 Quantitative variable

Las variables cuantitativas son aquellas variables que se miden en términos de números (Ziemer, 2007).

2.1.3 Qualitative variable

Las variables cualitativas son aquellas que expresan un atributo cualitativo como el color del cabello, el color de los ojos, la religión, la película favorita, el género, etc. (Ziemer, 2007).

2.2 Dataset

Un dataset generalmente consiste en vectores de características, donde cada vector de características es una descripción de un objeto mediante el uso de un conjunto de características (Zhou ,2012).

2.2.1 Dimensión

El número de características de un conjunto de datos se denomina dimensión o dimensionalidad (Zhou ,2012).

2.3 Open data

Los datos abiertos son la idea de que algunos datos deben estar disponibles libremente para que todos puedan usarlos y volver a publicarlos como lo deseen, sin restricciones de derechos de autor, patentes u otros mecanismos de control (Auer, 2007).

2.4 Data science

Data science es un campo interdisciplinario muy diverso que abarca múltiples campos. Data science se ocupa básicamente de principios, metodologías, procesos, herramientas y técnicas para recopilar conocimientos o información a partir de datos (estructurados y no estructurados). Data Science es más una compilación de procesos, técnicas y metodologías para fomentar una cultura basada en decisiones basada en datos (Sarkar & Bali & Sharma, 2018).

2.5 ETL process

Proceso usado para extraer datos de fuentes internas y externas de una organización, transformar estos datos, y cargarlos dentro de un almacén de datos (Vaisman & Zimányi, 2014).

2.5.1 Normalization

La normalización es el proceso de estandarización del rango de valores de atributos. Los algoritmos de machine learning en muchos casos utilizan métricas de distancia, atributos o características de diferentes escalas / rangos que pueden afectar adversamente los cálculos o sesgar los resultados (Sarkar & Bali & Sharma, 2018).

2.6 Data mining

Es el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas al excavar grandes cantidades de datos almacenados en un almacén, usando estadística, aprendizaje automático, inteligencia artificial y visualización de datos (Sumathi & Sivananda, 2006).

2.7 Machine learning

Es una rama de la inteligencia artificial. Usando la computación, diseñamos sistemas que pueden aprender de los datos por medio de entrenamiento. Los sistemas pueden aprender y mejorar con la experiencia y, con el tiempo, refinar un modelo que se puede usar para predecir los resultados de preguntas basadas en el aprendizaje anterior (Bell, 2015).

2.7.1 Algoritmo

El algoritmo es una serie de operaciones detalladas y no ambiguas para ejecutar paso a paso que conducen a la resolución de un problema, y se representan mediante una herramienta o técnica (Aguilar, 1993).

2.7.2 Unsupervised learning

Unsupervised learning estudia cómo los sistemas pueden aprender para representar patrones de entrada particulares de una manera que refleje la estructura estadística de la colección general de patrones de entrada (Tripathi, 2017).

2.7.3 Clustering

Clustering es una división de datos en grupos de objetos similares. Cada objeto(cluster) consiste en objetos que son similares entre sí y diferentes a los objetos de otros grupos (Tripathi, 2017).

2.7.4 Principal Component Analysis

Principal Component Analysis es uno de los métodos más comunes usados en el análisis de datos para proporcionar una descripción condensada y describir patrones de variación en dataset multivariadas (Gauch, 1982).

2.7.6 Regression

El análisis de regresión es una técnica de modelado predictivo que investiga la relación entre una variable dependiente y una variable independiente (Shukla, 2017).

2.7.6 Underfitting

Ocurre cuando un modelo estadístico no puede capturar adecuadamente la estructura subyacente de los datos. Un modelo mal equipado es un modelo en el que faltan algunos parámetros o términos que aparecen en un modelo correctamente especificado (Skrondal, 2010).

2.7.7 Overfitting

El ajuste excesivo es cuando su modelo aprendido funciona excepcionalmente bien en el conjunto de datos de entrenamiento, pero tiende a funcionar mal en el conjunto de datos de prueba (Shukla, 2017).

2.8 Data visualización

Visualmente despliegan cantidades medidas mediante el uso combinado de puntos, líneas, un sistema de coordenadas, numero, símbolos, palabras, sombreado y color (Tufte, 2001).

2.8.1 Geocoding

El proceso en el cual el software de GIS coincide cada registro en un atributo de la base de datos con el archivo geográfico (Melnick, 2002).

2.8.2 GIS software

Software que contiene operadores geográficos que puede identificar los objetos en un mapa (Melnick, 2002).

2.8.3. Shapefile

Es un formato de datos vectoriales geoespaciales para el software del sistema de información geográfica (ESRI, 1998).

2.9 Python

Python es un lenguaje de programación interpretado de alto nivel para programación de propósito general. Creado por Guido van Rossum y lanzado por primera vez en 1991, Python tiene una filosofía de diseño que enfatiza la legibilidad del código, especialmente al usar espacios en blanco significativos. Proporciona construcciones que permiten una programación clara tanto a pequeña como a gran escala (Kuhlman, 2012).

2.9.1 Numpy

Es una de las bibliotecas más importantes de Python para cálculos numéricos. Agrega soporte al núcleo Python para matrices multidimensionales (y matrices) y operaciones vectorizadas rápidas en estas matrices (Sarkar & Bali & Sharma, 2018).

2.9.2 Pandas

Pandas es una importante biblioteca de Python para manipulación de datos, disputas y análisis. Funciona como un conjunto de herramientas intuitivas y fáciles de usar para realizar operaciones en cualquier tipo de datos (Sarkar & Bali & Sharma, 2018).

2.9.3 Data Frame

Un Data Frame representa una tabla de datos rectangular y contiene una colección ordenada de columnas, cada una de las cuales puede ser un tipo de valor diferente (McKinney, 2018).

2.9.4 matplotlib

matplotlib es una popular biblioteca de ploteo. Proporciona interfaces y utilidades para generar visualizaciones de calidad de publicación (Sarkar & Bali & Sharma, 2018).

