Figure 1: Logo ITESO

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

PRESENTAN

Programas educativos y estudiantes

Lic. en Ingeniería Financiera. Frida María Hernández López

Lic. en Ingeniería Financiera. José Leonardo Aceves González

Lic. en Ingeniería Financiera. Jesús Iván Lafarga Lizárraga

Profesor PAP: Diana Paola Montoya Escobar

Tlaquepaque, Jalisco, Julio del 2022

Contenido

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración y/o coparticipación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.

El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).

El PAP es un proceso acotado en el tiempo en que los estudiantes, los beneficiarios externos y los profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, cocrear o transformar en distintos campos sociales. El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las disimetrías sociales. En otras palabras, se trata del reto de “saber y aprender a transformar”.

Objetivo

Vincular de forma integral, transferir y aplicar los conocimientos y habilidades que se han adquirido y desarrollado a lo largo de la carrera para proponer soluciones en materia de gestión de riesgos de tipo financiero

Actividades para lograr el resultado

Desarrollar un análisis exploratorio de los datos históricos (datos públicos), ayudados de herramientas estadísticas. Para lograrlo, se debe utilizar herramientas como R, Python, tableau, entre otras de business inteligence. Buscar tendencias y proponer modelos matemáticos de pronósticos del precio de las importaciones.

Resultado

Análisis de datos. Resultados de análisis de correlación para realizar modelo que nos ayude a predecir ciertas variables hacia el futuro para poder tomar mejores decisiones.

Metodología (Issues)

1. Exploración de datos
2. Limpieza de datos - texto
3. Ingeniería de características
4. Creación de modelos
5. Reporte PAP
6. Presentación.

Exploración de datos

Como es usual en ciencia de datos, uno de los primeros pasos a seguir es el de data knowledge and visualization. Para este apartado, decidimos trabajar en Python ya que, es el lenguaje en el que tenemos mayor expertiz y nos permite hacer una indagatoria más minuciosa.

Comenzamos analizando los datos de manera habitual, explorando a mano el DataFrame, columna por columna y a partir de ahí generar ciertas graficas como: histogramas sobre los precios, agrupadores por país de procedencia de las importaciones, matriz de correlación para ver que variables tienen una mayor correlación y demás métricas relevantes.

