INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Sustentabilidad y tecnología

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

ITESO, Universidad Jesuita de Guadalajara

4J07A Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

Modelo predictor de precios atípicos en importaciones de México

PRESENTAN

Programas educativos y Estudiantes

Lic. en Ingeniería Financiera. Iván Andrés Arellano Ruelas

Lic. en Ingeniería Financiera. Pablo Alejandro Rivera Sánchez

Lic. en Ingeniería Financiera. Andrés Ramírez Villanueva

Lic. en Ingeniería Financiera. Juan Pablo Rodríguez Alonso

Profesora PAP: Diana Paola Montoya Escobar

Tlaquepaque, Jalisco, Julio del 2022

Índice

Contenido

REPORTE PAP

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) (co-participación) 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, co-crear 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".

El Reporte PAP consta de tres componentes:

El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de este y la valoración de las incidencias en el entorno.

El segundo componente presenta los productos elaborados de acuerdo con su tipología.

El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

El propósito general del PAP "Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio" es que a partir de una base de datos de importaciones del estado de Jalisco se busca generar un modelo de predicción mediante caracteres y precios, para lograr identificar precios atípicos dentro de las importaciones. Con la base de datos que cuenta con más de un millón de registros, se pueden observar datos importantes, como la descripción que cada uno de los importadores deben llenar acerca de su producto. Este apartado de la base de datos llamado descripción no está obligada a seguir ningún formato ni palabras específicas, por lo que el grado de complejidad del modelo buscado es mayor al tener que realizar una limpieza exhaustiva de la base de datos. Como se mencionó, lo primero realizado fue una limpieza del texto contenido en la base de datos, para poder buscar la información en la que tenemos interés investigar, como la relación de los precios con los productos. Después de limpiar la base se fijó como objetivo el realizar un modelo que pueda predecir los precios atípicos para los productos.

1. Ciclo participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa construir sus conocimientos para dar respuestas a problemáticas de un contexto específico y en un tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

1.1 Entendimiento del ámbito y del contexto

El comercio exterior es uno de los factores económicos más importantes para México, siendo el país número 181 del ranking mundial, y puesto 45 en cuanto a las exportaciones. La ciencia de datos en una herramienta del presente y del futuro que puede ayudar a las empresas y al gobierno a realizar análisis exploratorio a las bases de datos que tienen y realizar pronósticos o poder identificar anomalías de manera automática, pero sabemos que, al ser tecnología reciente, no es aplicado en áreas donde podría ser de gran utilidad.

Las importaciones, de tantos productos, y en bases de datos sucias, pueden ser abrumadoras para analizar y observar precios que pueden llegar a ser atípicos, y que cause una señal de alerta para los empresarios en puestos de responsabilidad, pero con herramientas computacionales es posible.

1.2 Caracterización de la organización

Este proyecto estuvo conformado por un equipo de trabajo de 11 personas, las cuales actuaron en el rol de ser analista de datos. El equipo estuvo encargado de realizar diferentes análisis y obtener conclusiones de varias variables que se encontraban en una base de datos de las importaciones de México.

Para realizar dichos análisis, cada semana se realizaron diferentes entregas con el objetivo de obtener cada vez más información de la base de datos. Se realizaban 3 entregas cada semana por equipo, es decir, el equipo de trabajo de 11 personas se subdividió en 3 sub-equipos para realizar la misma entrega cada semana. Esto se realizó con el objetivo de que cada equipo tuviera resultados y métodos diferentes para poder comparar y aprender entre todos, además, esto permitió que al final se obtuvieran 3 diferentes tipos de modelos de inteligencia artificial.

El propósito final de realizar todos estos análisis fue entender la base de datos y obtener diferentes características de la descripción de todos los productos importados con el fin de obtener un modelo de clasificación que determine si el precio de un producto es atípico o no. Esto puede ayudar a detectar actividades ilícitas en las importaciones ya que se detectarán *outliers* en los precios.

1.3 Identificación de la(s) problemática(s)

Como ya mencionamos, la base de datos era un problema por lo sucia que estaba, tuvimos que utilizar herramientas diversas de Excel y de lenguajes de programación como R y Python, pudimos realizar las correcciones en las faltas de ortografía, sinónimos de palabras (como englobar pantalón y pantalones en la palabra pantalón, por ejemplo). El no tener una

base de datos limpia da a pie a complicar el análisis y observación de *outliers* que pudiesen enriquecer la información sobre lo que se está trabajando, y con bases de datos tan amplias como la que manejamos para este proyecto, se puede complicar más, pero viendo la relevancia de la base es primordial realizar la limpieza correspondiente.

