Instituto Tecnológico de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial Aplicada (MNA-V)

Análisis de grandes volúmenes de datos

Nombre del trabajo:

Proyecto

Autor:

Kevin Brandon Cruz Mejia

Objetivos:
Integrar el procesamiento, análisis, visualización y predicción con grandes volúmenes de datos mediante herramientas de código abierto utilizadas por los profesionales informáticos en sus bases de datos y sistemas.

Instrucciones:
Los modelos creados anteriormente, regresión y árboles de clasificación, se evalúan de manera diferente. El árbol de clasificación (al ser un método supervisado), se evalúa procesando el conjunto de prueba.

Crear un perfil socioeconómico de las diversas regiones del estado de Jalisco, México, incluyendo un modelo predictivo del crecimiento de sectores claves en el estado, a fin de dar una contextualización a la entrega final.

Considerando el trabajo anterior utilizar las variables detectadas que impactaron a la región que les fue asignada y buscar cubrir los siguientes elementos:

- Para el árbol de clasificación: mostrar imagen del árbol, mostrar la matriz de confusión al utilizar el conjunto de prueba.
- Una discusión final sobre el tema de evaluación de modelos.
- Presentar un reporte integrador del proyecto final con las siguientes secciones:
    - Introducción: descripción del objetivo y alcance.
    - Marco teórico: tecnologías utilizadas y panorama de la región de Jalisco.
    - Metodología: descripción de cada una de las etapas del proyecto: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado y evaluación de modelos.
    - Resultados y conclusiones: en esta sección se discuten los resultados del perfil socioeconómico y se reflexiona sobre el proyecto completo.
- De las fuentes utilizadas, incluyan citas y referencias en formato APA.

Reporte integrador

Introducción

Integrar los datos obtenidos en diferentes bases de datos del estado de Jalisco con herramientas de análisis de grandes volúmenes de datos para la generación de un modelo que permita generar un análisis socioeconómico de las diferentes regiones del estado de Jalisco.

Marco teórico

Panorama de Jalisco

Jalisco es un estado de la República Mexicana ubicado al Oeste de México. Dentro del estado se encuentra una zona económica importante a nivel nacional y es muy conocida por la producción de diferentes productos emblemáticos del país como el Tequila.

Su extensión territorial es de 78,588 km^2 y con una población de 8.348 millones de personas conforme el censo población realizado en 2020.

Ilustración 1: Estado de Jalisco en la República Mexicana

En el año de 1998 se instituyó la "Regionalización administrativa" en Jalisco, para promover el progreso de la entidad, congregándose los 125 municipios en 12 regiones, cada una asumiendo a un municipio sede, que hace la función de "capital" de la región. Estas regiones y sus municipios son los siguientes:

1. Región Norte
    a. Bolaños, Chimaltitlán, Cotolán, Huejúcar, Huejuquilla el Alto, Mezquitic, San Martin de los Bolaños, Totatiche, Villa Guerrero.
2. Región Altos Norte
    a. Encarnación de Díaz, Lagos de Moreno, Ojuelos de Jalisco, San Diego de Alejandría, San Juan de los Lagos, Teocaltiche, Unión de San Antonio, Villa Hidalgo.

3. Región Altos Sur
    a. Acatic, Arandas, Cañadas de Obregón, Jalostotitlán, Jesús María, Mexticacán, San Ignacio Cerro Gordo, San Julián, San Miguel el Alto, Tepatitlán de Morelos, Valle de Guadalupe, Yahaulica de González Gallo.
4. Región Ciénega
    a. Atotonilco el Alto, Ayotlán, Degollado, Jamay, La Barca, Ocotlán, Poncitlán, Tototlán, Zapotlán del Rey.
5. Región Sureste
    a. Chapala, Concepción de Buenos Aires, Jocotepec, La Manzanilla de la Paz, Mazamitla, Quitupan, Santa María del Oro, Tizapán el Alto, Tuxcueca, Valle de Juárez
6. Región sur
    a. Gómez Farías, Jilotlán de los Dolores, Pihuamo, San Gabriel, Tamazula de Gordiano, Tecalitlán, Tolimán, Tonila, Tuxpan, Zapotiltic, Zapotitlán de Vadillo, Zapotlán el Grande.
7. Región Sierra de Amula
    a. Atengo, Autlán de Navarro, Ayutla, Chiquilistlán, Cuautla, Ejutla, El Grullo, El limón, Juchitán, Tecolotlán, Tenamaxtlán, Tonaya, Tuxcacuesco, Unión de Tula.
8. Región Costa Sur
    a. Casimiro Castillo, Cihuatlán, Cuautitlán de García Barragán, La Huerta, Tomatlán, Villa Purificación.
9. Región Costa-Sierra Occidental
    a. Atenguillo, Cabo Corrientes, Guachinango, Mascota, Mixtlán, Puerto Vallarta, San Sebastián del Oeste, Talpa de Allende.
10. Región Valles
    a. Ahualulco de Mercado, Amatitán, Ameca, El Arenal, Etzatlán, Hostotipaquillo, Magdalena, San Juanito de Escobedo, San Marcos, Tala, Tequila, Teuchitlán,
11. Región Lagunas.
    a. Acatlán de Juárez, Amacueca, Atemajac de Brizuela, Atoyac, Cocula, San Martín Hidalgo, Sayula, Tapalpa, Techaluta de Montenegro, Teocuitatlán de Corona, Villa Corona, Zacoalco de Torres.
12. Región Centro
    a. Cuquío, El Salto, Guadalajara, Ixtlahuacán de los Membrillos, Ixtlahuacán del Río, Juanacatlán, San Cristóbal de la Barranca, San Pedro Tlaquepaque, Tlajomulco de Zúñiga, Tonalá, Zapopan, Zapotlanejo.

Instituto Nacional de Estadística y Geografía

El Instituto Nacional de Estadística y Geografía (INEGI) es un organismo público autónomo responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geográfica, así como de captar y difundir información de México en cuanto al territorio, los recursos, la población y economía, que permita dar a conocer las características de nuestro país y ayudar a la toma de decisiones.

El INEGI publica diferentes herramientas y estudios sobre datos estadísticos y geográficos de México, entre los que se encuentran:

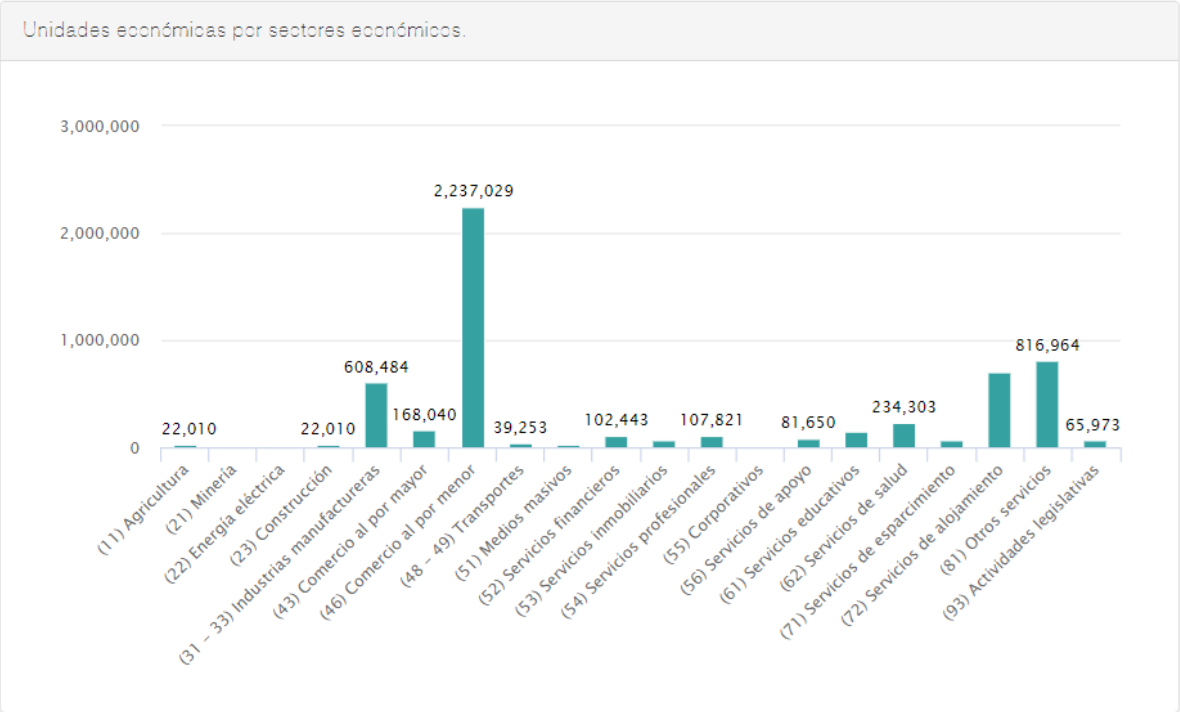
- Directorio Estadístico Nacional de Unidades Económicas

  En el Directorio Estadístico Nacional de Unidades Económicas (DENUE) se ofrecen los datos de identificación, ubicación, actividad económica y tamaño de los negocios activos en el territorio nacional, actualizados, fundamentalmente, en el segmento de los establecimientos grandes.