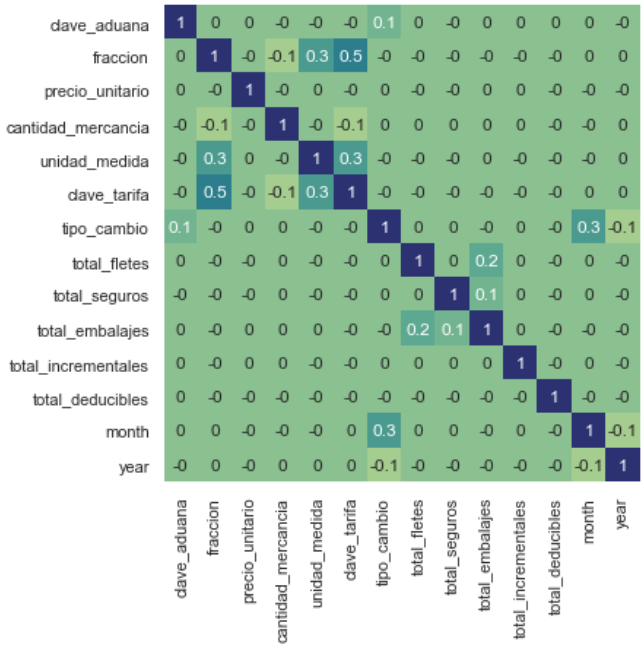

Figura. 2

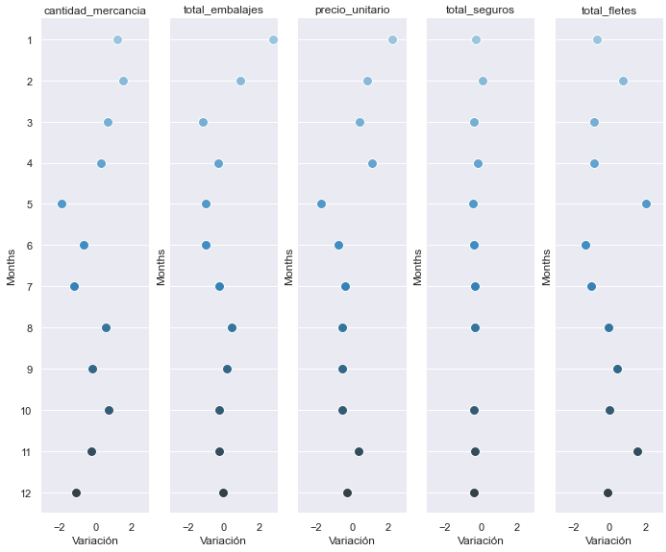

Figura. 3

En la figura 2, podemos observar que la mayoría de las variables tienen una correlación irrelevante, algunas incluso llegan a valores negativos. Sin embargo, hay un par como, ‘clave_tarifa’, ‘unidad_medida’ que guardan una fuerte relación con la columna de ‘fracción’, ‘month’ con ‘tipo_cambio’ y otras variables que guardan una correlación considerable.

En la figura 3, vemos como las variables relacionadas al envio de los productos como, la cantidad de mercancía, el gasto en embalajes, el gasto en primas de seguro y el flete (no especifica el medio de transporte) tienen ciertas particularidades, sin guardar una fuerte correlación entre sí.

Notamos que, independientemente del mes, el gasto en primas de seguro es constante, salvo el mes de septiembre donde, hay un aumento considerable en el gasto que, a primera vista, parece no tener una justificación ya que ni la cantidad de mercancía ni el gasto en embalaje aumenta. Concluimos que este aumento fue consecuencia del aumento en el tipo de cambio, lo cual hizo que aumentara el costo del seguro, más no la cantidad de unidades enviadas aseguradas.

Limpieza de datos - texto

La limpieza de datos es uno de los procesos más importantes y complicados dentro del procesamiento de información. Este proceso consta de corregir o eliminar datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de información. Cuando se combinan varias fuentes de datos, se dan muchas circunstancias para que los datos se dupliquen o se etiqueten incorrectamente. Para este caso en particular, podríamos considerar que la información estaba limpia. Todas las columnas tenían campos bien definidos, con parametros únicos constantes y tipos de datos uniformes en cada registro. Encontramos un par de datos NAN, que a mi parecer eran consecuencia de renglones residuales generados por la exportación de la información.

	word	abs_freq	abs_perc	abs_perc_cum	wtd_freq_perc	wtd_freq_perc_cum
0	pantalon	179709	0.0494	0.0494	0.0494	0.0494
1	punto	165344	0.0454	0.0948	0.0454	0.0948
2	camiseta	140938	0.0387	0.1335	0.0387	0.1335
3	dama	129360	0.0355	0.1690	0.0355	0.1690
4	algodon	99132	0.0272	0.1963	0.0272	0.1963
5	poliester	85301	0.0234	0.2197	0.0234	0.2197
6	hombre	78071	0.0214	0.2411	0.0214	0.2411
7	fibras	68682	0.0189	0.2600	0.0189	0.2600
8	mujer	66646	0.0183	0.2783	0.0183	0.2783
9	caballero	61070	0.0168	0.2951	0.0168	0.2951

Figura. 4

Sin duda, el mayor reto en la limpieza estuvo en estandarizar la columna ‘descripcion’. Esta columna, como su nombre lo indica, contiene textualmente la descripción del producto comercializado, conteniendo datos como: tipo de prenda, color, talla, tela, agregados (estampados, diamantes), identificador del comerciante y algunas veces, se describía el uso que se le debía de dar a la prenda (ropa de baño, traje de baño). Hubo dos principales obstáculos al trabajar con esta columna:

1. Las dimensiones del DataFrame, generaban una importante carga computacional que impedía probar formas para limpiar la información ya que nunca terminaba de correr. Al principio intentamos hacerlo mediante Python, sin embargo, terminamos trabajando la limpieza en Excel al sustituir todas las stopwords por espacios vacíos.
2. Las stopwords. Estas palabras hacen referencia a aquellas palabras que no están registradas por los robots de Google, las cuales carecen de sentido cuando se escriben solas o sin la palabra clave o keyword. Las palabras que pueden ser consideradas Stop Words, pueden depender del idioma. Son, básicamente, conjunciones, artículos, preposiciones y adverbios. “en”, “y” “como”.

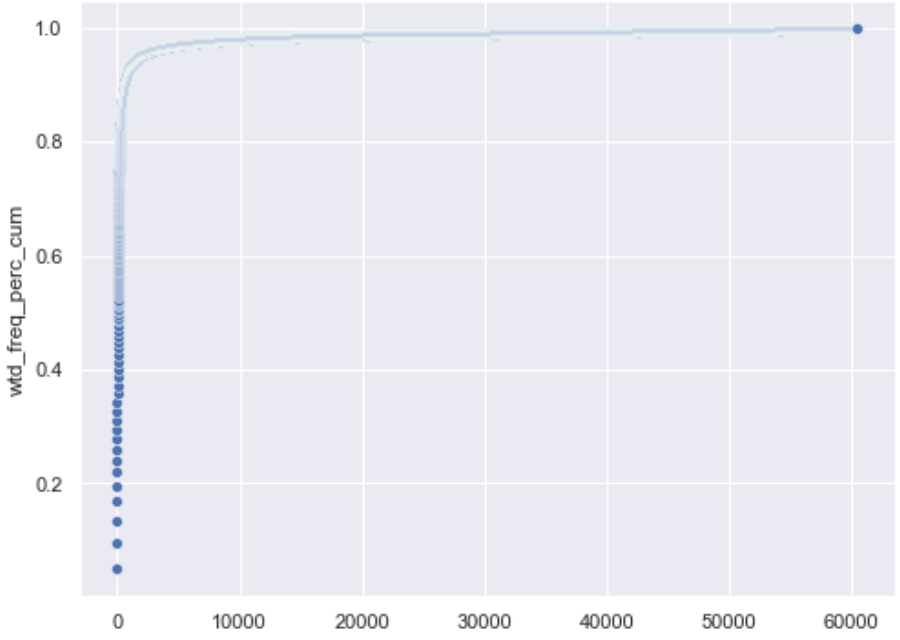

Figura. 5

Tokenize words

Como parte de este proyecto, una de las principales tareas era la de procesar el texto contenido dentro de la principal variable en cuestión, la descripción textual de los productos importados. Como ya mencionamos anteriormente, en cada registro, se complementaba con una frase, a veces más o a veces menos completa acerca del producto. Esta tarea se lleva mediante el PLN o Procesamiento de Lenguaje Natural consiste, básicamente, en el desarrollo de codigos capaces de interpretar los lenguajes humanos.

Es una disciplina muy amplia, relacionada con ámbitos tan complejos y dispares como la inteligencia artificial, la lingüística, los lenguajes formales y los compiladores. Las herramientas de procesamiento que usamos en ciencia de datos normalmente trabajan con números. Cuando queremos trabajar con texto, nos interesa transformar dicha información en datos numéricos. Es ahí donde entra en juego el proceso de tokenización. Consiste en la segmentación del texto en frases o palabras. Además de separar una frase en tokens o grupos de tokens (ngrams), podemos identificar su frecuencia de aparición; identificar si se trata de sustantivos, verbos, etc.; eliminar palabras habituales que no aportan significado (stop-words); realizar correcciones ortográficas, homologar conceptos, etc.