El no detectar precios atípicos, puede dar pie a permitir actividades ilícitas, como lavado de dinero, y un algoritmo que detecte estos precios puede llegar a ser una herramienta de gran valor, que evite problemas y consecuencias de gravedad.

Al comenzar con las soluciones, tuvimos problemas para aplicar modelos de clasificación que puedan identificar los precios atípicos por el tamaño de la base de datos y la representación no numéricas de las variables de cada producto (como país de origen, por ejemplo), entonces tuvimos que codificar como número estas variables y trabajar a partir de esto, además de que la capacidad computacional que conlleva adaptar un modelo clasificatorio a una base de datos tan amplia debe ser potente, y aun así solo tomar una muestra de la base. CONTINUAR

1.4 Planeación de alternativa(s)

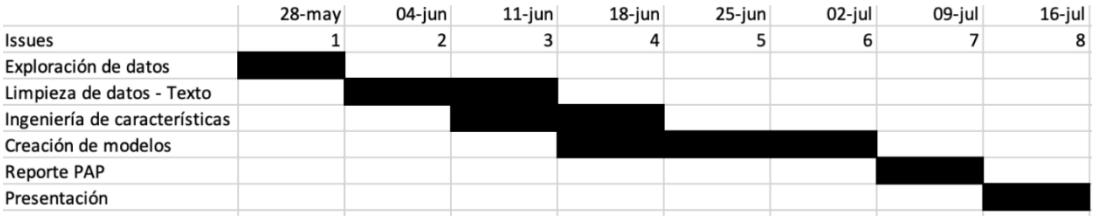
Se realizó en fases como Issues, en la que comprendimos un poco sobre la problemática con la base de datos y lo que podríamos solucionar.

Exploración de datos, que es de estudio sobre la base de datos, como está conformada y como nos puede añadir al estudio.

Limpieza de datos – Texto, que conformó la limpieza de la base de datos, como mencionamos previamente, la columna de descripción, que describe cada producto de manera no uniforme, por lo que la limpieza requirió trabajo.

Ingeniería de características, que consistió en la creación de nuevas variables a partir de texto, como lo es pantalón de dama o camisa de hombre, que no estaban como tal en la base, pero se pudieron identificar ya con la base de datos limpia.

Creación de modelos, aplicar modelos de clasificación, como el *random forest* o el *support vector machine*, que nos pueda identificar precios atípicos de las variables creadas en la ingeniería de características.



1.5 Desarrollo de la propuesta de mejora

Entrega 1

Introducción

Para esta entrega, el trabajo realizado consistió en llevar a cabo una breve exploración de la base de datos con la que trabajaremos a lo largo del proyecto. Dicha exploración consiste tanto en un análisis por medio de tablas con características de los datos, así como visualizaciones que ayudan a ver cómo se comportan los datos.

Procedimiento

La exploración de los datos se llevó a cabo realizando principalmente los siguientes análisis:

- Realizar un reporte de calidad de datos
- Realizar un análisis de correlación entre las variables
- Obtener gráficas (histogramas) de algunas de las variables
-