2.9.5 Sklearn

Sklearn es la biblioteca de aprendizaje automático más popular y ampliamente utilizada. Está construido sobre SciPy y presenta una gran cantidad de algoritmos de aprendizaje supervisados y no supervisados (Swamynathan, 2017).

2.9.6 SciPy

SciPy es una colección de paquetes que abordan varios dominios de problemas estándar diferentes en la computación científica (McKinney, 2018).

2.10 Jupyter Notebook

Es una herramienta que permite a los científicos de datos registrar su proceso de análisis completo, de la misma manera que otros científicos usan un cuaderno de laboratorio para registrar pruebas, avances, resultados y conclusiones (Toomey, 2016).

2.11. KML

Es un lenguaje XML centrado en la visualización geográfica, incluida la anotación de mapas e imágenes. La visualización geográfica incluye no solo la presentación de datos gráficos en el mundo, sino también el control de la navegación del usuario en el sentido de dónde ir y dónde mirar (Wilson, 2007).

3 PROCEDIMIENTO

El proceso de análisis para esta investigación implica el desarrollo de distintas etapas donde se trabaja con datos, que son el objeto de nuestro estudio, cada una de ellas está orientada a la búsqueda del conocimiento en la información. Esto se puede lograr con el uso de técnicas y herramientas que nos proporciona data mining, la cual juega un papel fundamental dentro de data science. Sin embargo, antes de poder utilizar cualquiera de estas técnicas es importante que se haga una integración de los datos y prepararlos para que puedan ser procesados. Eso se puede conseguir a través de los procesos ETL.

3.1 Proceso ETL.

Los procesos ETL son una parte fundamental en la integración de datos, están constituidos por las fases Extract, Transform y Load. De acuerdo al objetivo que se plantea en la investigación, ETL permite extraer información de distintas fuentes de datos, transformarla y cargarla para que posteriormente pueda ser procesada por los algoritmos de machine learning y conseguir un mejor resultado.

En la figura 1 se puede apreciar de manera general las etapas del proceso ETL.



Figura 1. Proceso ETL.

3.1.1 Extracción.

En esta etapa del proceso ETL se busca extraer información ya sea de una o varias fuentes, existen múltiples opciones que la pueden proporcionar, entre ellas se encuentran las páginas web, las redes sociales, encuestas, censos, bases de datos, y otras más. Esta última fue la que se eligió para la investigación. Se encontraron las bases de datos del Instituto Nacional de Estadística y Geografía, mejor conocido como INEGI, el cual mantiene la política de datos abiertos. El instituto brinda una gran cantidad de encuestas y bases de datos para distintos temas ya sea estadísticos o geográficos. Para este caso, se optó por utilizar la información generada en el año 2015 por la encuesta ENDUTIH que significa Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares, ya que se necesita saber si existe alguna relación entre el uso de tecnologías de información en el hogar y la delincuencia, o si su uso repercute sobre los índices de delitos en el país. Se seleccionó el año 2015 ya que es el que guarda más registros en comparación con otros, y para este tipo de estudios es importante explorar una gran cantidad de datos puesto que pueden generar mayor conocimiento.



En la figura 2 se muestra información obtenida de la página oficial del INEGI sobre la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares del año 2015.

Figura 2. Información de encuesta ENDITUH.

Una vez seleccionada la fuente de datos, se prosigue descargando la base de datos de la página oficial de INEGI. Es importante mencionar que todas las fuentes de información son distintas, la manera en la que se capturan los datos de una base de datos a otra cambia y los formatos de cada una de ellas también. En este caso el formato de los archivos estaba en DBF (Data Base File), por lo que se optó por hacer una conversión del archivo al formato CSV (Comma Separated Value) pues las tecnologías de transformación de datos lo requieren, para ello se utilizaron las librerías simpledbf y Pandas de Python, con ellas se realiza la tarea de conversión.

Se encontró que dentro de la base de datos existían las tablas viviendas, hogares, residentes y usuarios, de las cuales se toma la tabla de viviendas, siendo esta la que servirá como objeto de análisis.

En la figura 3 se muestran los archivos con la información de las viviendas para descargar de la página oficial de INEGI.

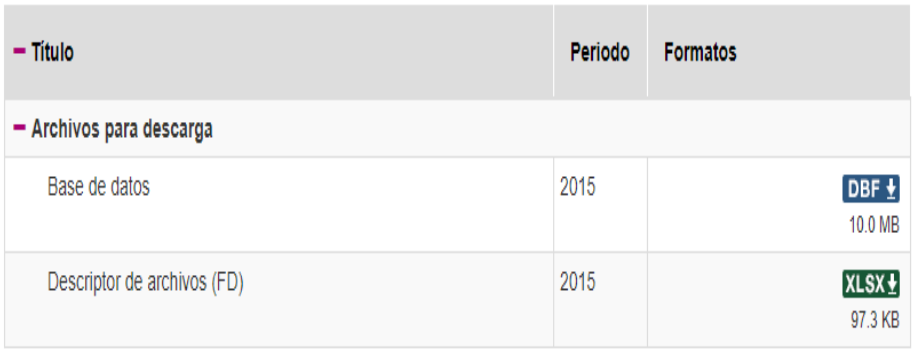


Figura 3. Base de datos ENDITUH.

3.1.2 Transformación.

Esta es la fase del proceso donde se debe limpiar la información de las tablas que se seleccionaron en la fase de extracción, con el fin de que los datos sean íntegros y puedan ser procesados por los algoritmos de machine learning. Para esto se utilizó Python3 con las librerías de Pandas, Matplotlib y Numpy, cada una de estas con un propósito distinto, Pandas para crear data frames o tablas, Matplotlib para la graficación de la información y Numpy para funciones matemáticas.