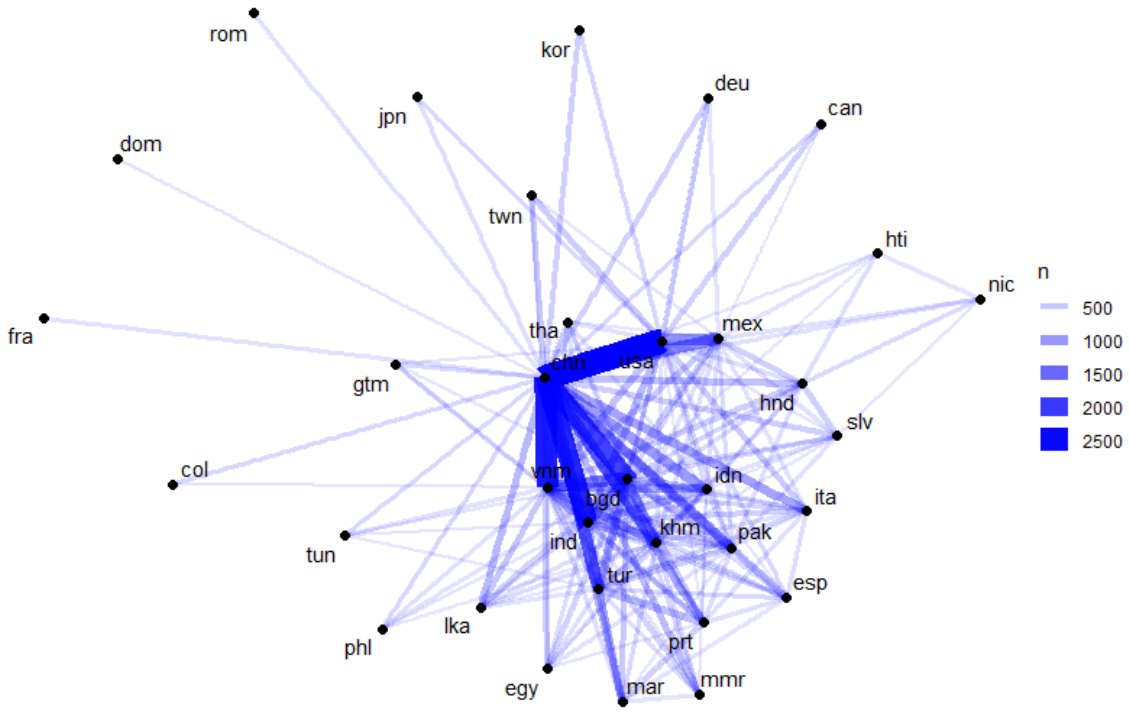

Figura. 6

Word Relationships

En nuestro caso, el proceso de tokenización fue relativamente fácil. Sin embargo, como mencionamos anteriormente, el proceso conlleva descartar una serie de palabras que, en caso contrario, solo generarían ruido e impedirían llegar a conclusiones optimas.

En la gráfica siete que presentamos a continuación podemos observar como las principales palabras, contenidas dentro del DataFrame, son aquellas que hacen referencia al género para el cual está hecha la prenda, después siguen las palabras que describen los materiales de estas y por último algunos métodos de confección como ‘punto’, ‘tejido’ y demás métodos de costura.