Resultados

Data Quality Report

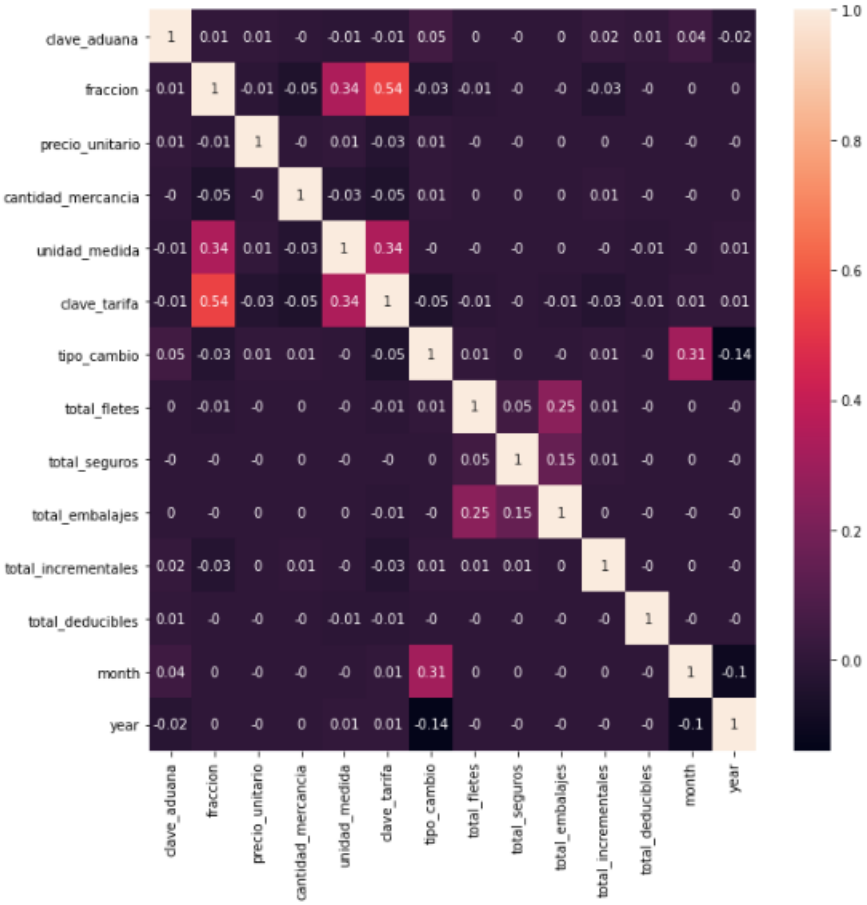
Primero que nada, realizamos un breve reporte de calidad de datos en donde obtenemos información del *Dataset* importante, como por ejemplo los valores mínimos, valores máximos, valores únicos, entre otras características. A continuación, se muestra dicho reporte.

	Nombres	Data_Types	missing_values	present_values	unique_values	min	max
Unnamed: 0	Unnamed: 0	int64	0	1021082	112102	0	118643
clave_aduana	clave_aduana	int64	0	1021082	59	20	840
fraccion	fraccion	int64	0	1021082	1546	50040001	63109099
subdivision	subdivision	object	510433	510649	26	NaN	NaN
descripcion	descripcion	object	0	1021082	166955	''T-SHIRTS'', CAMISETAS DE PUNTO DE ALGODON	Ã¿PANTALON LARGO
precio_unitario	precio_unitario	float64	0	1021082	340232	0.0	14720000.0
cantidad_mercancia	cantidad_mercancia	float64	0	1021082	66177	0.001	47707792.0
unidad_medida	unidad_medida	int64	0	1021082	16	1	21
clave_tarifa	clave_tarifa	int64	0	1021082	5	1	9
valor_agregado	valor_agregado	float64	0	1021082	1	0.0	0.0
pais_origen_destino	pais_origen_destino	object	0	1021082	169	ABW	ZYA
tipo_cambio	tipo_cambio	float64	0	1021082	1197	9.3578	25.1185
total_fletes	total_fletes	float64	0	1021082	41776	0.0	1453186846.0
total_seguros	total_seguros	float64	0	1021082	7108	0.0	292964799.0
total_embalajes	total_embalajes	float64	0	1021082	1915	0.0	1314976.0
total_incrementales	total_incrementales	float64	0	1021082	21868	0.0	71089108.0
total_deducibles	total_deducibles	float64	0	1021082	21	0.0	1003824.0
month	month	int64	0	1021082	12	1	12
year	year	int64	0	1021082	2	2020	2021

Gracias al reporte, nos dimos cuenta de que hay 2 columnas que no aportan información alguna, la primera, llamada "Unnamed: 0", es únicamente un contador de las filas de la base de datos. Por otra parte, la segunda columna eliminada fue la de "valor_agregado", esto ya que tiene, como se observa en el reporte, tanto el valor mínimo como el máximo es 0, por lo que indica que esta columna está compuesta únicamente de ceros.

Análisis de correlación

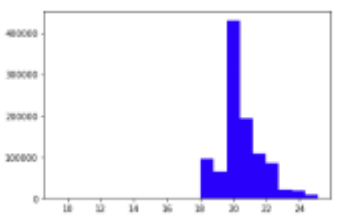
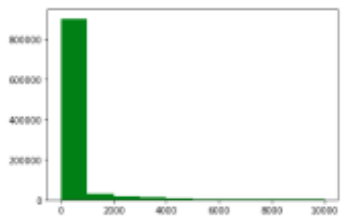
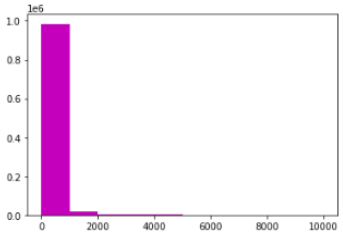
Posteriormente, realizamos un análisis de correlación entre las variables de la base de datos, estos fueron los resultados:



Podemos observar que los dos pares de variables con una correlación más alta son clave_tarifa con fracción, con una correlación positiva de 0.54, así como fracción con unidad_medida, con una correlación también positiva de 0.34. Ambas correlaciones son relativamente bajas como para considerar la eliminación de alguna de estas variables, de igual manera, no encontramos ninguna correlación negativa alta.

