Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Campus LATAM

Nombre del trabajo:

5.3 Fase 3b. Ingeniería de datos: enriquecer, agregar, generar features

Curso:

Análisis de Grandes Volúmenes de Datos

Alumnos:

Juan Carlos Alvarado Carricarte, A01793486

Bryan Rodolfo Alvarado Cruz, A01793670

Eduardo Gabriel Arévalo Aguilar, A01793897

Profesor:

Alberto De Obeso Orendain

Tutor:

Luis Angel Lozano Medina

Fecha de entrega:

12 de febrero

Resumen

En este trabajo se generaron nuevas características para enriquecer los datos demográficos de los municipios de la zona de Jalisco. Se diseñaron y describieron nuevas características para seis fuentes de datos diferentes: DENUE, educación, hogares, natalidad, mortalidad y población. Algunas de las nuevas características incluyen el tiempo de creación de las empresas y el tamaño de las compañías en DENUE, el promedio de personas por hogar en los municipios de Jalisco en el último año en la tabla hogares, la tasa de alfabetización por municipios y la cantidad de empresas en la zona con la tabla de educación educación, y la tasa de crecimiento poblacional sin inmigrantes usando las tablas de natalidad, mortalidad y población. Las nuevas características se implementaron utilizando Apache Spark.

El enriquecimiento de datos es fundamental para mejorar la calidad de los modelos y análisis pertinente que se puede realizar. Las nuevas características pueden dar una nueva perspectiva de los municipios de la zona de Jalisco, identificando posibles relaciones entre diferentes variables y ayudando a obtener mejores respuestas.

Palabras clave: ingeniería de características, limpieza de datos, enriquecimiento de datos, Apache Spark, Databricks, datos demográficos, fuentes de datos, integración de datos, México, Jalisco, DENUE.

Ingeniería de Features

La ingeniería de características es una de las etapas clave en la construcción de modelos de machine learning y minería de datos. Esta etapa en general es la que más demanda tiempo cuando se realiza la limpieza de datos y la exploración para su entendimiento, también se trata de generar nuevas características o atributos que permitan mejorar la calidad de los datos para mejorar los resultados de los modelos. En este trabajo se presenta la ingeniería de características realizada sobre datos demográficos de los municipios de la zona de Jalisco, con el simple objetivo de enriquecer los datos disponibles y mejorar la capacidad de futuros modelos. Haciendo uso de Apache Spark, se diseñaron nuevas características a partir de las fuentes de datos con las que ya se vienen trabajando.

En el presente trabajo, se describe el proceso de ingeniería de características realizado para enriquecer los datos demográficos de los municipios de la zona de Jalisco. Para cada una de las fuentes de datos (DENUE, educación, hogares, natalidad, mortalidad y población), se identificaron características relevantes que permitieran obtener información adicional para la creación de nuevas características a partir de las mismas, y que puedan mejorar los futuros modelos de análisis. Algunas características se pueden analizar en tablas independientes, mientras que otras, se pueden agregar en algunas de las tablas ya existentes.

Tabla DENUE

Para la tabla DENUE se agregaron dos nuevas características: el *tiempo de creación de las empresas* y el *tamaño de la compañía por número de empleados*.

Tabla DENUE:

1. Tiempo en años que llevan las compañías:

Cmd 29

```
%sql
SELECT id AS id_denue, (YEAR(GETDATE()) - CAST(SUBSTR(fecha_alta,1,4) AS INT)) AS anos_de_la_compania
FROM denue
LIMIT 10
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [id_denue: integer, anos_de_la_compania: integer]

Table +

	id_denue	anos_de_la_compania
1	1789827	9
2	9337856	2
3	9203369	4
4	9316641	3
5	1957928	13
6	6845953	5
7	8433744	4

10 rows | 1.08 seconds runtime

2. Tamaño de la compañía por número de empleados:

Cmd 31

```
%sql
SELECT id AS id_denue, per_ocu,
  CASE
    WHEN per_ocu = '0 a 5 personas' OR per_ocu = '6 a 10 personas' OR per_ocu = '11 a 30 personas' OR per_ocu = '31 a 50 personas' THEN 'Pequeña'
    WHEN per_ocu = '51 a 100 personas' OR per_ocu = '101 a 250 personas' THEN 'Mediana'
    WHEN per_ocu = '251 y más personas' THEN 'Grande'
    ELSE 'No Clasificada'
  END AS tamano_empresa
FROM denue
LIMIT 10
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [id_denue: integer, per_ocu: string ... 1 more field]

Table +

	id_denue	per_ocu	tamano_empresa
1	1789827	31 a 50 personas	Pequeña
2	9337856	0 a 5 personas	Pequeña
3	9203369	0 a 5 personas	Pequeña
4	9316641	6 a 10 personas	Pequeña
5	1957928	101 a 250 personas	Mediana
6	6845953	251 y más personas	Grande
7	8433744	0 a 5 personas	Pequeña

Tabla Hogares

En hogares, se obtuvo el *promedio de personas por hogar en Jalisco en cada municipio en el último año*.