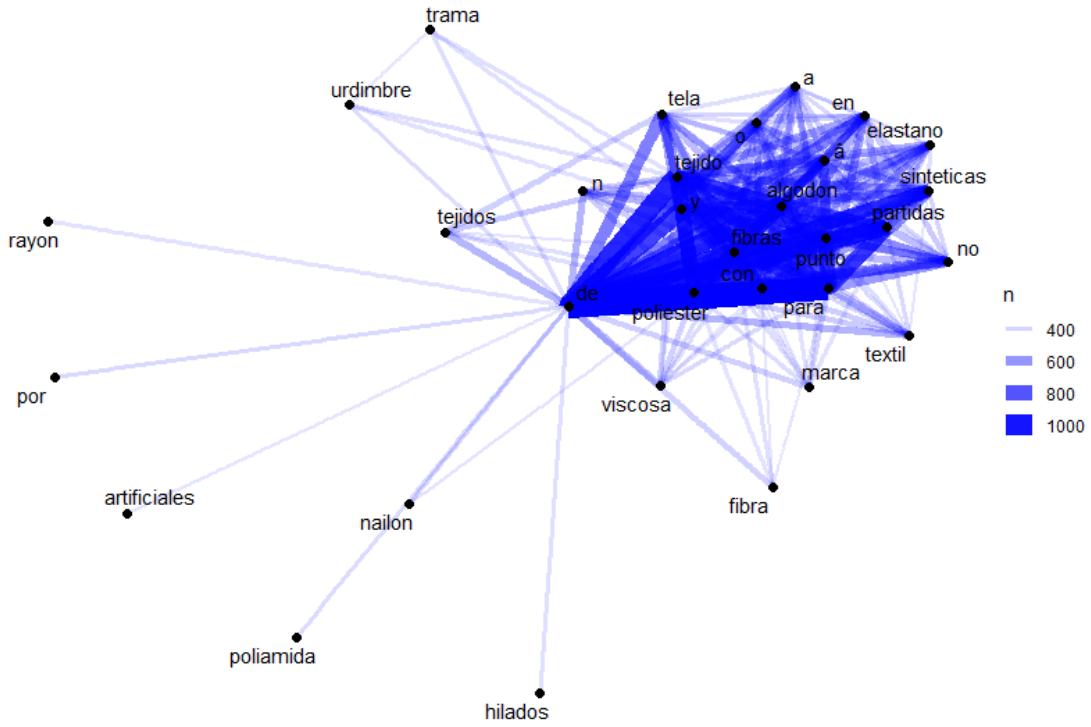

Figura. 7

Lo interesante de este método es que permite tener una idea general bastante acertada del contenido del texto. MWE (multi-world expression). Una vez que concluimos el proceso de manera individual, palabra a palabra, cambiamos la dinámica a incluir frases. Este tokenizador agrupa frases de N palabras (gráfica ocho) las cuales suelen ir juntas. Mayormente, igual que en el caso anterior, los ejemplos más comunes son aquellos en los que se agrupa el género de la prenda, seguido del material y por último en tipo de tejido.