La limpieza de cada tabla se debe llevar a cabo columna por columna, teniendo un total de 75 columnas dentro de la tabla de Disponibilidad y Uso de Tecnologías de la Información en los Hogares, que fue la que se seleccionó anteriormente. Al observar la tabla seleccionada, se puede notar que esta presenta algunos problemas que pueden tener repercusiones en los resultados finales. A continuación, se mencionan los defectos que presenta la tabla en cuestión y las transformaciones que se puede llevar a cabo para resolverlos.

- Existen columnas que almacenan información que no es relevante para el análisis, por lo que se opta por eliminarlas.

- Existen columnas con datos de tipo string que es necesario eliminar o convertir a tipo numérico ya que los algoritmos que se utilizaran para la investigación sólo procesan números. En los casos en los que el valor string representa un número, puede ser convertido, en caso contrario la columna se elimina.
- Existen muchos campos con valores nulos que no son importantes para la investigación, aquí tenemos dos opciones, eliminar la columna con valores nulos o sustituir ese valor por otro que sea representativo del mismo. En este caso se eliminaron las columnas que presentaban nulos en más de la mitad de sus registros ya que no aportan información de valor al análisis, por el contrario, en aquellas con un número de nulos menor a la mitad de sus registros se optó por sustituirlos.
- Existen columnas con poca información, para este caso se decide eliminarlas ya que realmente no aportan nada al análisis y pueden generar un resultado equivocado.
- Los valores numéricos de algunas columnas presentan una diferencia enorme con respecto a los de otras columnas, esto puede causar un resultado erróneo, para resolverlo se hace una normalización de los datos, de modo que todos estén a la misma escala, con el fin de que no hubiera una dispersión de datos muy grande, además de que esto minimiza los errores que se producen en los cálculos de los algoritmos.

  En la figura 4 se muestra la tabla con la información de las viviendas después de haber sido transformada atendiendo a los puntos mencionados anteriormente.



Limpieza a tabla viviendas de 2015

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('TIC_2015_VIVIENDAS.csv', encoding='latin', low_memory=False)
df.head()
```

	Unnamed: 0	D_R	UPM	VIV_SEL	ENT	MUN	LOC	P1_1	P1_2	P1_3	...	P1_7	P1_8	P2_1	P2_2	P2_3	AREAM	TLOC
0	0	NaN	101250	5	1	1	1	2	2	4	...	1	1	4	1	1	14	1
1	1	NaN	101267	1	1	1	1	3	3	4	...	1	1	7	1	1	14	1
2	2	NaN	101267	3	1	1	1	3	4	6	...	1	1	8	1	1	14	1
3	3	NaN	101267	4	1	1	1	3	3	4	...	1	1	5	1	1	14	1
4	4	NaN	101267	5	1	1	1	3	3	7	...	1	1	6	1	1	14	1

5 rows × 25 columns

Figura 4. Tabla viviendas 2015.

3.1.3 Carga.

Esta es la fase final del proceso ETL, simplemente se toma la tabla que previamente fue limpiada y se continúa creando los archivos limpios, estos son exportados a formato CSV ya que, como se mencionó anteriormente, es un tipo de formato sencillo de manejar y es el que procesan las