  Este directorio nos permite obtener información de todas las empresas y negocios a nivel nacional que han sido registrados. En la siguiente imagen, podemos observar que el estado de Jalisco cuenta con un número importante de empresas a nivel nacional.

Ilustración 2: Distribución de establecimientos en México

De la misma forma, de manera general, podemos identificar que en México comercio al por menor es el negocio más recurrente. Esto es porque en México se cuenta con cadenas de tiendas de conveniencia con gran presencia en todo el territorio nacional.

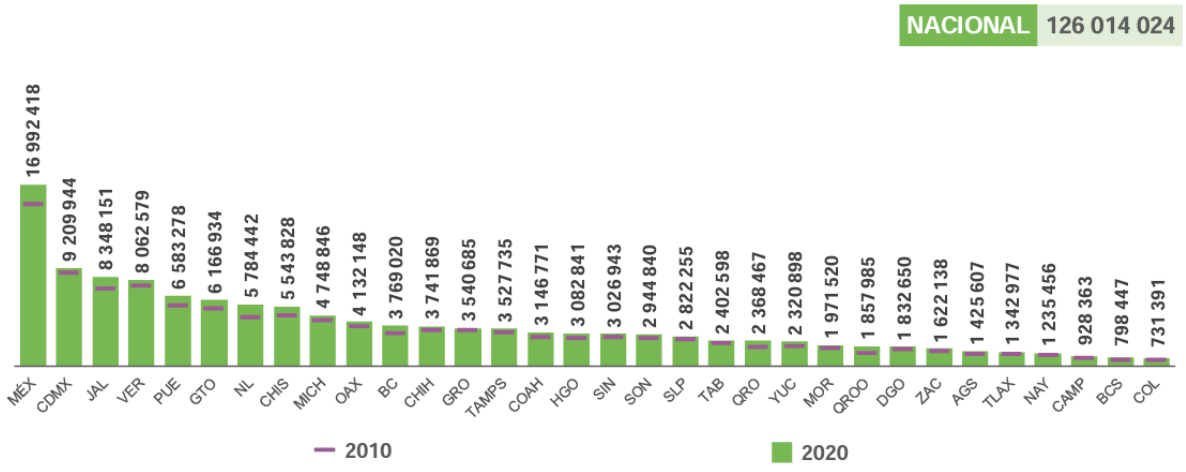


- <u>Censo de Población y Vivienda 2020</u>

  El Censo de Población y Vivienda 2020 (Censo 2020) se realizó del 2 al 27 de marzo del 2020; en éste participaron poco más de 147 mil entrevistadoras y entrevistadores que recorrieron los cerca de dos millones de kilómetros cuadrados del territorio nacional, visitando cada una de las viviendas para obtener información sobre estas, contar a la población que vive en México e indagar sobre sus principales características demográficas, socioeconómicas y culturales.

  El objetivo del censo fue producir información sobre la dimensión, estructura y distribución espacial de la población, así como de sus principales características socioeconómicas y culturales. Obtener la cuenta de viviendas y algunas de sus características. El censo muestra la continuación de tendencias demográficas, económicas y sociales.

  Conforme los datos del Censo, podemos identificar que Jalisco es el tercer estado más poblado a nivel nacional.



Entre otros datos, el Censo poblacional nos proporciona información sobre la educación, acceso a servicios y diferentes características de la población mexicana que nos permitirán generar un estudio socioeconómico.

Estudio Socioeconómico

Es una investigación que permite conocer el entorno social y económico de una persona, empresa o zona geográfica en particular, se trata de una investigación que tiene como objetivo dilucidar los aspectos propios de un sujeto de investigación. Los rubros para desplegar en el estudio socioeconómico dependerán de la finalidad de este.

Regresión lineal

En la estadística, la regresión lineal es el método para estimar la relación entre dos o más variables basado en datos observados. La regresión linear es el método utilizado cuando esta relación puede ser expresada por una ecuación linear, es decir el formato básico de la ecuación de la recta.

$$Y = ax + b$$

La línea descrita por esta ecuación es frecuentemente referida como la línea de regresión o línea de mejor ajuste, significando que cuando los datos son graficados en un diagrama de dispersión se agrupan sobre esta línea.

Árbol de clasificación

Un árbol de clasificación es una herramienta utilizada para ayudar en el análisis de opciones y determinar los pasos en un proceso. El árbol de clasificación es un método del aprendizaje supervisado.

Kmeans

El método de agrupamiento por K-means es uno de los algoritmos de agrupamiento más antiguos y utilizados. K-means es un algoritmo particional de agrupamiento que intenta encontrar K no encimados grupos (clusters) Estos clusters son representados por sus centroides (un centroide de un cluster normalmente es el promedio de los puntos en este cluster). El proceso de agrupamiento de K-means es el siguiente:

1. Primero los K centroides iniciales son seleccionados, donde K es especificada por el usuario e indica el número deseado de clusters.
2. Cada punto en los datos es asignado a él centroide más cercano y cada colección de puntos asignados a un centroide forma un cluster.
3. El centroide de cada cluster es después actualizado.

Metodología

Para el desarrollo del modelo de estudio socioeconómico de las regiones de Jalisco, se dividió el proceso en 7 fases:

1. Historias de usuario (SCRUM)
   El desarrollo de las historias de usuario con la metodología SCRUM, consistió en definir los pasos a seguir del proyecto, así como la funcionalidad del modelo. Este es un resumen de lo generado:
   - Título del proyecto: Proyecto del Curso - Fase 1 Entendimiento del proyecto
     - Descripción:
       Como equipo realizaremos el entendimiento de los datos, así como su tipo, que se requiere extraer para poder tener en la base la información unificada y así crear un estudio socioeconómico. El foco escogido es la zona céntrica del estado de Jalisco. Utilizando datos de empresas, censos y bases de datos públicas.
   - Título del proyecto: Proyecto del Curso - Fase 1 Entendimiento del proyecto
     - Descripción:
       Como equipo crearemos las historias con el framework de scrum en un documento para
       organizar las actividades del proyecto enfocándonos en los objetivos y alcances de este.
   - Título del proyecto: Proyecto del Curso - Fase 2. Entendimiento de los datos
     - Descripción:
       Como analistas de negocio vamos a buscar relaciones entre el éxito de las empresas que haya dentro de nuestra zona de enfoque para poder con eso determinar si es que las empresas de diversos sectores están siendo influenciadas por sectores de zona, inversión, giro empresarial o capital humano.
   - Título del proyecto: Proyecto del Curso - Fase 3. Procesamiento de datos
     - Descripción:
       Con las herramientas de Hadoop como HDFS, poder aplicar el paradigma MapReduce para su limpieza, extracción y descripción. Usar Pentaho y ApachePig para poder agregar valor a los datos con la creación de variables y agregar características.
   - Título del proyecto: Proyecto del Curso - Fase 4. Modelado
     - Descripción:
       Generación de modelos de segmentación, KPIs desglosados y con valor para la
       toma de decisiones, uso de regresiones, árboles de decisión con sus respectivas métricas las cuales van a medir la predicción tanto como de predicciones como de modelados del sistema actual.
   - Título del proyecto: Proyecto del Curso - Fase 5. Presentación de resultados
     - Descripción:
       Creación de tablero de presentación de los datos con los KPIs más relevantes en
       modo de presentación interactiva.