Histogramas

Finalmente, realizamos algunos histogramas de algunas variables tales como el tipo de cambio, cantidad de mercancía y precio unitario, obtuvimos lo siguiente.

Tipo de Cambio	Cantidad de Mercancía	Precio Unitario
		

La primera gráfica nos muestra que el tipo de cambio, a lo largo del tiempo, ha estado en un rango de entre 18 y 25. Las otras dos gráficas nos indicaron que tanto la cantidad de mercancía como el precio unitario tiene outliers, es decir, hay valores extremadamente grandes que sesgan la gráfica.

Conclusiones

Gracias al pequeño análisis realizado, pudimos obtener información importante de algunas de las variables de la base de datos, tanto por medio del reporte de calidad de datos como por medio de los histogramas. Eliminamos algunas variables que no parecían relevantes y observamos que ninguna variable está lo suficientemente correlacionada con otra para poder eliminarla. Finalmente, los histogramas nos indican que hay valores muy grandes en algunas variables, por lo que sería pertinente realizar un análisis más a detalle para determinar por qué sucede esto. En las entregas siguientes analizaremos profundamente algunos aspectos de las bases de datos.

Entrega 2

Introducción

Para esta entrega, realizamos la primera limpieza de palabras de la base de datos y nos quedamos con la información más concreta posible. Realizamos un diccionario para relacionar las palabras similares o mal escritas y tener la definición que nos parecía más adecuada. Analizamos las palabras más repetidas para conocer más a fondo la base de datos, y observamos el comportamiento del precio con los productos principales, con los países listados y por mes, separando los datos del 2020 y 2021.

Procedimiento

Primero hicimos una primera limpieza de los datos en excel, con herramientas de diccionario que ya tiene integrado el programa, para reducir el espectro de palabras mal escritas, plural, números, todo esto trabajando en la columna de descripción, que básicamente define el producto. Después de esto, realizamos una segunda limpieza en python, eliminando caracteres que no nos importaban, como ·, #, ?, ¡, de igual manera, eliminamos los conectores del enunciado (de, las, los) para tener únicamente la descripción del producto. Ya con la base más limpia realizamos el primer análisis de precio por país, para proseguir con el precio promedio de algunos productos como las playeras, y finalmente analizar el monto de dinero transaccionado, mes a mes de ambos años.

Resultados

Diccionario

El diccionario, debido a que no lo pudimos hacer de manera automática, es decir, las palabras del mismo había que introducirlas de manera manual, fue realizado mediante Excel. En pocas palabras, utilizamos las herramientas de ortografía de Excel para limpiar la columna (primera palabra) de descripción, se podría decir que Excel fue nuestro diccionario para limpiar la base de datos.

Eliminación de palabras innecesarias (limpieza)

En cuanto a la eliminación de palabras innecesario, la siguiente gráfica muestra un antes y después de la columna de descripción, ya después de haber eliminado ciertos símbolos y ciertas palabras conectoras.