```
df_copy.to_csv('viviendas_2015_clean.csv')
```

bibliotecas de python que se utilizarán. En la figura 5 se muestra la manera de exportar la información de la tabla a formato CSV usando el lenguaje de programación python.

Figura 5. Archivo csv.

3.2 Machine Learning.

Una vez que los datos han sido integrados es posible aplicar cualquier técnica de data mining. Para este caso, lo que interesa es hacer una predicción que arroje el número de delitos que se puede cometer en un estado y encontrar los grupos formados por los estados de la República que comparten características similares en relación con el uso de tecnologías de la información. Para realizar el agrupamiento de los datos y la predicción es necesario el uso de los algoritmos de Machine Learning, estos brindan una metodología para llegar al resultado deseado, mediante algoritmos de clustering y regresión que aprenden y resuelven problemas gracias a la experiencia.

Los algoritmos pueden aprender de formas diferentes mediante un aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado o aprendizaje por refuerzo. Para este estudio se utilizan los algoritmos de aprendizaje supervisado y aprendizaje no supervisado.

3.2.1 Aprendizaje no supervisado.

Lo que interesa es encontrar los grupos formados por los estados de la República que comparten características similares en relación con el uso de tecnologías de la información. Esto puede ser posible con las técnicas de aprendizaje no supervisado que permiten descubrir estructuras ocultas en los datos donde no se sabe que existe. Permite afrontar problemas de los que apenas se conoce el aspecto que tienen los resultados ya que falta un conocimiento previo.

3.2.2 Clustering.

Los algoritmos de clustering, están clasificados como algoritmos no supervisados, esto porque no hay identificación de grupos en los datos. Para este estudio se busca conocer los grupos en los que se pueden clasificar los municipios mexicanos de acuerdo con determinadas características.

Existen diferentes algoritmos que permiten hacer clustering dentro de la biblioteca sklearn de python, cada uno de ellos con sus ventajas y desventajas, la selección del algoritmo correcto depende de ciertos factores que se deben tomar en cuenta, a continuación, se analizarán dos de ellos para posteriormente hacer una elección.

3.2.3 Exploración de algoritmos de clustering.

De la gran variedad de algoritmos que ofrece la biblioteca sklearn se seleccionaron dos de ellos para hacer una comparación y escoger el idóneo. Las dos opciones que considerar son MeanShift y Kmeans.

Ambos algoritmos son muy fáciles de implementar, pero también computacionalmente muy eficientes comparado con otros algoritmos, sin embargo, KMeans es más popular. Al realizar algunas prácticas donde se utilizaron dataset con un gran número de registros, se pudo observar que tanto KMeans como MeanShift necesitaban de un tiempo considerable para procesar una gran cantidad de información, este es un punto muy importante ya que para esta investigación se necesita explorar un conjunto enorme datos.

Otra característica que se debe tomar en cuenta es que mientras MeanShift genera un número de clusters estimados con respecto a los datos que le fueron proporcionados, sin la necesidad de indicarle cuantos grupos deseamos que se formen, KMeans requiere que se le indique cuántos clusters debe generar, esto puede ser un problema debido a que no se sabe realmente cuántos grupos se pueden producir, una elección inadecuada puede dar como resultado un rendimiento de clúster deficiente.

Después de examinar ambos algoritmos se optó por utilizar Kmeans. A pesar de las desventajas que supone su implementación.

3.2.4 Algoritmo KMeans.

Este algoritmo es muy popular gracias a su sencilla implementación, también es computacionalmente muy eficiente comparado con otros algoritmos. KMeans es muy bueno para identificar clusters con forma esférica. Trabaja iterativamente para asignar cada objeto a un grupo de acuerdo con sus características. El agrupamiento se hace minimizando la suma de distancias cuadráticas entre cada objeto y el centroide de su clúster.