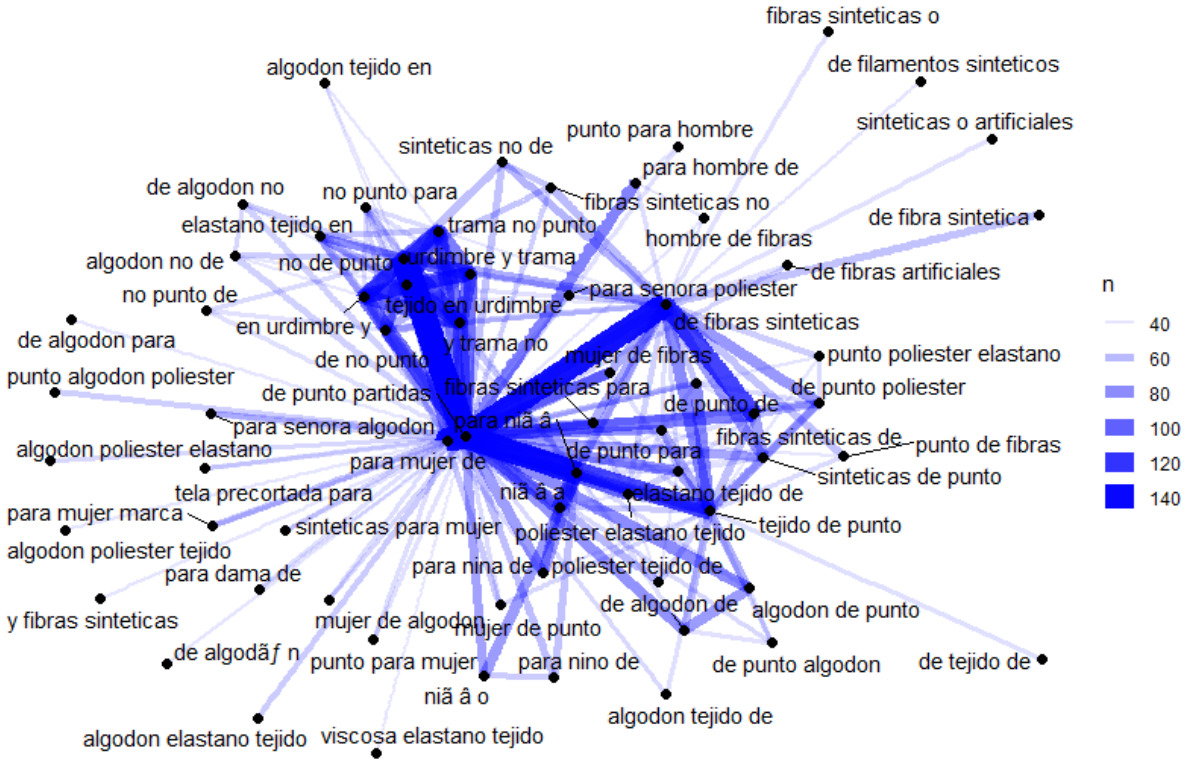

Figura. 8

Ingeniería de características

La ingeniería de características es el proceso de selección y/o transformación de los datos, de forma que puedan ser entradas del mayor provecho posible, para las técnicas de machine learning. La ingeniería de características persigue un objetivo claro: obtener el mejor desempeño posible de los modelos de machine learning, por esta razón es que los científicos de datos consumen más del 80% del tiempo en esta etapa. Algunas de las razones por las que se justifica la dedicación suficiente y necesaria en esta fase de un proyecto de ciencia de datos, se exponen en el siguiente apartado.

1. Mejores características se traducen en modelos simples, con características bien diseñadas, se pueden elegir “los parámetros incorrectos” (óptimos) y aun así obtener buenos resultados.

2. Mejores características significan mejores resultados, el esfuerzo en la etapa de transformación de los datos, es una cuesta que nos conduce con cierta propiedad a los mejores resultados posibles.