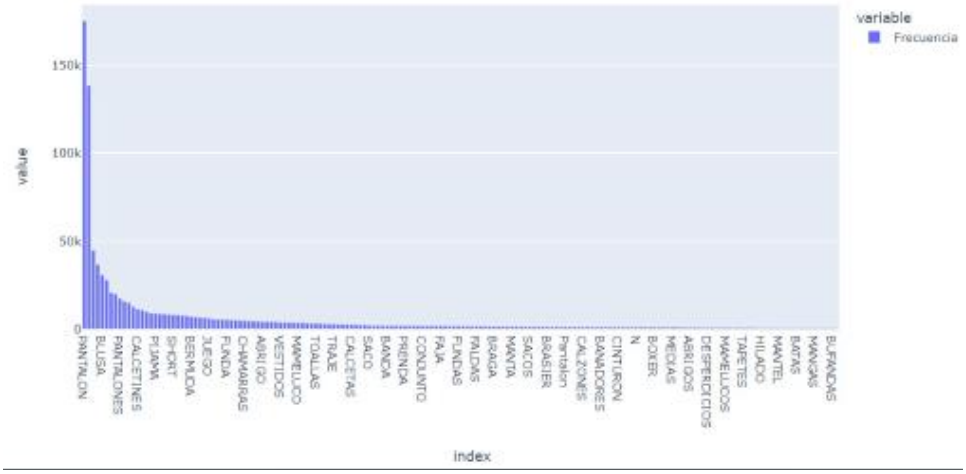
Antes Después

descripcion	new_desc
tejidos de fibras sinteticas 100% poliester (t...	tejidos fibras sinteticas poliester teâ¥ido
tejidos de fibras sinteticas 100% poliester(bl...	tejidos fibras sinteticas poliester blanqueados
manoplas de kevlar/piel id#:mx4kev	manoplas kevlar/piel
batas de 100% poliester id#:3470-l	batas poliester
batas de 100% poliester id#:3470-xl	batas poliester

Como se puede observar, después de realizar la limpieza la columna de descripción es más entendible, corta y concreta, haciendo que su análisis sea más sencillo.

Frecuencia de palabras

Se realizó un conteo de la frecuencia de cada palabra, cabe recalcar que nuevamente se eliminaron símbolos y palabras conectoras, estos fueron los resultados.



Vemos que la palabra más repetida es pantalón, seguida de blusa, pantalones, calcetines, entre otras.

Diccionario (actualización)

Para esta segunda entrega, logramos crear un diccionario que nos corrigiera la mayor cantidad de palabras posibles de manera semi-automática, utilizando Python y Excel, a continuación, se muestra el proceso.

Teniendo ya la columna de descripción relativamente limpia, sin símbolos y palabras innecesarias, y también teniendo la frecuencia de palabras de dicha columna ya limpia, tomamos todas las palabras que se repitieran al menos 50 veces para trabajar sobre las mismas.

	Frecuencia
pantalon	179820
punto	172112
camiseta	141179
dama	129396
algodon	102499
poliester	91324
hombre	78156
fibras	68728
mujer	66732
caballero	61080
tejido	59083
sinteticas	58315
bebe	56683
niâ¥a	55730
niâ¥o	53881

Tomamos cada una de estas palabras y, con ayuda de una librería de Python, buscamos todas las palabras que iniciaran de manera similar a la palabra elegida e iteramos sobre cada una para encontrar todas las coincidencias. Obtuvimos el siguiente resultado.

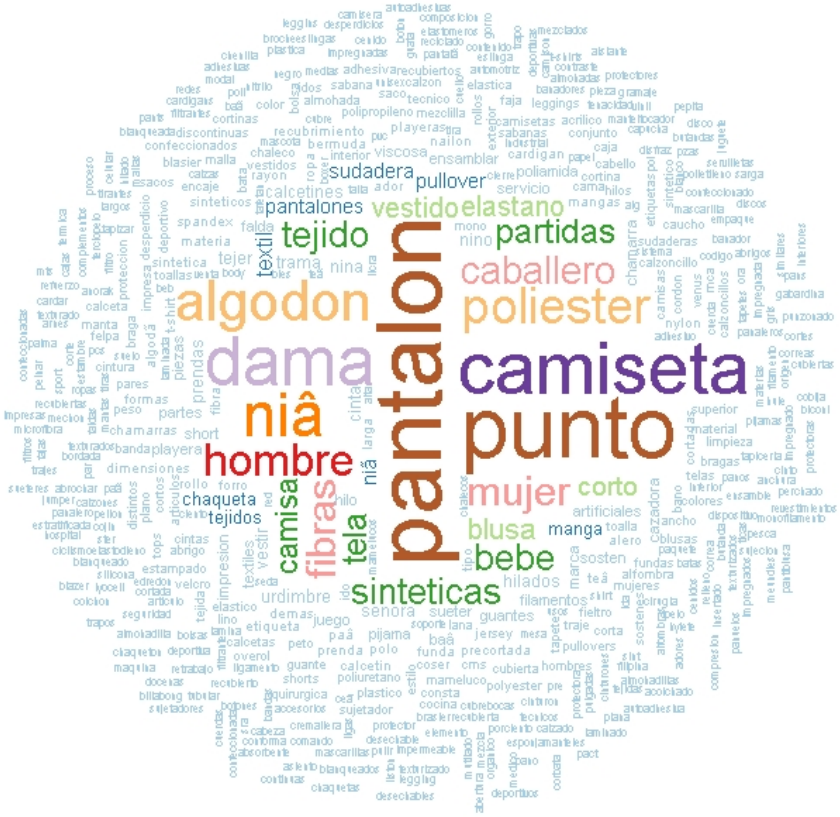
	0	1	2	3	4	5	6	7	8	9
0	pantalon	pantalones	None	None	None	None	None	None	None	None
1	punto	punta	puntas	puntos	None	None	None	None	None	None
2	camiseta	camisetas	None	None	None	None	None	None	None	None
3	algodon	algodon	algodon	None	None	None	None	None	None	None
4	poliester	poliesteres	poliester	poliester	poliester	poliester	poliester	None	None	None
...	...	...	...	...	...	...	...	...	...	...
556	gramaje	granel	gramos	grande	graduada	gray	None	None	None	None
557	reforzada	reforzado	None	None	None	None	None	None	None	None
558	copas	copetuda	copa	None	None	None	None	None	None	None
559	salvavidas	salida	saltos	saldos	None	None	None	None	None	None
560	mascadas	mascada	None	None	None	None	None	None	None	None