Uno de los inconvenientes de este algoritmo es que tenemos que especificar el número de clusters, k, a priori. Una elección inadecuada para k puede dar como resultado un rendimiento de

clúster deficiente. Es por ello por lo que para elegir un número de grupos óptimos se puede utilizar el método de elbow como apoyo para encontrar el número k óptimo.

El método de elbow es una técnica que muestra en una gráfica los diferentes clusters que se pueden generar. La figura 6 muestra la gráfica generada a partir de la tabla viviendas del año 2015.

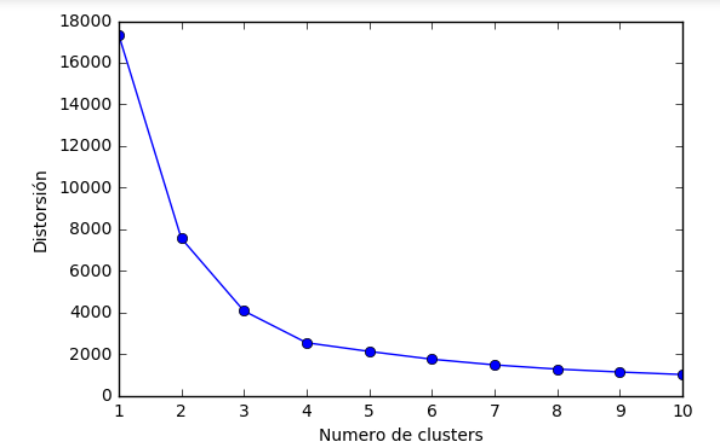


Figura 6. Método de Elbow.

Cada punto de la gráfica anterior representa los grupos que se pueden generar, para este ejemplo se decidió tomar como número de clusters óptimo el 4 ya que se recomienda elegir un conjunto en el que la distorsión empiece a ser mínima.

Una vez establecido el número de clusters óptimo, se continúa ejecutando el algoritmo KMeans asignando como valores de entrada el dataset e indicando que se esperan 4 agrupaciones. Su funcionamiento consta de varias etapas en la primera se leen los objetos del dataset. Una vez escogido el número de grupos k, se establecen k centroides en el espacio de los datos, de forma aleatoria. Cada objeto es asignado a su centroide más cercano. Se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición de promedio de los objetos pertenecientes a dicho grupo. El proceso se repite hasta que los centroides no se mueven o se mueven a una distancia mínima. En la figura 7 se puede visualizar claramente el proceso.

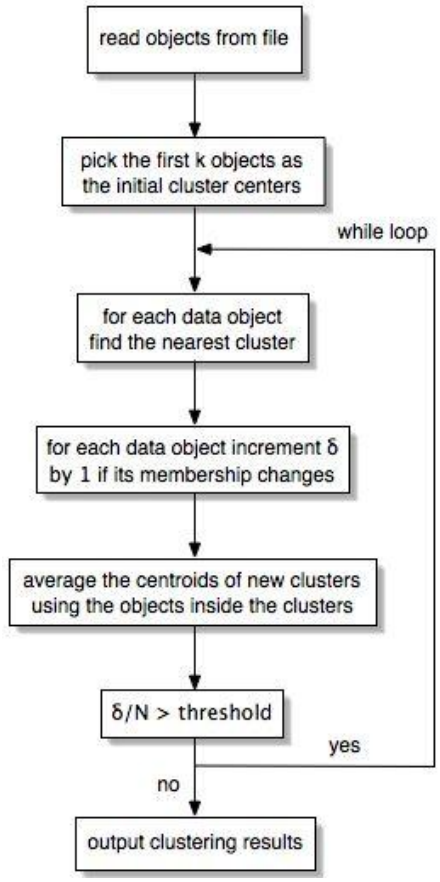


Figura 7. Proceso para generar clusters.

Al terminar su ejecución obtenemos la gráfica mostrada en la figura 8, en ella están plasmados los objetos del dataset, la forma y color de cada uno indican el cluster al que pertenecen, también se muestran los centroides de cada agrupación.

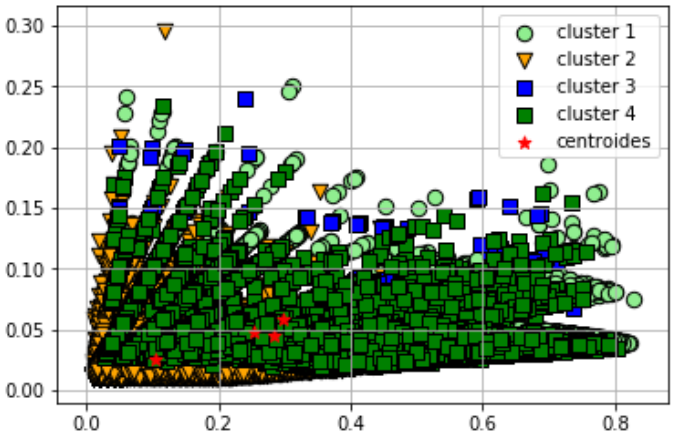


Figura 8. Clusters obtenidos.

El producto obtenido no es muy claro, a pesar de que los objetos pertenecientes a cada grupo se distinguen entre sí, no parecen tener realmente un orden por lo que es difícil visualizar. Esto se puede solucionar aplicando lo que se conoce como Principal Component Analysis, por sus siglas en ingles PCA.

3.2.5 Aplicación de PCA

Una vez que se han generado los clusters, visualmente es complicado distinguir los grupos y los objetos pertenecientes a cada uno, lo recomendable es aplicar lo que se conoce como PCA, una técnica que sirve para reducir la dimensionalidad de un conjunto de datos. La imagen mostrada en la figura 9 presenta el resultado de aplicar PCA a los clusters generados, ahora es posible identificar un cluster de otro.

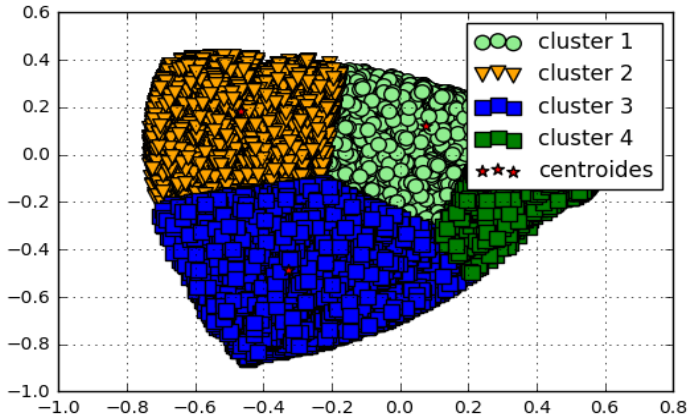


Figura 9. Clusters con PCA.

3.2.6 Adición de columna Cluster.

Esta es la etapa final del clustering, una vez que se crearon los grupos de objetos de la tabla de Viviendas, lo siguiente es crear una columna nueva que lleve por nombre “Cluster”, en ella se almacenará el número de cluster al que pertenece cada objeto. Ahora se conocen a qué grupo pertenece cada registro de acuerdo con la Disponibilidad y Uso de Tecnologías de la Información en los Hogares. La figura 10 muestra la tabla Viviendas con los clusters.

	VIV_SEL	P1_1	P1_2	P1_3	P1_4	P1_5_1	P1_5_2	P2_1	P2_2	P2_3	cve_est	cve_mun	Cluster
0	5	2	1	1	1	1	1	4	1	1	1	1	2
1	1	3	1	1	1	1	1	7	1	1	1	1	2
2	3	3	1	1	1	1	1	8	1	1	1	1	2
3	4	3	1	1	1	1	1	5	1	1	1	1	2
4	5	3	1	1	1	1	1	6	1	1	1	1	2

Figura 10. Columna Cluster.

3.2.7 Generación de diccionario.

Una vez que la tabla de Viviendas está completa se toman las columnas clave de estado, clave de municipio y número de cluster para crear una tabla nueva, a esta se le llamará diccionario. Posteriormente se hace una agrupación de los datos a partir de la clave de municipio y clave de

estado, esto con la finalidad de calcular la moda de la columna clusters en relación con el estado y municipio, es decir, obtener el número de cluster que aparece con más frecuencia en un estado, el valor que arroja la moda es el que se asigna a la columns.

En la figura 11 se muestra la manera en la que se obtiene la moda usando python. Posteriormente en la figura 12 se puede visualizar como queda el diccionario de Viviendas.