2. Diccionario de datos y esquemas de bases de datos

Para la elaboración del proyecto, se nos solicitó consultar el DENUE de la región de Jalisco y enriquecer los datos con otras fuentes. En este sentido se seleccionaron los siguientes datos:

- Censo poblacional 2020.
- índice de mortalidad
- Nivel educativo de la población
- Parque vehicular.

Esta información está segregada por municipio, por lo que fue necesaria una tabla adicional, donde se incorporaran los municipios por cada región de Jalisco.

Durante la elaboración del diccionario de datos, se estandarizó la información y se eliminaron los datos que no aportarían al análisis.

Conforme el análisis realizado y aplicando el paradigma MapReduce de HortonWorks se crearon los diccionarios que se encuentran al calce, por lo que se definió como key la variable “desc_municipio” para propósitos de entidad relación para la base de datos.

Registro de Mortalidad del INEGI:

Field Name	Type	Mode	Description	Example
desc_municipio	Alfanumeric	Required	Es el sustantivo propio que identifica al Municipio	Estatal
indicador	Alfanumeric	Optional	Número de Defunciones por año	Defunciones generales
año	Numeric	Optional	Año en que fue medido el indicador	1994
valor	Numeric	Optional	Valor numérico de la característica medida en el indicador	29906

Registro de DENUE:

Field Name	Type	Mode	Description	Example
nombre_act	alfanumeric	Optional	Nombre del código de actividad conforme al SCIAN 2018.	Piscicultura y otra acuicultura, excepto camaronicultura
per_ocu	alfanumeric	Optional	Comprende al personal contratado directamente por la razón social y al personal ajeno suministrado por otra razón social, que trabajó para la unidad económica, sujeto a su dirección y control y que cubrió, como mínimo, una tercera parte de la jornada laboral. Puede ser personal de planta, eventual remunerado o no remunerado. Las unidades económicas se clasifican por rangos de personal ocupado, que permiten identificar el tamaño de Unidades Económicas por el número de personal que emplean; es decir, según su personal ocupado total. Para el Directorio las Unidades Económicas están agrupadas en rangos o estratos de personal ocupado, con base en el personal ocupado total reportado por éstas, como se indica en la siguiente tabla. Personal Ocupado 1 = 0 a 5 2 = 6 a 10 3 = 11 a 30 4 = 31 a 50 5 = 51 a 100 6 = 101 a 250 7 = 251 y más	11 a 30 personas
municipio	alfanumeric	required	Es el sustantivo propio que identifica al Municipio, y en el caso de la Ciudad de México los que identifican a las Alcaldías.	Zapopan
tipoUniEco	alfanumeric	Optional	Grupo de categorías que sirve para identificar si el establecimiento es fijo o semifijo, bajo las siguientes definiciones: Establecimiento fijo: es la unidad que en una sola ubicación física, asentada en un lugar de manera permanente y delimitada por construcciones e instalaciones fijas, combina acciones y recursos bajo el control de una sola entidad propietaria o controladora para realizar alguna actividad económica con y sin fines de lucro, excepto las destinadas al autoconsumo. En este tipo de establecimiento están comprendidas las viviendas cuando en algún espacio de la casa-habitación, que también está destinado a otras actividades cotidianas, se realiza alguna actividad económica. Establecimiento semifijo: establecimientos que están fijos, enclavados en el suelo, y que permanecen de día y de noche en el mismo sitio aunque sus instalaciones sean frágiles o rústicas. También aquellas instalaciones que aun cuando no estén enclavadas al suelo no pueden ser removidas de donde están, aunque lleven a cabo actividades que tienen carácter temporal o estacional, siempre y cuando permanezcan activas y en un solo lugar durante un periodo de por lo menos tres meses seguidos.	Fijo

Registro Censo Poblacional y Vivienda

Field Name	Type	Mode	Description	Example
desc_municipio	Alfanumeric	required	Es el sustantivo propio que identifica al Municipio	El Salto
indicador	Alfanumeric	optional	Característica medida por el indicador	Viviendas particulares habitadas que disponen de drenaje
1900	Numeric	optional	Valor numérico de la característica medida en el indicador	14.4
1910	Numeric	optional	Valor numérico de la característica medida en el indicador	1208855
1921	Numeric	optional	Valor numérico de la característica medida en el indicador	1191957
1930	Numeric	optional	Valor numérico de la característica medida en el indicador	1255346
1940	Numeric	optional	Valor numérico de la característica medida en el indicador	1418310
1950	Numeric	optional	Valor numérico de la característica medida en el indicador	1746777
1960	Numeric	optional	Valor numérico de la característica medida en el indicador	2443261
1970	Numeric	optional	Valor numérico de la característica medida en el indicador	3296586
1980	Numeric	optional	Valor numérico de la característica medida en el indicador	4371998
1990	Numeric	optional	Valor numérico de la característica medida en el indicador	5302689
1995	Numeric	optional	Valor numérico de la característica medida en el indicador	34679
2000	Numeric	optional	Valor numérico de la característica medida en el indicador	41554
2005	Numeric	optional	Valor numérico de la característica medida en el indicador	55815
2010	Numeric	optional	Valor numérico de la característica medida en el indicador	69220
2015	Numeric	optional	Valor numérico de la característica medida en el indicador	7880539
2020	Numeric	optional	Valor numérico de la característica medida en el indicador	111162
unidad_medida	Alfanumeric	optional	Nombre sustantivo que define a la medida del valor numérico	Número de personas