Como podemos observar, los resultados no fueron del todo correctos, esto ya que la forma de encontrar coincidencias en la palabra no es perfecta. Por lo tanto, teniendo ya solamente 561 filas en el DataFrame anterior, utilizamos Excel para hacer las correcciones necesarias y crear nuestro diccionario.

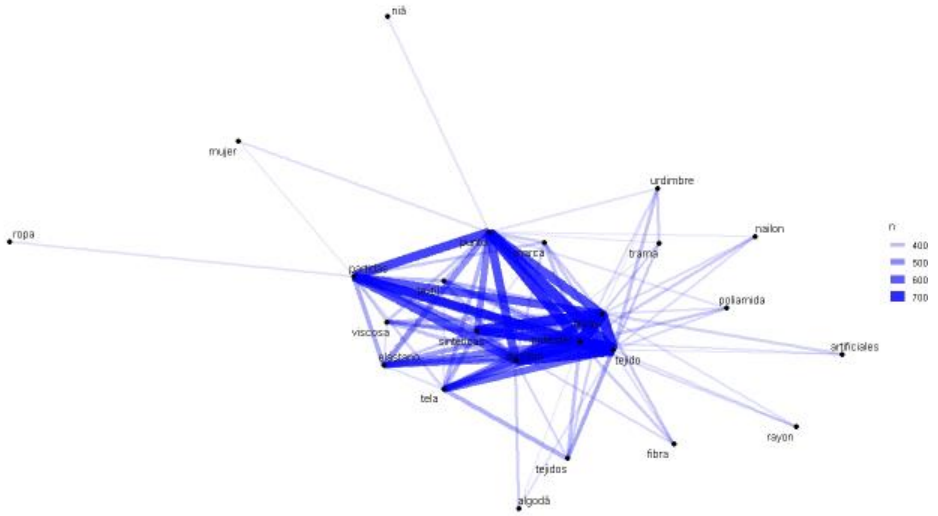
```
dict<-dictionary(list(
  pantalon=c('pantalon','pantalones','panalon'),
  punto=c('punto','punta','puntas','puntos'),
  camiseta=c('camiseta','camisetas','camiseta','camisa','camisas','camisetas','camison','camisera','camisola','camison','camisones',
  algodon=c('algodon','algodon','algodon'),
  poliester=c('poliester','poliesteres','poliester','poliester','poliester','poliester','poliester'),
  hombre=c('hombre','hombres','hombreras'),
  mujer=c('mujer','mujeres','mujer'),
  tejido=c('tejido','tejidos','tejida','tejidas'),
  sinteticos=c('sinteticas','sinteticos','sintetica','sintetico'),
  bebe=c('bebe','bebes'),
  ninia=c('ni??a','ni??as','ni??a','nina','ninas','nina','ninas','ni?'),
  ninio=c('ni??os','ni??o','nino','ninos','nino','ninos'),
  tela=c('tela','telas'),
  blusa=c('blusa','blusas','bluson','blusa'),
  vestido=c('vestido','vestidos'),
  textil=c('textil','textiles'),
  pullover=c('pullover','pullovers'),
  sudadera=c('sudadera','sudaderas'),
```

Análisis entre palabras (actualización)