```
df_mode = df.groupby(['cve_est', 'cve_mun']).CLUSTER.apply(lambda x: x.mode())
df_mode.head(50)
```

Figura 11. Proceso de obtención de moda.

	cve_est	cve_mun	CLUSTER
0	1	1	2
1	1	2	2
2	1	3	2
3	1	4	2
4	1	5	2

Figura 12. Diccionario de viviendas.

3.3 Tabla de incidencias.

En esta fase del proyecto se toma una tabla de incidencias delictivas del año 2015, la cual recopila información de diferentes delitos cometidos en los municipios de los estados mexicanos. Los datos que se pueden observar son totales de delitos efectuados por mes, total de delitos por año, tipo de delito, modalidad, clave de estado y clave de municipio. Esta tabla y el diccionario de viviendas comparten las columnas clave de estado y clave de municipio lo que permite relacionarlas entre sí para efectuar las predicciones.

En la figura 13 se puede apreciar la información de la tabla de incidencias del año 2015.

CVE_ENT	CVE_MUN	CVE_MOD	CVE_TIPO	CVE_SUBTIP	ENERO	FEBRERO	MARZO	ABRIL	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	NOVIEMBRE
1	1	6	15	16	4	3	0	7	4	6	4	3	4	1	4
1	1	6	15	17	9	10	10	11	11	6	10	18	8	3	14
1	1	6	15	18	4	1	5	3	0	1	1	4	2	5	6
1	1	6	15	19	0	0	0	0	1	0	0	0	0	0	1
1	1	6	15	20	34	32	46	44	49	34	49	56	34	24	49
1	1	6	15	10	4	2	1	2	4	4	3	4	8	1	3
1	1	6	15	6	0	0	0	0	0	0	0	0	0	0	0
1	1	6	16	16	172	151	181	155	170	182	173	168	158	144	201
1	1	6	16	17	96	68	107	97	96	89	80	105	91	101	79
1	1	6	16	18	114	106	122	124	138	94	89	93	126	105	152
1	1	6	16	19	0	0	0	0	0	1	0	0	0	0	0
1	1	6	16	20	86	131	143	106	94	86	78	79	87	56	79
1	1	6	16	10	190	206	231	188	201	177	176	186	185	203	226
1	1	6	16	6	0	0	0	0	0	0	0	0	0	0	0
1	1	7	17	21	4	6	2	6	5	0	9	5	7	7	11
1	1	9	15	25	1	0	1	0	0	0	0	0	0	0	0
1	1	9	15	26	0	0	0	0	0	0	0	0	0	0	0
1	1	9	15	27	0	0	0	0	0	0	0	0	0	0	0
1	1	9	15	28	0	0	0	0	0	0	0	0	0	0	0
1	1	9	15	10	0	0	0	0	0	0	0	0	0	0	0
1	1	9	15	6	0	0	0	0	0	0	0	0	0	0	0
1	1	9	16	25	0	0	0	0	0	0	0	0	0	0	0

Figura 13. Tabla de incidencias.

Lo siguiente es tomar las columnas clave de estado, clave de municipio y total de delitos de la tabla Incidencias para crear un nuevo data frame. Enseguida se agrega una columna con el nombre VIVIENDAS en la que se registrará la información del diccionario. En la figura 14 se muestra el data frame creado.

	CVE_ENT	CVE_MUN	TOTAL_DELITOS	VIVIENDAS
0	1	1	43	0
1	1	1	115	0
2	1	1	32	0
3	1	1	2	0
4	1	1	449	0

Figura 14. Adición de columna Cluster.

Para agregar los valores a la columna VIVIENDAS se deben iterar tanto el data frame generado como el diccionario, cuando los campos de clave de municipio y clave de estado coincidan, se tomará el valor de la columna CLUSTER del diccionario y se registra en columna. Entonces se obtiene lo que muestra la figura 15.

	CVE_ENT	CVE_MUN	TOTAL_DELITOS	VIVIENDAS
0	1	1	43	2
1	1	1	115	2
2	1	1	32	2
3	1	1	2	2
4	1	1	449	2

Figura 15. Resultado de iteración.

La información contenida en esta tabla será utilizada para generar un mapa geo codificado donde se podrán visualizar de forma gráfica las agrupaciones de los municipios de acuerdo con sus características similares.

3.4 Aprendizaje supervisado.

El aprendizaje supervisado se suele usar en problemas de clasificación, como identificación de dígitos, diagnósticos, detección de fraude de identidad, etc. También se emplea en problemas de regresión, como predicciones meteorológicas, de expectativa de vida, de crecimiento entre otros. Estos dos tipos principales se distinguen por el tipo de variable objetivo, es decir, el tipo de dato que se desea obtener, para los casos de clasificación, se busca encontrar una variable categórica que se puede definir como cualitativa, mientras que, en los casos de regresión, la variable objetivo es numérica definida como cuantitativa.

En vista de que uno de los objetivos de esta investigación es el de predecir el número de delitos cometidos en cada estado del país mexicano, es obvio que se debe emplear la regresión, puesto que el valor que se desea encontrar es de carácter numérico.

3.4.1 Regresión.

Después de haber generado los clusters e integrado su información a la tabla de incidencias delictivas, el próximo paso nos lleva a la predicción del número delitos por estado. Esto es posible

aplicando uno de los tantos algoritmos de regresión que ofrece la biblioteca sklearn de python. Su funcionamiento consiste en recibir como parámetros de entrada los datos con los que se debe entrenar, para este caso la última tabla generada es la que se utilizará como entrada. Los valores contenidos en ella son analizados por el algoritmo y a partir de estos debe responder a nuestra interrogante, ¿cuál es el número de delitos que se cometen por estado? Esto lo hace basándose en los datos que ya analizó, como si se tratara de un examen que debe resolver, y le fuera proporcionada una guía para estudiar.

En un proyecto de esta naturaleza se recomienda dividir la información que se tomará como parámetro de entrada para entrenar el algoritmo. Se debe coger un 80% para entrenamiento y un 20% para realizar pruebas.

3.4.1.3 Algoritmo Gaussian Naive Bayes.

Entre los algoritmos explorados se descubrió Gaussian Naive Bayes, este es simple y fácil de implementar, además es útil cuando se trabaja con valores continuos. Para comenzar con el entrenamiento se proporciona la tabla de incidencias como parámetro de entrada. Los registros contenidos en ella se dividen en dos partes, 80 y 20 por ciento, se tomará la mayor fracción de información para entrenar el algoritmo y el resto para realizar pruebas.

En la figura 16 se muestra la selección de la información que será utilizada para entrenar el algoritmo, usando python.