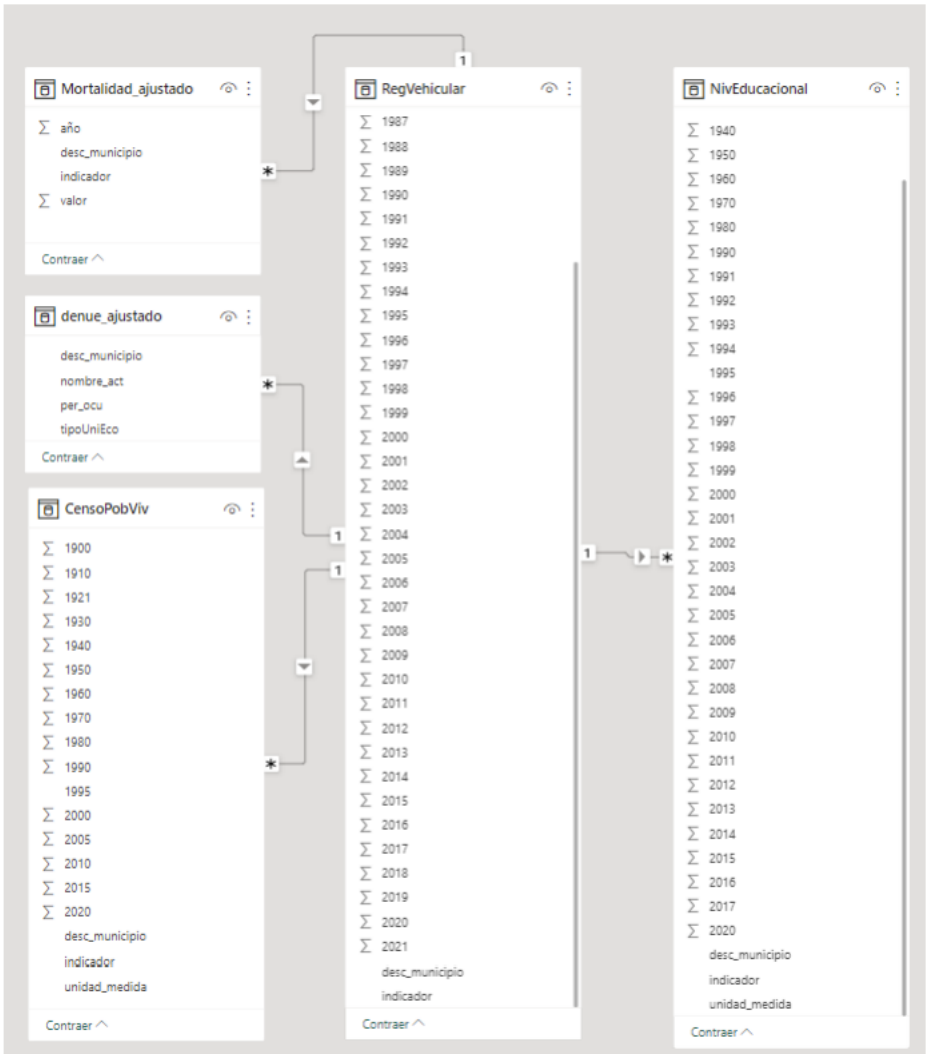
Registro de Parque Vehicular

Field Name	Type	Mode	Description	Example
desc_municipio	Alfanumeric	Required	Es el sustantivo propio que identifica al Municipio	Ameca
indicador	Alfanumeric	Optional	Es la característica del nivel educativo de la población	Vehículos de motor registrados en circulación
1980	Numeric	Optional	Valor numérico de la característica medida en el indicador	1762
1981	Numeric	Optional	Valor numérico de la característica medida en el indicador	2193
1982	Numeric	Optional	Valor numérico de la característica medida en el indicador	2529
1983	Numeric	Optional	Valor numérico de la característica medida en el indicador	2804
1984	Numeric	Optional	Valor numérico de la característica medida en el indicador	2944
1985	Numeric	Optional	Valor numérico de la característica medida en el indicador	2968
1986	Numeric	Optional	Valor numérico de la característica medida en el indicador	3117
1987	Numeric	Optional	Valor numérico de la característica medida en el indicador	3209
1988	Numeric	Optional	Valor numérico de la característica medida en el indicador	3363
1989	Numeric	Optional	Valor numérico de la característica medida en el indicador	3548
1990	Numeric	Optional	Valor numérico de la característica medida en el indicador	4474
1991	Numeric	Optional	Valor numérico de la característica medida en el indicador	4908
1992	Numeric	Optional	Valor numérico de la característica medida en el indicador	5299
1993	Numeric	Optional	Valor numérico de la característica medida en el indicador	18464
1994	Numeric	Optional	Valor numérico de la característica medida en el indicador	8807
1995	Numeric	Optional	Valor numérico de la característica medida en el indicador	9580
1996	Numeric	Optional	Valor numérico de la característica medida en el indicador	9915
1997	Numeric	Optional	Valor numérico de la característica medida en el indicador	10258
1998	Numeric	Optional	Valor numérico de la característica medida en el indicador	10433
1999	Numeric	Optional	Valor numérico de la característica medida en el indicador	9728
2000	Numeric	Optional	Valor numérico de la característica medida en el indicador	10875
2001	Numeric	Optional	Valor numérico de la característica medida en el indicador	12657
2002	Numeric	Optional	Valor numérico de la característica medida en el indicador	12229
2003	Numeric	Optional	Valor numérico de la característica medida en el indicador	13985
2004	Numeric	Optional	Valor numérico de la característica medida en el indicador	15165
2005	Numeric	Optional	Valor numérico de la característica medida en el indicador	15928
2006	Numeric	Optional	Valor numérico de la característica medida en el indicador	17611
2007	Numeric	Optional	Valor numérico de la característica medida en el indicador	18854
2008	Numeric	Optional	Valor numérico de la característica medida en el indicador	20179
2009	Numeric	Optional	Valor numérico de la característica medida en el indicador	20972
2010	Numeric	Optional	Valor numérico de la característica medida en el indicador	21467
2011	Numeric	Optional	Valor numérico de la característica medida en el indicador	22735
2012	Numeric	Optional	Valor numérico de la característica medida en el indicador	23946
2013	Numeric	Optional	Valor numérico de la característica medida en el indicador	24884
2014	Numeric	Optional	Valor numérico de la característica medida en el indicador	25897
2015	Numeric	Optional	Valor numérico de la característica medida en el indicador	26232
2016	Numeric	Optional	Valor numérico de la característica medida en el indicador	26720
2017	Numeric	Optional	Valor numérico de la característica medida en el indicador	28118
2018	Numeric	Optional	Valor numérico de la característica medida en el indicador	31031
2019	Numeric	Optional	Valor numérico de la característica medida en el indicador	32723
2020	Numeric	Optional	Valor numérico de la característica medida en el indicador	34425
2021	Numeric	Optional	Valor numérico de la característica medida en el indicador	36490

Registro de Nivel Educativo INEGI

Field Name	Type	Mode	Description	Example
desc_municipio	alfanumeric	Required	Es el sustantivo propio que identifica al Municipio	Zapopan
indicador	alfanumeric	Optional	Es la característica del nivel educativo de la población	Porcentaje de personas de 15 años y más alfabetas
1910	Numeric	Optional	Valor numérico de la característica medida en el indicador	64.1
1921	Numeric	Optional	Valor numérico de la característica medida en el indicador	57.5
1930	Numeric	Optional	Valor numérico de la característica medida en el indicador	58
1940	Numeric	Optional	Valor numérico de la característica medida en el indicador	47.3
1950	Numeric	Optional	Valor numérico de la característica medida en el indicador	38.8
1960	Numeric	Optional	Valor numérico de la característica medida en el indicador	30
1970	Numeric	Optional	Valor numérico de la característica medida en el indicador	21
1980	Numeric	Optional	Valor numérico de la característica medida en el indicador	13.2
1990	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.998796377
1991	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.001600723
1992	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.989483251
1993	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.983130953
1994	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.977377109
1995	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.970252348
1996	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.975543652
1997	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.98168333
1998	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.990909202
1999	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.98971899
2000	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.997943725
2001	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.994293414
2002	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.995160081
2003	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.000651063
2004	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.003072557
2005	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.003130542
2006	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.00259345
2007	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.007515535
2008	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.005098526
2009	Numeric	Optional	Valor numérico de la característica medida en el indicador	1.00382564
2010	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.997746467
2011	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.985726254
2012	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.996346978
2013	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.996861119
2014	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.99829556
2015	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.988819495
2016	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.987025057
2017	Numeric	Optional	Valor numérico de la característica medida en el indicador	32.53652598
2020	Numeric	Optional	Valor numérico de la característica medida en el indicador	0.180507018
unidad_medida	alfanumeric	Optional	Nombre sustantivo que define a la medida del valor numérico	Porcentaje

Por lo que, al generar la relación de la Base de Datos, el esquema se refleja de la siguiente forma:



3. Preprocesamiento de datos, integración de datos

Diseño de ETL:
A partir de los diccionarios de datos, en los archivos .csv se eliminaron los acentos, así como el municipio como llave por cada fuente para integrar los datos.

En Hive

Se uso el siguiente SQL por cada fuente de datos:

```
SELECT * FROM CENSOPOBLACIONVIVIENDA D
INNER JOIN REGIONJALISCOCENTRO RJ
ON D.DESC_MUNICIPIO = RJ.MUNICIPIO
WHERE RJ.REGION = 'CENTRO' LIMIT 10;
```

Utilizando Pig

```
DRIVERS = LOAD './JALISCOCENTRO/MORTALIDAD.CSV' USING PIGSTORAGE(',');
DRIVERS = FILTER DRIVERS BY $0>1;
DRIVERS_DETAILS = FOREACH RAW_DRIVERS GENERATE $0 AS DESC_MUNICIPIO, $1 AS VALOR;
TIMESHEET = LOAD './JALISCOCENTRO/REGIONESJALISCO.CSV' USING PIGSTORAGE(',');
RAW_TIMESHEET = FILTER TIMESHEET BY $0>1;
TIMESHEET_LOGGED = FOREACH RAW_TIMESHEET GENERATE $0 AS MUNICIPIO, $2 AS CLAVE, $3 AS
REGION;
GRP_DENUE = GROUP DRIVERS_DETAILS BY DESC_MUNICIPIO;
SUM_LOGGED = FOREACH GRP_DENUE GENERATE GROUP AS DESC_MUNICIPIO,
SUM(DRIVERS_DETAILS.VALOR) AS SUM_VALOR,
SUM(DRIVERS_DETAILS.VALOR) AS SUM_MILESLOGGED;
JOIN_SUM_LOGGED = JOIN SUM_LOGGED BY DESC_MUNICIPIO, TIMESHEET_LOGGED BY MUNICIPIO;
JOIN_DATA = FOREACH JOIN_SUM_LOGGED GENERATE $0 AS DESC_MUNICIPIO, $3 AS REGION, $2 AS
CLAVE, $1 AS VALOR;
DUMP JOIN_DATA;
```

Filtra después del header del archivo. Donde con dos fuentes de consulta se agrupa por municipio para contarlos para Mortalidad.csv y RegionJalisco.csv.