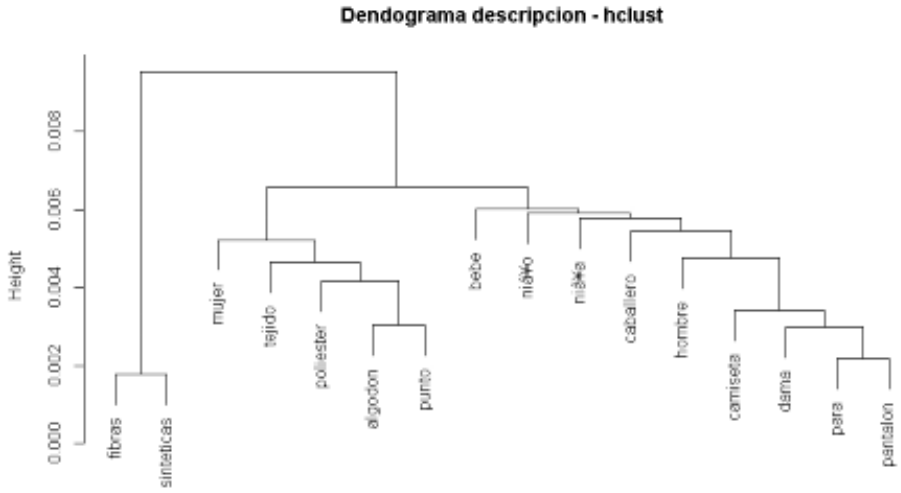
Teniendo entonces la columna de descripción limpia y habiendo ya aplicado nuestro diccionario a dicha columna, procedimos a realizar los siguientes análisis para ver la relación entre todas las palabras. Primero, realizamos un Wordcloud, donde podemos observar que el diccionario se aplicó correctamente, al únicamente aparecer pantalón y no pantalones.

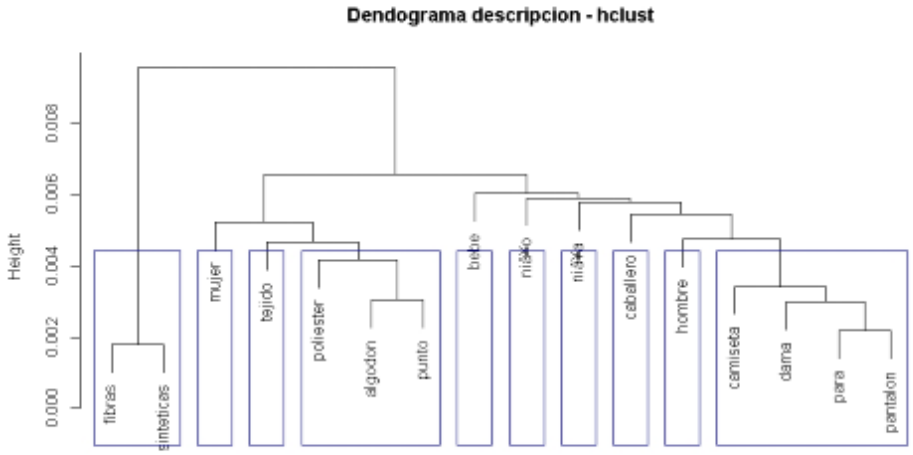


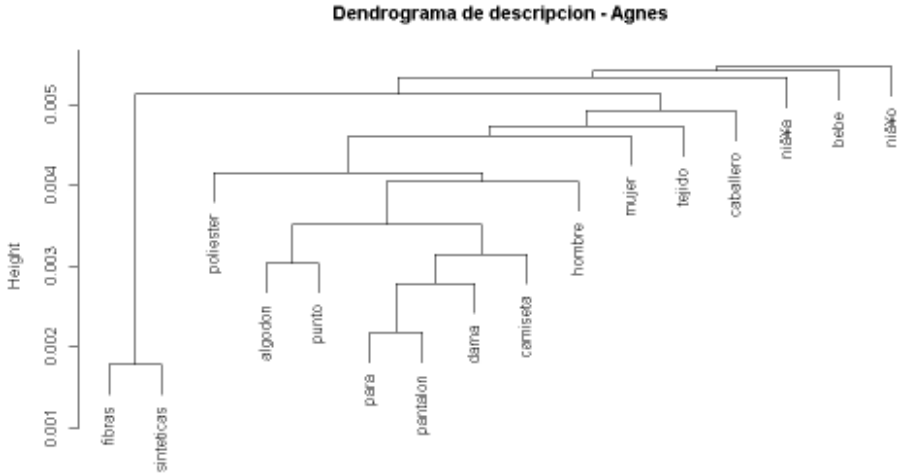
Posteriormente, realizamos un análisis de pares en donde estos fueron los resultados.