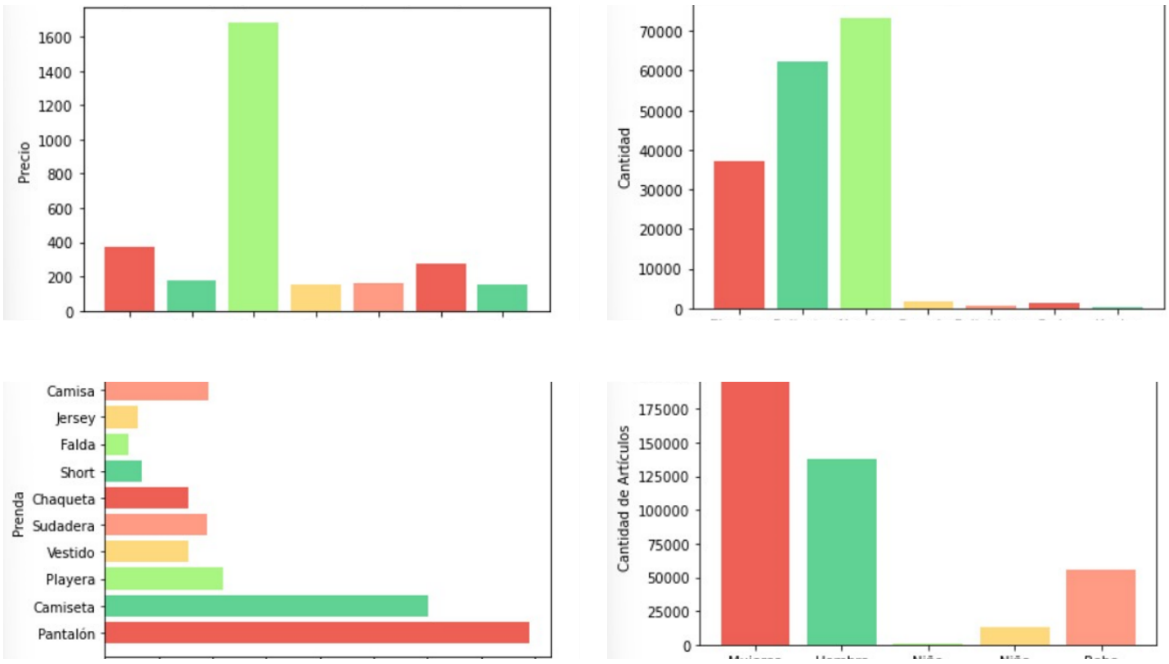

Figura. 9

Variables dummy

La idea detrás de las variables ficticias (dummy) es reemplazar una variable categórica con una o más características nuevas que pueden tener los valores 0 y 1.

A fin de obtener una visualización gráfica de características, para lograr mayor comprensión de los datos con los que se cuenta, se procedió a clasificar los datos dependiendo de rasgos distintivos, como lo es el tipo de prenda, tipo de tela que se utiliza y las prendas que están especificadas a un público (mujeres, hombres o niños).

Siendo el propósito del modelo final enfocado al precio, se consideraron tambien características como países más caros de importar y precios más altos por prenda. Con estas visualizaciones gráficas se logró un mayor entendimiento de los datos, así como características generales de los mismos. Se pudo concluir que el 95% de las importaciones textiles viene de únicamente 25 países, de los cuales Alemania, Italia y EUA ocupan los primeros lugares en cuestión de precios más altos. Así como también se observó que los pantalones es la prenda de la que se importa en más cantidad, en cuestión de telas el primer lugar lo ocupa el algodón, seguido del poliéster.

Creación de modelos

Para realizar la siguiente lectura de los datos no fue necesario utilizar ninguna herramienta de Machine Learning debido a que la categorización de los datos pudo ser desarrollada de manera más optima por métodos sencillos de análisis de datos, por lo tanto, se comenzó por la limpieza de los datos para poder utilizarlos de manera adecuada. Se organizaron creando inicialmente una variable en la cual se localice el precio de cada artículo. Las variables de entrada dentro del proceso son específicos artículos los cuales son las prendas pantalón, camiseta, camisa, playera, short, falda, chaqueta, sudadera, suéter y vestido, se guardaron a manera de variables dummies para la contabilización de estos.

	fraccion	clase_precio
0	54024601	(59.488, 145.72]
1	62044391	(513.6, 49438.0]
2	62044391	(513.6, 49438.0]
3	55092101	(30.660999999999998, 35.145]
4	61091002	(89.0, 19272.0]
5	61051003	(24.799, 97.881]
6	62121004	(141.951, 2796.0]
7	61099003	(55.0, 72.218]
8	62029391	(476.87, 12710.0]
9	62046291	(199.0, 30162.0]

Figura. 10

	Artículo	# por encima del 75 %
0	Pantalón	5947
1	Camiseta	36518
2	Camisa	13619
3	playera	15526
4	short	5532
5	falda	3002
6	chaqueta	11413
7	sudadera	12307
8	sueter	4354
9	vestido	9818

Figura. 11

En la figura número 10, podemos ver la cantidad de veces que un artículo presenta un precio atípico. Este resultado se obtuvo tras realizar la categorización de los artículos de acuerdo con su precio. Un elemento importante por destacar es que las camisetas son las que presentan mayor cantidad de valores atípicos, con un total de 36,518, por otro lado, el artículo que menos valores atípicos presenta es la falda con un total de 3002.

Estos resultados, en un escenario aplicado nos hablan de la variabilidad de los precios, es más probable encontrar una camisa con un valor atípico, que una falda. Este modelo nos sirvió para identificar dentro de los artículos seleccionados los valores que representan más precios atípicos de estas prendas, es funcional para descubrir la varianza de precios de cada uno de ellos , su utilidad dentro del proyecto es formar un orden por fracción para la facil identificación de la relación de precio con cualesquier articulo deseado dentro de los artículos de la base.