Tabla Hogares:

1. Promedio de personas por hogar en Jalisco en cada municipio en el último año:

Cmd 27

Untitled

```
%sql
SELECT cve_municipio,
       sum(CASE WHEN indicador = 'Población en hogares' THEN valor ELSE 0 END) /
       sum(CASE WHEN indicador = 'Hogares' THEN valor ELSE 0 END) AS promedio_poblacion_por_hogar,
       "Promedio de población por hogar" AS unidad_promedio_poblacion
FROM indicadoresHogares
WHERE ano = (SELECT MAX(ano) FROM indicadoresHogares) AND cve_entidad = "14"
GROUP BY cve_municipio
SORT BY promedio_poblacion_por_hogar;
```

(4) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [cve_municipio: integer, promedio_poblacion_por_hogar: double ... 1 more field]

Table +

	cve_municipio	promedio_poblacion_por_hogar	unidad_promedio_poblacion
1	34	2.9650986342943852	Promedio de población por hogar
2	54	2.9905027932960895	Promedio de población por hogar
3	104	3.130925507900677	Promedio de población por hogar
4	117	3.1432664756446993	Promedio de población por hogar
5	56	3.167539267015707	Promedio de población por hogar
6	12	3.185354691075515	Promedio de población por hogar
7	65	3.1907303370786515	Promedio de población por hogar

Tabla Educación

En educación se cruzaron las tablas DENUE y educación para obtener la *cantidad de empresas en la zona* y averiguar si existe alguna correlación con la tasa de alfabetización en personas mayores a 15 años.

```
%sql
SELECT D.cve_mun, D.municipio, count(D.id) AS cantidad_empresas,
      IE.promedio_alfabetas_mayores_15
FROM denue D
JOIN (
   SELECT IE.cve_municipio, IE.desc_municipio,
          SUM(CASE WHEN IE.indicador = 'Porcentaje de personas de 15 años y más alfabetas' THEN valor ELSE 0 END) AS promedio_alfabetas_mayores_15
   FROM indicadoresEducacion IE
   WHERE IE.ano = (SELECT MAX(ano) FROM indicadoresEducacion) AND IE.cve_entidad = "14"
   GROUP BY IE.cve_municipio, IE.desc_municipio
) IE
ON D.cve_mun = IE.cve_municipio
GROUP BY D.cve_mun, D.municipio, IE.promedio_alfabetas_mayores_15
SORT BY cantidad_empresas DESC
LIMIT 50
```

(6) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [cve_mun: integer, municipio: string ... 2 more fields]

Table +

	cve_mun	municipio	cantidad_empresas	promedio_alfabetas_mayores_15
1	39	Guadalajara	99738	98.271947982632
2	120	Zapopan	52309	98.1738527443924
3	98	San Pedro Tlaquepaque	24625	97.5796467155587
4	101	Tonalá	19420	97.6907482179232
5	67	Puerto Vallarta	16610	97.8478863494636
6	97	Tlajomulco de Zúñiga	15709	98.2430971837565
7	93	Tepatitlán de Morelos	7714	95.6773987886516

Tablas Natalidad, Mortalidad y Población

Finalmente, para las tablas de natalidad, mortalidad y población, se calculó la *tasa de crecimiento poblacional* sin inmigrantes o tasa de crecimiento natural por cada mil habitantes en cada municipio para el año 2020.