```
In [57]: from sklearn.model_selection import train_test_split

In [58]: X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.2, random_state=42)

In [28]: X_train.shape
Out[28]: (42926, 18)

In [29]: X_test.shape
Out[29]: (10732, 18)

In [30]: y_train.shape
Out[30]: (42926, 1)

In [42]: y_test.shape
Out[42]: (10732, 1)
```

Figura 16. Proceso de separación de datos.

Una vez que se determina los valores con que se entrenará, se prosigue con el entrenamiento, para esto se utiliza el método fit de la clase GaussianNB, dándole como parámetro el 80% de los registros.

En la figura 17 se muestra el entrenamiento del algoritmo Gaussian Naive Bayes usando python.

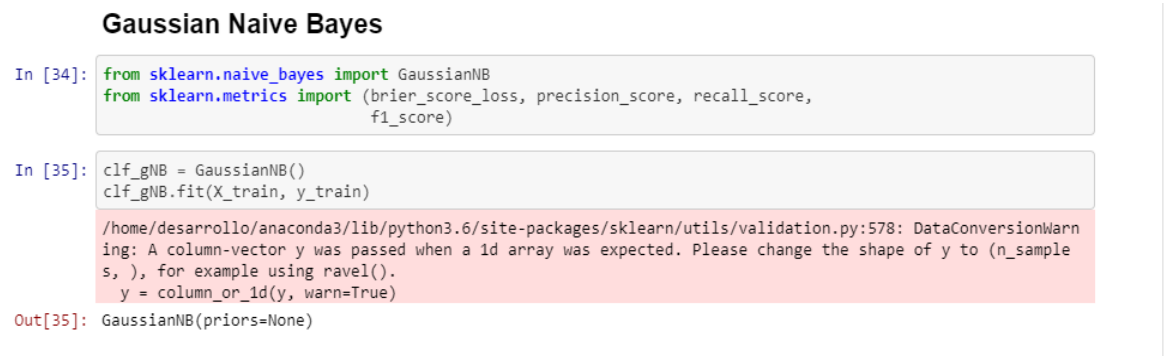


Figura 17. Entrenamiento con algoritmo de regresión.

Cuando se ha terminado de entrenar, lo que sigue es realizar las predicciones. Para ello se toma el 20% de los registros que se habían dividido en un principio para insertarlos como parámetro del método predict de la clase GaussianNB de sklearn. El resultado que arroja el método es un array de números enteros que representan la cantidad de delitos que se cometen por estado tal como lo muestra la figura 18.



Figura 18. Predicción con algoritmo de regresión.

Después de que un algoritmo arroja un resultado surge la duda de si es correcto o no. En el caso del aprendizaje supervisado existe una manera de conocer en qué porcentaje acertó o falló. A esto es a lo que se le conoce como performance, puede ser calculado con el método score de la clase GaussianNB.

Es importante tener en cuenta que un porcentaje muy alto de performance no es un resultado positivo, ya que puede indicar que el algoritmo fue sobreentrenado y se ajustó demasiado a los parámetros que le fueron proporcionados que al hacer pruebas con información distinta será más difícil responder de forma acertada, esto es lo que conocemos como overfitting. Por el contrario, underfitting se da cuando el porcentaje es muy bajo, lo cual es igualmente negativo indicando que el entrenamiento fue deficiente.

El performance de esta investigación presenta un grado de 78%, lo cual quiere decir que se equivocó en un 22% al predecir el número de delitos cometidos por estado. Es considerado un número positivo ya que indica que no presenta índices de overfitting o underfitting.

En la figura 19 se muestra la manera en la que se calcula el performance del algoritmo empleado por medio de python.

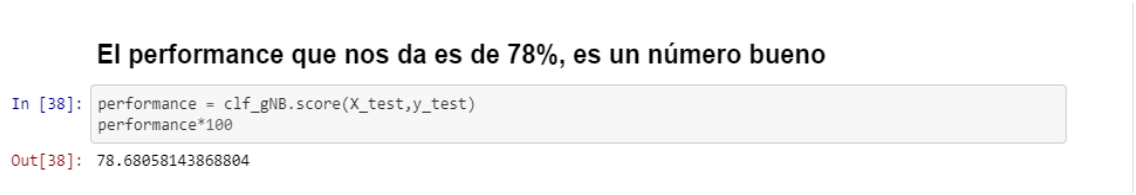


Figura 19. Performance del algoritmo de regresión.

3.5 Data Visualization.

La visualización de la información obtenida es la etapa final de data science. El objetivo de esta es representar de manera gráfica los resultados obtenidos, de modo que quien los observa pueda interpretarlos y entenderlos.

Existen diversas maneras de presentar de forma visual el producto final del análisis, por medio de gráficas, mapas geo codificados, tablas, etc. Todo depende de lo que se necesite mostrar.

3.5.1 Geocoding.

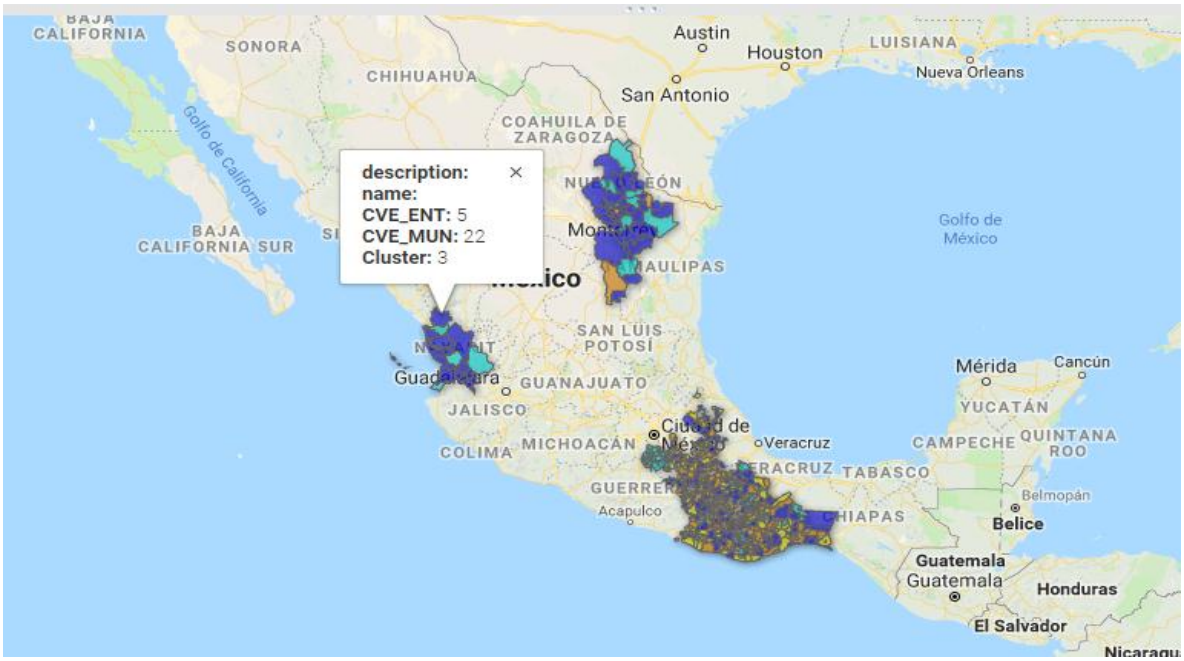
El geocoding permite asignar coordenadas geográficas a puntos de un mapa, en este caso se necesita mostrar los municipios de cada estado de México en un plano, indicando el cluster al que pertenecen, de acuerdo con el uso de tecnologías en los hogares.

Teniendo los clusters se debe proceder a convertir la base de datos en un sistema de información geográfica, para ello se utiliza un GIS software que provee la tecnología necesaria para hacer la conversión, además de archivos Shapefile de INEGI que almacenan la ubicación geométrica de las entidades geográficas, de él se toman las divisiones territoriales de México.

Se hace una unión entre el archivo Shapefile y la última tabla de incidencias generada que contiene los registros de clusters e incidentes de delincuencia, por medio de las claves de estado y claves de municipio del dataset y los del archivo de INEGI. A partir de esta unión se crea el mapa. Una vez creado se debe dar color a los objetos de los clusters para distinguirlos entre sí, posteriormente se hace la exportación a Google Maps, para ello es necesario exportar a un archivo KML, con el fin de hacerlo compatible con los servicios de Google.

Como resultado se obtiene un mapa de México donde se puede visualizar los municipios y su respectiva información como clave de entidad, clave de municipio y número del cluster al que pertenece.