Conclusiones

Frida Hernández

Durante la realización del proyecto enfrenté retos académicos a nivel personal como en conjunto con mis compañeros. Retos que me permitieron obtener una visión mucho más completa y real de la implementación de la Ciencia de Datos en problemas reales. Desde analizar una base de datos grande y limpieza de esta, hasta la creación del modelo. En cuestión del análisis de datos, aprendí muchísimas herramientas de visualización que mejoran la presentación al usuario, así como también me dan a mí claridad acerca de los datos a trabajar.

Fortalecí también habilidades de comunicación y trabajo en equipo al llevar a cabo la totalidad de las prácticas en modalidad 100% virtual, obligándonos así a delegar tareas entre todos y compartir resultados enriqueciendo el trabajo propio. Esto también resaltó las fortalezas de cada integrante y la importancia del enriquecimiento grupal.

Finalmente, mi destreza en cuestión de programación se incrementó. Al enfrentar problemas de tecnicismos y/o adecuarme a un lenguaje que no domino, me vi obligada a recurrir a la búsqueda de información y aplicación de las potenciales soluciones que pudiera encontrar. Desde la primera sesión los datos me parecieron muy interesantes, sin embargo, no pensé que se podría obtener tanta información de estos.

Leonardo Aceves

A lo largo de este proyecto me enfrente a una serie de cuestiones que hicieron muy enriquecedor el programa. Al igual que me sucedió en el PAP anterior, al principio me veía con una renuencia a utilizar nuevas tecnologías y lenguajes de programación, como en este caso R.

Creo que uno de los principales aprendizajes que me llevo es el de tener apertura para poder implementar nuevas herramientas que ayudan a explorar data desde otro enfoque. Hablando desde el punto de vista de los datos. Pienso que fue un reto importante el trabajar con texto no estructurado como lo fue el campo de descripción de los artículos comercializados. Como ya es bien sabido, el mundo laboral genera cantidades exorbitantes de información que no siempre vienen de manera estructurada y este proyecto me abrió el panorama para poder enfrentar estas situaciones a futuro.

Me hubiese gustado tener más tiempo para poder explorar más técnicas, comparar los métodos del resto de los grupos dentro del proyecto y tener más tiempo para explorar el dataframe a profundidad.

Iván Lafarga

Durante el desarrollo de este pap me di cuenta una de las cosas más importantes que tienen que tener en cuenta todos los analistas al revisar los casos, esto es la sensibilidad del tema, desarrollando el proyecto tenemos muchos resultados, tratamiento de datos, desarrollo de conclusiones, lo más importante fue tener en cuenta que los números que estamos trabajando tienen un impacto social en referencia al uso de las prendas que considero un uso inconsciente hoy en día , así como las lecturas de estos mismos son personas con realidades, vida, familia y seres queridos las cuales estan detras de todo el proceso del desarrollo de estas prendas. Tener en cuenta que la base de datos mayormente se trata de prendas de uso común y hay que tener en cuenta su debida categorización respecto a que hay un mal manejo de los nombres de prendas pues sus registros no tienen estandarización en la base de datos respecto a su nombramiento.

Bibliografía

the pandas development team. (2008-2022). Pandas. julio 4, 2022, de Pandas Documentation Sitio web: https://pandas.pydata.org/docs/index.html

The Carpentries. (2018). DataCarpentry. junio 28, 2022, de Data Carpentry Sitio web: https://datacarpentry.org/

DelfStack. (2020). Los Howto's de Python. junio 13, 2022, de DelfStack Sitio web: https://www.delftstack.com/

Python Software Foundation. (2022). Python 3.10.5 documentation. julio 20, 2022, de Python Sitio web: https://docs.python.org/3/

Hyndman, R., Athanasopoulos, G. (2018). Forecasting: Principles and Practice. Monash University, Australia: Texts.