Obtenemos el Job, así como el resultado, en este ejemplo con 0 records como se espera al no tener los key en igualdad, así como su agrupamiento, por ello el uso del map reduce con el <key,value> para obtener mejores resultados. Cíclicamente se usará el script para agrupar las fuentes por municipio y realizar el análisis para tener la región centro en un perfil socioeconómico.

- Hive para realizar consultas tipo SQL

- Pig para realizar las mismas en un script con map reduce así tener mejores tiempos de procesamiento

Cumpliendo con el ETL, extraer de diferentes fuentes de consulta, transformar con scripts y/o queries a datos de interés para armar el perfil socioeconómico y cargar los resultados en un formato como el, csv para generar el perfil y si es necesario regresar al paso 1 para depurar e interpretar mejor los datos para mejorar la visualización de datos.

4. Ingeniería de datos: enriquecer, agregar, generar features

   Analizando los datos, se detectó que es posible generar nuevas variables (features) que nos sean de utilidad, esto se hace generando una operación matemática o filtrado con los datos ya existentes de nuestras bases de datos.

   En el conjunto de datos de DENUE podemos obtener las siguientes nuevas características:

   - Número de empresas por municipio.
     El número de empleados de una empresa refleja su tamaño e ingresos, por la capacidad de crecimiento. Se puede contar el número de empresas agrupándolas por tamaño, esto considerando el número de empleados, y así tener un volumen de empresas por municipio por tamaño de la empresa.
     El siguiente Query nos sirve para obtener los datos:

```
SELECT COUNT(DENUE.ID), DENUE.PER_OCU, DENUE.MUNICIPIO
FROM DENUE
GROUP BY DENUE.MUNICIPIO, DENUE.PER_OCU
```

   - Antigüedad en meses de las empresas
     La antigüedad de las empresas nos puede ayudar a interpretar si el desarrollo es reciente, o si se trata de una evolución constante. Se puede calcular la antigüedad de las empresas en meses al consultar la fecha de alta y el número de meses entre la fecha actual y la fecha anteriormente mencionada.
     El siguiente Query nos sirve para obtener los datos:

```
SELECT
TEMP.FECHA_ALTA_DATE AS FECHA_ALTA,
MONTHS_BETWEEN(CURRENT_DATE(), TEMP.FECHA_ALTA_DATE) AS ANTIGUEDAD_MESES
FROM
(+SELECT CAST(CONCAT(SUBSTR(FECHA_ALTA, 1, 4), '-', SUBSTR(FECHA_ALTA, 6, 2), '-01')
AS DATE) AS FECHA_ALTA_DATE
 FROM DENUE) AS TEMP
LIMIT 10
```

   En el conjunto de datos de Mortalidad podemos obtener las siguientes nuevas características:

   - Defunciones de niños menores de un año
   -

     Las muertes infantiles son un indicador del nivel de desarrollo y acceso a la salud, por lo que su cálculo pudiera ayudar para el estudio socio económico. Se requiere consultar de la base de datos de mortalidad y filtrar por las muertes para menores de un año.
     El siguiente Query nos sirve para obtener los datos:

```
SELECT count(mortalidad.id), mortalidad.indicador, mortalidad.desc_municipio
FROM mortalidad
GROUP BY mortalidad.desc_municipio, mortalidad.indicador
WHERE mortalidad.indicador CONTAINS ‘Defunciones de menores de un año’
```

- Defunciones generales entre población de 15 años y más con escolaridad básica.

  Las comunidades en las que las personas no tienen acceso a la educación son más propensas a participar en actividades que ponen en riesgo su vida, así como la dificultad para el acceso a servicios médicos. Podemos tomar datos de diferentes bases de datos para realizar este cálculo.
  De la Base de datos de Mortalidad consultamos las defunciones generales por municipio. De la base de datos de Censo Poblacional de Vivienda, consultamos la Población Total por municipio. Finalmente, de la base de datos de Mortalidad, consultamos las defunciones generales por municipio.

  Para generar el cálculo, Multiplicamos el porcentaje de población de 15 años y más con escolaridad básica por municipio, por la población total por municipio en cada año. Con este dato como denominador, dividimos las defunciones generales entre la población calculada para obtener una proporción.

  El siguiente Query que se está desarrollando para la nueva característica es de la siguiente forma:

  SELECT count(mortalidad.id), mortalidad.indicador, mortalidad.desc_municipio
  FROM mortalidad
  GROUP BY mortalidad.desc_municipio, mortalidad.indicador
  WHERE mortalidad.indicador CONTAINS ‘Defunciones generales’
  FROM CensoPoblacionVivienda
  GROUP BY CensoPoblacionVivienda.desc_municipio,
  CensoPoblacionVivienda.indicador
  WHERE CensoPoblacionVivienda.indicador CONTAINS ‘Población total’
  FROM NiveEducativo
  GROUP BY NiveEducativo.desc_municipio, NiveEducativo.indicador
  WHERE NiveEducativo.indicador CONTAINS ‘Porcentaje de población de 15 años y más con escolaridad básica

5. Almacenamiento de información en bases de datos NoSql
   La información se encontraba en bases de datos no SQL, por lo que se tuvo que proceder con la integración de la información en consultas en SQL y se subieron a MongoDB.
6. Modelado de datos, regresión lineal, perfiles socioeconómicos de las regiones de Jalisco

7. Modelado de datos, árboles de clasificación, perfiles socioeconómicos de las regiones de Jalisco

databricks FASe4a

(https://databricks.com)

```
import pandas as pd
import numpy as np
import sys
from pyspark.sql import SparkSession
```

```
%fs ls FileStore/tables
```

Table

	path	name	size	modificationTime
1	dbfs:/FileStore/tables/Censo_2020.xlsx	Censo_2020.xlsx	7898	1678063262000
2	dbfs:/FileStore/tables/clee_denue.csv	clee_denue.csv	65653	1678071114000
3	dbfs:/FileStore/tables/denue_inegi_14_-1.csv	denue_inegi_14_-1.csv	176534554	1678065888000
4	dbfs:/FileStore/tables/denue_inegi_14_-2.csv	denue_inegi_14_-2.csv	176534554	1678067798000
5	dbfs:/FileStore/tables/denue_inegi_14_-3.csv	denue_inegi_14_-3.csv	176534554	1678068695000
6	dbfs:/FileStore/tables/denue_inegi_14_.csv	denue_inegi_14_.csv	176534554	1678064529000
7	dbfs:/FileStore/tables/denue_inegi_14_clean.csv	denue_inegi_14_clean.csv	155776826	1678069110000