- Tasa de crecimiento natural = [(Nacimientos - Defunciones) / Población inicial] x 1000

Reemplazando valores, Tasa de crecimiento natural: = [(Nacimientos(2020) - Defunciones(2020)) / Población inicial(2020)] x 1000

PD: El ejercicio se hace con la población del año 2020, ya que el censo no registra población del año anterior para sacar la media

```
%sql
SELECT INat.cve_municipio, INat.desc_municipio, INat.nacimientos, IM.muertes, IP.poblacion_total,
       ((INat.nacimientos - IM.muertes) / IP.poblacion_total)*1000 AS tasa_crecimiento_natural
FROM (
  SELECT cve_municipio, desc_municipio,
         CASE WHEN indicador = 'Nacimientos registrados' THEN valor ELSE 0 END AS nacimientos
  FROM indicadoresNatalidad
  WHERE cve_entidad = '14' AND ano = '2020' AND indicador = 'Nacimientos registrados'
) INat
JOIN (
  SELECT cve_municipio, desc_municipio,
         CASE WHEN indicador = 'Defunciones registradas' THEN valor ELSE 0 END AS muertes
  FROM indicadoresMortalidad
  WHERE cve_entidad = '14' AND ano = '2020' AND indicador = 'Defunciones registradas'
) IM ON INat.cve_municipio = IM.cve_municipio
JOIN (
  SELECT cve_municipio, desc_municipio,
         CASE WHEN indicador = 'Población total' THEN valor ELSE 0 END AS poblacion_total
  FROM indicadoresPoblacion
  WHERE cve_entidad = '14' AND ano = '2020' AND indicador = 'Población total'
) IP ON INat.cve_municipio = IP.cve_municipio
SORT BY tasa_crecimiento_natural DESC
LIMIT 10;
```

Table +

	cve_municipio	desc_municipio	nacimientos	muertes	poblacion_total	tasa_crecimiento_natural
1	19	Bolaños	291	43	7043	35.21226749964504
2	61	Mezquitic	770	115	22083	29.660825069057648
3	125	San Ignacio Cerro Gordo	522	139	18341	20.882176544354177
4	71	San Cristóbal de la Barranca	63	16	2924	16.073871409028726
5	106	Tuxcacuesco	122	35	5482	15.870120394016782
6	111	Valle de Guadalupe	176	71	6627	15.844273426889995
7	86	Tapalpa	471	138	21245	15.674276300305955

La ingeniería de características permitió obtener información adicional y relevante para el análisis de los datos, lo que contribuirá a la creación de modelos más precisos y eficientes.

.

Implementación con Apache Spark

Para la generación de las nuevas características, se hizo uso de la herramienta Apache Spark en la plataforma cloud de *Databricks*. Se cargaron todos los archivos resultantes del proceso anterior de ETL donde se limpió, transformó e implementaron algunas pequeñas integraciones con los datos. Gracias a la versatilidad de Databricks, se cargaron las tablas y se convirtieron en dataframes usando Python y Pandas, herramientas con las cuales se hicieron un análisis y exploración de todos los datos. Posteriormente, para las consultas y generación de nuevas tablas y características se usó SQL, ya que Databricks permite el uso de varios lenguajes en sus libros de ejecución de comandos.

En conclusión, el proceso de ingeniería de características nos permite realizar varias secciones como la limpieza y exploración de datos para su transformación, analizar las características de nuestras fuentes de datos, agregar nuevas características que nos permitan enriquecer los datos y puedan dar una mejor calidad a nuestros modelos de análisis y aprendizaje automático. En este estudio, se aplicó la ingeniería de características en seis fuentes de datos demográficos de los municipios de la zona de Jalisco: DENUE, educación, hogares, natalidad, mortalidad y población. Se diseñaron y describieron nuevas características para cada fuente de datos y se implementaron utilizando Apache Spark. Los resultados de la aplicación de estas características en los datos demográficos mostraron que se puede mejorar la calidad de los datos existentes para un posterior análisis, que se pueden sacar algunas pequeñas conclusiones, pero requiere de análisis más profundos con procesos de machine

learning para poder sacar detalles a fondo. En general, la ingeniería de características es un proceso que puede mejorar significativamente el rendimiento de los modelos de aprendizaje automático.

Referencias

Apache Spark™ - Unified Engine for large-scale data analytics. (s. f.). https://spark.apache.org

Databricks. (2022, 26 abril). *Contact Us*. https://www.databricks.com/

Geografía, E. D. N. I. Y. (s. f.-a). Directorio Estadístico Nacional de Unidades Económicas. DENUE. *https://www.inegi.org.mx/app/mapa/denue/default.aspx*

Regiones de Jalisco | Gobierno del Estado de Jalisco. (s. f.). https://www.jalisco.gob.mx/es/jalisco/regiones

Patel, H. (2022, 5 enero). *What is Feature Engineering — Importance, Tools and Techniques for Machine Learning*. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10