Al observar el resultado que se muestra en la figura 20 se puede apreciar que no aparecen todos los municipios, la causa de ello es que no todas las claves de los municipios y estados del dataset



de incidencias coinciden con los del Shapefile.

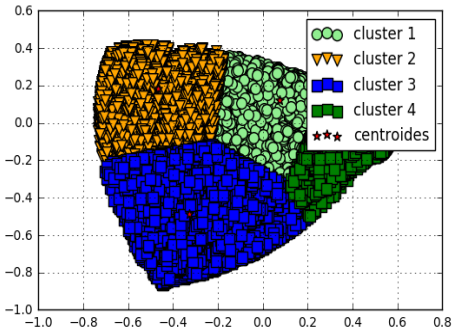
Figura 20. Mapa geocodificado de incidencias delictivas.

4 RESULTADOS

Los resultados obtenidos mediante el proceso de ciencia de datos fueron, la generación de cuatro grupos de datos, y la predicción del número de delitos en los municipios registrados en las encuestas de INEGI.

El producto final deja ver que el hecho de que dos municipios pertenezcan a un mismo estado no determina que estos posean las mismas características en cuestión de uso de tecnologías e incidencia delictiva. Municipios de estados del norte, del centro y sur del país pueden pertenecen a un mismo grupo. Sin embargo, es importante recordar que no todos los municipios están representados en el mapa, a causa de esto no es posible tener una visión completa de los resultados.

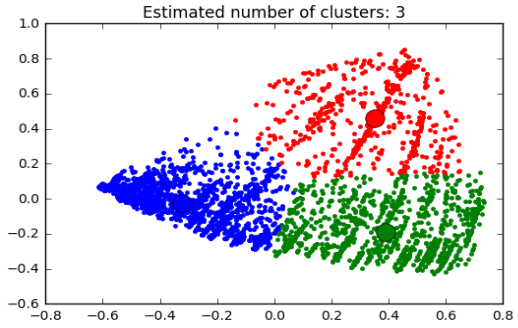
De acuerdo con el algoritmo KMeans, se determinó que el número de clusters óptimo es cuatro, sin embargo, podría haberse elegido otro valor y el producto habría tomado un camino distinto. Siendo este un algoritmo de aprendizaje no supervisado podemos percatarnos que no existe una métrica que indique cual es el porcentaje de error o acierto como en el aprendizaje supervisado, esto porque no le se le proporcionó ninguna clase de entrenamiento, por el contrario, debe aprender por si solo por medio del cálculo de distancias. Lo mismo sucede cuando se utiliza Mean Shift, la única diferencia radica en que este es quien determina el número óptimo de clusters, generando así un total de tres clusters como se ve en el proyecto “Caracterización Geográfica para temas delictivos con Data Science” aplicado a la misma información.



forma diferente

proyecto y del ti

s y conocer su



Un aspecto fundamental que se debe tomar en cuenta para un proyecto de esta naturaleza es el hardware con el que se trabaja. El procesamiento de una enorme cantidad de información como la que se utilizó en esta investigación requiere de un equipo con suficiente capacidad de memoria. Este es uno de los inconvenientes principales que pueden producir un análisis incompleto.

6 CONCLUSIONES

Las fases de este proyecto nos dirigen a un único objetivo, el de generar y compartir conocimiento. De nada sirve poseer o almacenar una gran cantidad de datos si estos no son explotados, ya sea para fines de investigación o lucrativos. El desarrollo de la tecnología ha permitido encontrar y mejorar la manera en la que se obtiene y almacena la información que generamos con nuestras actividades diarias. Conforme pasa el tiempo, el volumen de datos va creciendo exponencialmente, es por ello que, se necesita de una ciencia para trabajarlos, pues con su estudio, las posibilidades de mejorar procesos, predecir sucesos y encontrar comportamientos es exponencial, teniendo esto, será muy fácil actuar ante cualquier situación que se necesite. En el ámbito de la delincuencia, este proceso será de mucha ayuda ya que la información día con día es recolectada ya sea por videos, imágenes, sensores, etc. Teniendo esa información y aplicando las tecnologías de la ciencia de datos, los resultados cada vez serán mejores ayudando así a la erradicación de la delincuencia.

7 FUENTES DE INFORMACIÓN

1. Ulrich, K., & Eppinger, S. (2013). Diseño y desarrollo de productos (5ª edición ed.). México, D.F.: Mc Graw Hill.

2. Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python Exercises". Archived from the original on 23 June 2012.

3. Dipanjan Sarkar, Raghav Bali, Tushar Sharma. (2018). Practical Machine Learning with Python. India: Apress.

4. Manohar Swamynathan. (2017). Mastering Machine Learning with Python in Six Steps. India: Apress.

5. Nishant Shukla. (2017). Machine Learning with TensorFlow. India: Manning Publications.

6. William McKinney. (2018). Python for Data Analysis. United States of America: ORREILY.

7. David M. Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, Heidi Zimmer. (2007). Introduction to Statistics. University of Houston: n/a.

8. Dan Toomey. (2016). Learning Jupyter. Australia: Packt Publishing.