7 rows

```
path = "/FileStore/tables/denue_inegi_14_clean.csv"
data = spark.read.csv(path, encoding="ISO-8859-1")
data.take(20)
```

```
Out[39]: [Row(_c0='id', _c1='clee', _c2='nom_estab', _c3='raz_social', _c4='codigo_act', _c5='per_ocu', _c6='tipo_vial',
_c7='nom_vial', _c8='tipo_v_e_1', _c9='nom_v_e_1', _c10='tipo_v_e_2', _c11='nom_v_e_2', _c12='tipo_v_e_3', _c13='nom_v_e_
3', _c14='numero_ext', _c15='letra_ext', _c16='edificio', _c17='edificio_e', _c18='numero_int', _c19='letra_int', _c20='t
ipo_asent', _c21='nomb_asent', _c22='tipoCenCom', _c23='nom_CenCom', _c24='num_local', _c25='cod_postal', _c26='cve_ent',
_c27='entidad', _c28='cve_mun', _c29='municipio', _c30='cve_loc', _c31='localidad', _c32='ageb', _c33='manzana', _c34='te
lefono', _c35='correoelec', _c36='www', _c37='tipoUniEco', _c38='latitud', _c39='longitud', _c40='fecha_alta'),
 Row(_c0='8624390', _c1='14120112512000022000000000U3', _c2='ACUACULTORES LOS CASTRO S.C. DE R.L. DE C.V.', _c3='ACUACULT
ORES LOS CASTRO SC DE RL DE CV', _c4='112512', _c5='0 a 5 personas', _c6='CIRCUITO', _c7='DE LAS CAÑAS', _c8='CALLE', _c9
='MOLINOS DEL VALLE', _c10='CALLE', _c11='AV DE LA MANCHA', _c12='CALLE', _c13='SIN REFERENCIA', _c14='71', _c15=None, _c
16=None, _c17=None, _c18='0', _c19=None, _c20='FRACCIONAMIENTO', _c21='FRACCIONAMIENTO LOS MOLINOS', _c22=None, _c23=Non
e, _c24=None, _c25='45200', _c26='14', _c27='Jalisco', _c28='120', _c29='Zapopan', _c30='430', _c31='Campestre las Paloma
s [Fraccionamiento]                                                                                    ', _c32='6940', _c33='13', _c34
='3317429415', _c35='ELDOTE@HOTMAIL.COM', _c36=None, _c37='Fijo', _c38='20.8206615', _c39='-103.4467258', _c40='2019-1
1'),
 Row(_c0='8838371', _c1='14039112511000014000000000U4', _c2='ACUACULTURA DEL PACIFICO SPR DE RL', _c3='ACUACULTURA DEL PA
CIFICO SPR DE RL', _c4='112511', _c5='11 a 30 personas', _c6='CALLE', _c7='PABLO VALDEZ', _c8='CALLE', _c9='PEDRO A. GALV
AN', _c10='CALLE', _c11='PEDRO TAMEZ', _c12='CALLE', _c13='ESTEBAN ALATORRE', _c14='606', _c15=None, _c16=None, _c17=Non
e, _c18=None, _c19=None, _c20='COLONIA', _c21='LA PERLA', _c22=None, _c23=None, _c24=None, _c25='44360', _c26='14', _c27
='Jalisco', _c28='39', _c29='Guadalajara', _c30='1', _c31='Guadalajara
', _c32='1170', _c33='3', _c34=None, _c35='CONSTRULOGA@PRODIGY.NET.MX', _c36=None, _c37='Fijo', _c38='20.68168604', _c39
='-103.3299623', _c40='2019-11'),
```

```
df=spark.read.load(path,format='com.databricks.spark.csv',header='true',inferSchema='true',encoding="ISO-8859-1")
display(df)
```

Table

	id	clee	nom_estab
1	8624390	14120112512000022000000000U3	ACUACULTORES LOS CASTRO S.C. DE R.L. DE C.V.
2	8838371	14039112511000014000000000U4	ACUACULTURA DEL PACIFICO SPR DE RL
3	9233864	14095112512000022000000000U7	ACUIC0LA LA CABAÑA
4	8341990	14095114119000082000000000U3	ACUICOLA DE VILLA CORONA S.P.R. DE R.L.
5	8908807	14079112512000011000000000U4	ACUÍCOLA EL DURAZNO
6	8901362	14093112512000012000000000U3	ACUICOLA LA PERSEVERANCIA SC DE RL DE CV
7	8274470	14085112512000051000000000U7	ACUÍCOLA LOS RUCIOS

1,000 rows | Truncated data

```
df_pandas = df.toPandas()
type(df_pandas)
```

```
Out[41]: pandas.core.frame.DataFrame
```

```
df_pandas.columns
```

```
Out[42]: Index(['id', 'clee', 'nom_estab', 'raz_social', 'codigo_act', 'per_ocu',
       'tipo_vial', 'nom_vial', 'tipo_v_e_1', 'nom_v_e_1', 'tipo_v_e_2',
       'nom_v_e_2', 'tipo_v_e_3', 'nom_v_e_3', 'numero_ext', 'letra_ext',
       'edificio', 'edificio_e', 'numero_int', 'letra_int', 'tipo_asent',
       'nomb_asent', 'tipoCenCom', 'nom_CenCom', 'num_local', 'cod_postal',
       'cve_ent', 'entidad', 'cve_mun', 'municipio', 'cve_loc', 'localidad',
       'ageb', 'manzana', 'telefono', 'correoelec', 'www', 'tipoUniEco',
       'latitud', 'longitud', 'fecha_alta'],
      dtype='object')
```

```
path = "/FileStore/tables/clee_denue.csv"
data = spark.read.csv(path, encoding="ISO-8859-1")
data.take(20)
df=spark.read.load(path,format='com.databricks.spark.csv',header='true',inferSchema='true',encoding="ISO-8859-1")
display(df)
clee_df = df.toPandas()
type(df_pandas)
clee_df.head()
```

Table

	clee	Clase
1	111110	Cultivo de soya
2	111121	Cultivo de cÃ¡rtamo
3	111122	Cultivo de girasol
4	111129	Cultivo anual de otras semillas oleaginosas
5	111131	Cultivo de frijol grano
6	111132	Cultivo de garbanzo grano
7	111139	Cultivo de otras leguminosas

1,000 rows | Truncated data

	clee	Clase
0	111110	Cultivo de soya
1	111121	Cultivo de cÃ¡rtamo
2	111122	Cultivo de girasol
3	111129	Cultivo anual de otras semillas oleaginosas
4	111131	Cultivo de frijol grano

```
df_pandas = pd.merge(df_pandas, clee_df, how= 'inner', left_on= 'codigo_act', right_on = 'clee')
```

```
df_pandas['per_ocu'].unique()
```

```
Out[45]: array(['0 a 5 personas', '6 a 10 personas', '11 a 30 personas',
       '31 a 50 personas', '251 y más personas', '101 a 250 personas',
       '51 a 100 personas'], dtype=object)
```

```
encod_per_ocu= {'0 a 5 personas':1, '6 a 10 personas':2, '11 a 30 personas':3,
'31 a 50 personas':4, '51 a 100 personas':5, '101 a 250 personas':6, '251 y más personas':7}
encod_per_ocu
```

```
Out[46]: {'0 a 5 personas': 1,
 '6 a 10 personas': 2,
 '11 a 30 personas': 3,
 '31 a 50 personas': 4,
 '51 a 100 personas': 5,
```

```
 '101 a 250 personas': 6,
 '251 y más personas': 7}

encode_clee = df_pandas.groupby('clee_y')[['id']].count().sort_values(by= 'id', ascending= False)
encode_clee['rep_scian'] = [x for x in range(len(encode_clee))]
encode_clee.reset_index(inplace= True)
encode_clee.head()
```

	clee_y	id	rep_scian
0	461110	37505	0
1	812110	18478	1
2	463211	13324	2
3	722514	13315	3
4	465311	7583	4

```
df_pandas['per_ocu_enc'] = df_pandas['per_ocu'].map(encod_per_ocu)
df_pandas = df_pandas.merge(encode_clee, how= 'inner', left_on= 'clee_y', right_on= 'clee_y')
```

```
df_pandas.columns
```

```
Out[49]: Index(['id_x', 'clee_x', 'nom_estab', 'raz_social', 'codigo_act', 'per_ocu',
       'tipo_vial', 'nom_vial', 'tipo_v_e_1', 'nom_v_e_1', 'tipo_v_e_2',
       'nom_v_e_2', 'tipo_v_e_3', 'nom_v_e_3', 'numero_ext', 'letra_ext',
       'edificio', 'edificio_e', 'numero_int', 'letra_int', 'tipo_asent',
       'nomb_asent', 'tipoCenCom', 'nom_CenCom', 'num_local', 'cod_postal',
       'cve_ent', 'entidad', 'cve_mun', 'municipio', 'cve_loc', 'localidad',
       'ageb', 'manzana', 'telefono', 'correoelec', 'www', 'tipoUniEco',
       'latitud', 'longitud', 'fecha_alta', 'clee_y', 'Clase', 'per_ocu_enc',
       'id_y', 'rep_scian'],
      dtype='object')
```

```
columns_to_use = ['cod_postal','cve_ent', 'cve_mun' , 'cve_loc','rep_scian', 'per_ocu_enc']
```

```
df_pandas.dropna(subset= columns_to_use, axis=0, inplace = True)
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

Show result

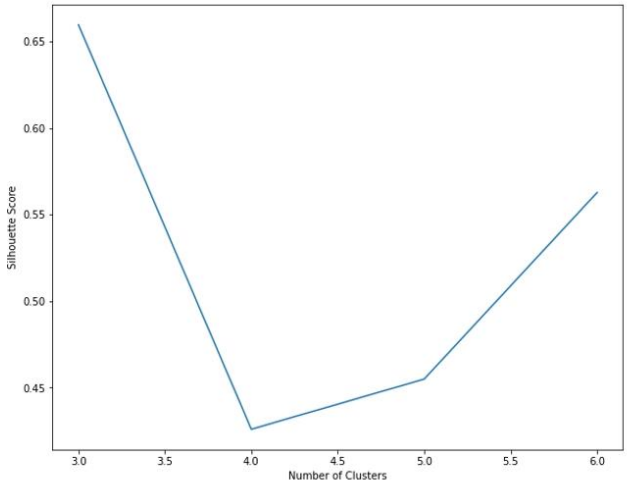
```
df_pandas[columns_to_use].shape
```

```
Out[53]: (375547, 6)
```

```
ss = StandardScaler()
standarized_df = pd.DataFrame(data = ss.fit_transform(df_pandas[columns_to_use]),
                              columns = columns_to_use)
standarized_df.head()
```

	cod_postal	cve_ent	cve_mun	cve_loc	rep_scian	per_ocu_enc
0	-0.569702	0.0	1.438734	3.854508	2.254383	-0.346970
1	0.416583	0.0	0.739093	-0.184610	2.254383	-0.346970
2	-0.234618	0.0	1.270820	-0.184610	2.254383	0.967169
3	1.927623	0.0	0.291323	-0.052797	2.254383	-0.346970
4	-1.164003	0.0	0.683122	-0.184610	2.254383	0.967169

```
jalisco_data = spark.createDataFrame(standarized_df)


# create a vector assembler and transform raw features into a single set of features
from pyspark.ml.feature import VectorAssembler
assemble=VectorAssembler(inputCols=['cod_postal','cve_ent', 'cve_mun' , 'cve_loc','rep_scian', 'per_ocu_enc'],outputCol =
'jals_features')

assembled_data=assemble.transform(jalisco_data)



from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator

silhouette_scores=[]
evaluator = ClusteringEvaluator(featuresCol='jals_features', \
metricName='silhouette', distanceMeasure='squaredEuclidean')

for K in range(3,7):
    print(K)

    KMeans_=KMeans(featuresCol='jals_features', k=K)

    KMeans_fit=KMeans_.fit(assembled_data)

    KMeans_transform=KMeans_fit.transform(assembled_data)

    evaluation_score=evaluator.evaluate(KMeans_transform)

    silhouette_scores.append(evaluation_score)


3
4
5
6




# plot
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1,1, figsize =(10,8))
ax.plot(range(3,7),silhouette_scores)
ax.set_xlabel('Number of Clusters')
ax.set_ylabel('Silhouette Score')

Out[58]: Text(0, 0.5, 'Silhouette Score')
```




```
KMeans_=KMeans(featuresCol='jals_features', k=3)
KMeans_Model=KMeans_.fit(assembled_data)
KMeans_Assignments=KMeans_Model.transform(assembled_data)
```

```
from pyspark.ml.feature import PCA as PCAml
pca = PCAml(k=2, inputCol="jals_features", outputCol="pca")
pca_model = pca.fit(assembled_data)
pca_transformed = pca_model.transform(assembled_data)
```

```
cluster_assignment = np.array(KMeans_Assignments.rdd.map(lambda row: row.prediction).collect()).reshape(-1,1)
```

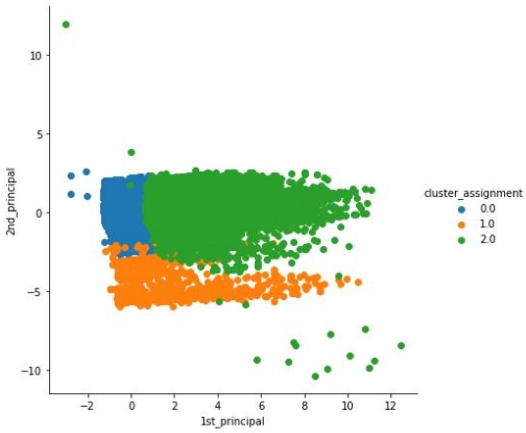
```
len(cluster_assignment)
```

```
Out[62]: 375547
```

```
x_pca = np.array(pca_transformed.rdd.map(lambda row: row.pca).collect())
```

```
import seaborn as sns
import matplotlib.pyplot as plt

pca_data = np.hstack((x_pca,cluster_assignment))

pca_df = pd.DataFrame(data=pca_data, columns=("1st_principal", "2nd_principal","cluster_assignment"))
sns.FacetGrid(pca_df,hue="cluster_assignment", height=6).map(plt.scatter, '1st_principal', '2nd_principal' ).add_legend()
```

```
Out[64]: <seaborn.axisgrid.FacetGrid at 0x7fd63e67b2e0>
```




```
resultados = df_pandas[columns_to_use]
resultados['cluster'] = cluster_assignment
resultados
```

```
<command-4408032627284666>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  resultados['cluster'] = cluster_assignment
```

	cod_postal	cve_ent	cve_mun	cve_loc	rep_scian	per_ocu_enc	cluster
0	45200.0	14	120	430	302	1	1
1	46760.0	14	95	1	302	1	2
2	45730.0	14	114	1	302	2	2
3	49150.0	14	79	15	302	1	0
4	44260.0	14	93	1	302	2	2
...	...	...	...	...	...	...	...
378434	44140.0	14	39	1	597	2	2
378435	47180.0	14	8	1	597	2	2
378436	45500.0	14	98	1	597	3	2
378437	46500.0	14	36	1	597	2	2
378438	46200.0	14	25	1	597	1	2

375547 rows × 7 columns

```
encod_per_ocu
```

```
Out[66]: {'0 a 5 personas': 1,
 '6 a 10 personas': 2,
 '11 a 30 personas': 3,
 '31 a 50 personas': 4,
 '51 a 100 personas': 5,
 '101 a 250 personas': 6,
 '251 y más personas': 7}
```

```
    encode_clee.iloc[242:248,:]
```

	clee_y	id	rep_scian
242	433110	194	242
243	434319	192	243
244	311813	192	244
245	327420	191	245
246	433220	191	246
247	321920	191	247

```
for i in resultados['cluster'].unique():
    print(f'Información del cluster {i}')
    display(resultados.loc[resultados.cluster == i].describe())
```

```
Información del cluster 1
```

Table

	cod_postal	cve_ent	cve_mun	cve_loc	rep_scian	per_ocu_enc	cluster
1	6508	6508	6508	6508	6508	6508	6508
2	45674.5503995083	14	96.9422249539029	772.1166256914566	50.92947141979103	1.2937922556853103	1
3	366.9786598514659	0	9.196823913236653	139.63258362972138	95.16908397936427	0.8949999598309407	0
4	44100	14	8	417	0	1	1
5	45647	14	97	822	3	1	1
6	45654	14	97	822	17	1	1
7	45655	14	97	843	51	1	1

8 rows

```
Información del cluster 2
```

Table

	cod_postal	cve_ent	cve_mun	cve_loc	rep_scian	per_ocu_enc	cluster
1	44512	44512	44512	44512	44512	44512	44512
2	45719.12832494608	14	67.94459920920201	7.847501797268152	246.45805625449316	2.72847771387491	2
3	1666.909549274674	0	35.750439094906845	37.11215507281151	174.63528990453005	1.3777648488564382	0
4	0	14	1	1	0	1	2
5	44657.75	14	39	1	108	2	2
6	45140	14	58	1	224	3	2
7	46470	14	100	1	346.25	3	2

8 rows

```
Información del cluster 0
```

Table

	cod_postal	cve_ent	cve_mun	cve_loc	rep_scian	per_ocu_enc	cluster
1	324527	324527	324527	324527	324527	324527	324527
2	46162.037922884694	14	68.11026509350532	7.287227256900042	44.64937277946057	1.0625679835576085	0
3	1576.0364349248343	0	35.835098970389744	30.80649055385062	55.789217758427604	0.24271845989020305	0
4	28200	14	1	1	0	1	0
5	44980	14	39	1	4	1	0
6	45588	14	67	1	23	1	0
7	47473	14	98	1	64	1	0

8 rows

.

Descripción de clusters

- Cluster 0 Son aquellas empresas con un número chico de empleados que se dedican principalmente a las actividades de:
  - Comercio al por menor de cerveza
  - Comercio al por menor de vinos y licores
  - Comercio al por menor de otros alimentos

Lo que puede ser traducido a tiendas de conveniencia chicas.

- Cluster 1 Desde negocios chicos hasta los más grandes pero con actividades económicas más diversas
- Cluster 2 En este cluster tenemos a las empresas que tienen un número mayor de empleados +50 personas en actividades como:
  - Comercio al por mayor de productos farmacéuticos
  - Fabricación de yeso y productos de yeso
  - Fabricación de productos para embalaje y envases de madera

CART

Se usará un árbol de clasificación para poder tener los valores del cluster como la target a predecir

```
from pyspark import SparkContext
from pyspark.sql import SQLContext
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.feature import VectorAssembler
from sklearn.metrics import confusion_matrix
import pandas as pd
```

Motivo de uso de estas variables

Nosotros creamos variables que son significativas como el tamaño de las empresas y también la representatividad de cada industria dentro de zonas geográficas cercanas porque lo que ya tenemos clara esa parte con las columnas creadas y solo tocaría ahora agregar las variables que logran identificar geográficamente el lugar de donde pertenece cada una

```
features = ['cod_postal', 'cve_ent', 'cve_mun', 'cve_loc', 'rep_scian','per_ocu_enc']
```

```
data = spark.createDataFrame(resultados)
va = VectorAssembler(inputCols = features, outputCol='features')
va_df = va.transform(data)
va_df = va_df.select(['features', 'cluster'])
va_df.show(3)
```

```
+--------------------+-------+
|            features|cluster|
+--------------------+-------+
|[45200.0,14.0,120...|      1|
|[46760.0,14.0,95....|      2|
|[45730.0,14.0,114...|      2|
+--------------------+-------+
only showing top 3 rows
```

```
import random

random.seed(42)
(train, test) = va_df.randomSplit([0.7, 0.3])
```

```
dtc = DecisionTreeClassifier(featuresCol="features", labelCol="cluster")
dtc = dtc.fit(train)
```

```
pred = dtc.transform(test)
pred.show(3)
```

```
+--------------------+-------+--------------------+--------------------+----------+
|            features|cluster|       rawPrediction|         probability|prediction|
+--------------------+-------+--------------------+--------------------+----------+
|[0.0,14.0,120.0,1...|      2|   [0.0,0.0,5703.0]|       [0.0,0.0,1.0]|       2.0|
|[4564.0,14.0,120....|      2|   [0.0,0.0,5703.0]|       [0.0,0.0,1.0]|       2.0|
|[4725.0,14.0,116....|      2|[13858.0,0.0,358.0]|[0.97481710748452...|       0.0|
+--------------------+-------+--------------------+--------------------+----------+
only showing top 3 rows
```

```
evaluator=MulticlassClassificationEvaluator(predictionCol="prediction", labelCol= 'cluster')
acc = evaluator.evaluate(pred)

print("Prediction Accuracy: ", acc)
```

```
Prediction Accuracy:  0.9909778678775001
```

```
y_pred=pred.select("prediction").collect()
y_orig=pred.select("cluster").collect()

cm = confusion_matrix(y_orig, y_pred)
print("Confusion Matrix:")
print(cm)
```

```
Confusion Matrix:
[[97172     3   377]
 [   88  1793    63]
 [  427    58 12961]]
```

Resultados y conclusiones

Como lo vimos durante el desarrollo del proyecto podemos encontrar dadas las zonas geográficas de Jalisco, no solamente una tendencia de desarrollo por actividad económica, el crecimiento de estas o su número de empleados.

Logramos hacer una unión de todas estas variables para poder así identificar a las regiones a lo largo del tiempo y eso es muy valioso siendo que aplicando lo visto durante el curso: recabar datos de diferentes fuentes en el ecosistema de Hadoop para usar el map reduce con pig, el almacenamiento con HDFS, el ETL con Hive, Databricks y MongoDB para cambiar a formato a JSON, machine learning con regresión lineal y arboles de clasificación con su respectiva métrica de desempeño para evaluar nuestro modelo propuesto.

Se logró una clusterización de las características de las empresas, con una fuente general (DENUE) y encontrando más bases de datos que nos permitieran identificar geográficamente y por el giro del negocio a las mismas, se hizo una visualización correcta donde la separación con 3 clusters (elegidos dada la información que veíamos en la silueta-score) y es claro cómo es que la información si muestra esa tendencia de agrupación porque es visualmente separable la información. Junto con un perfilamiento ya que al reducir las dimensiones y clusterizar tenemos que ver;

¿Qué información es la que estamos obteniendo de estas agrupaciones? No es solamente tener un grupo, sino que podamos interpretar que tipo de grupo es. Esto dado con una revisión estadística de las características de cada uno. Y usando la desviación estándar para poder encontrar la proyección inversa de los métodos ocupados, pudimos saber a qué grupo pertenecían tanto geográficamente, su giro y los trabajadores encontrando resultados tan significativos como que dentro de las zonas céntricas los negocios con pocos empleados, de giro de autoservicios junto con la venta de bebidas alcohólicas es tan relevante y que rápidamente se puede traquear este comportamiento a los Oxxos, 7-Eleven y algunas tiendas de autoservicio tienen una representación en todos estos puntos evitente.

Después con ayuda de Arboles de Decisión logramos encontrar cuales son las características que pueden llevar a las empresas a ser parte de alguna de estas categorías y con ello poder planear desde antes el mercado en el que se van a encontrar compitiendo. Un ejemplo es que si hay una empresa que después de pasar sus características por la predicción del árbol de decisión nos da una predicción de cluster 3 sabemos que las empresas con las que va a estar compitiendo dentro de las zona en la que inicie sus actividades serán muy grandes en número de empleados por lo que tendrá que apuntar a fidelización de sus clientes más que a la cantidad debido a que tendría que competir contra un capital mucho más sólido desde el inicio evitando el *churn* donde los costos por atraer nuevos clientes disminuye.

La métrica de desempeño que ocupamos fue el *accuracy* con un resultado de 0.9 lo que nos indica que posiblemente nuestro modelo este sobre entrenado ya que está aprendiendo de la clase positiva por ello se recomienda realizar balancear las clases especialmente los FN o FP.

Como comentario final, la actividad nos enseñó desde una perspectiva que nosotros podemos encontrar los datos de prácticamente todo lo que busquemos investigar y con apoyo de las herramientas vistas en el curso las conclusiones que demos irán más allá de opiniones, nos basamos en datos sólidos y públicos para que nuestro punto de vista siempre sea data-driven.

Bibliografía

- Secciones Introduction to Spark, Spark Framework y Spark Architecture, del capítulo 7, del libro: Achari, S. (2015). Hadoop Essentials: Delve Into the Key Concepts of Hadoop and Get a Thorough Understanding of the Hadoop Ecosystem. Packt Publishing. https://0-search-ebscohost-com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=e000xww&AN=986713&lang=es&site=ehost-live

- Secciones Introduction, Using linear regression y Understanding cost function, del capítulo 7, del libro: Yadav R. (2015) Spark Cookbook: Over 60 recipes on Spark, covering Spark Core, Spark SQL, Spark Streaming, MLlib, and GraphX libraries. Packt Publishing. https://0-search-ebscohost-com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=nlebk&AN=1044814&lang=es&site=eds-live&scope=site

- Kane, F. (2017). Hands-On Data Science and Python Machine Learning. Packt Publishing.

- Junjie Wu. (2012). Advances in K-means clustering : a data mining thinking. Springer.

- Portal Consejería Jurídica y de Servicios Legales del DF. (n.d.). Data.consejeria.cdmx.gob.mx. Retrieved March 19, 2023, from https://data.consejeria.cdmx.gob.mx/index.php/component/glossary/Glosario-Consejer%C3%ADa-1/E/ESTUDIO-SOCIOECON%C3%93MICO-67/#:~:text=II%2DEs%20un%20documento%20que

- Geografía (INEGI), I. N. de E. y. (n.d.). Quiénes somos. www.inegi.org.mx. https://www.inegi.org.mx/inegi/quienes_somos.html#:~:text=Somos%20un%20organismo%20p%C3%BAblico%20aut%C3%B3nomo

- Regiones de Jalisco | Gobierno del Estado de Jalisco. (n.d.). Www.jalisco.gob.mx. Retrieved March 19, 2023, from https://www.jalisco.gob.mx/jalisco/regiones