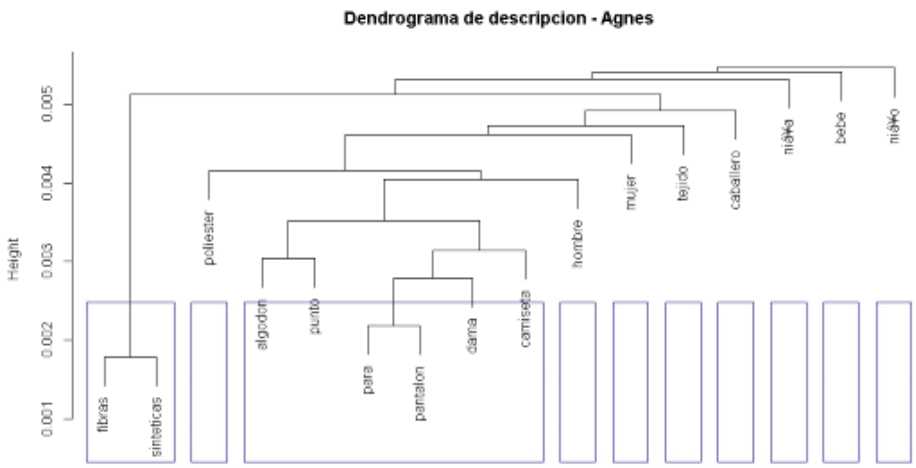


De igual manera, se realizaron matrices de distancia y *agglomerative nesting* de la columna de descripción, lo cual nos dio las siguientes gráficas.



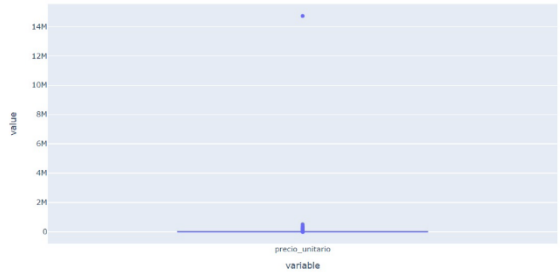
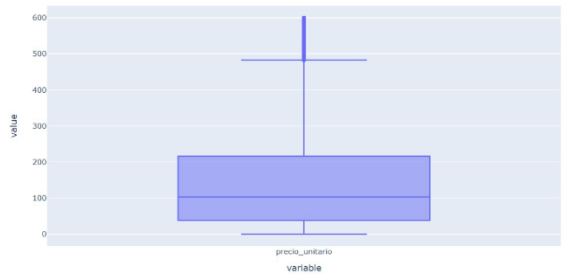
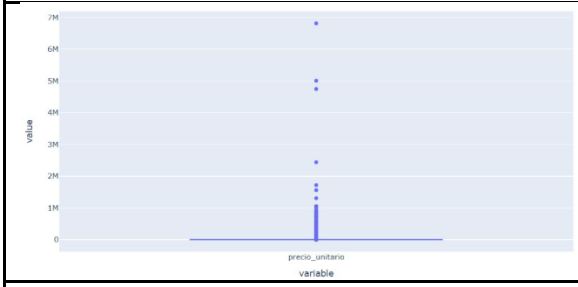
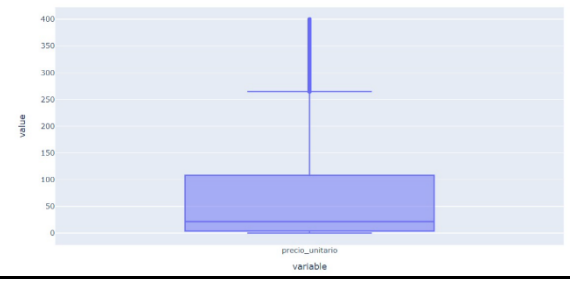

Dendograma descripcion - hclust
Height
0.000
0.002
0.004
0.006
0.008
fibras
sinteticas
mujer
tejido
poliester
algodon
punto
bebe
niño
niña
caballero
hombre
camiseta
dama
para
pantalon

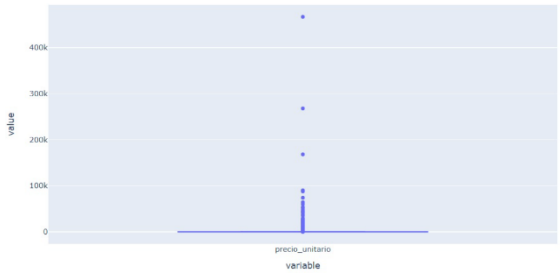
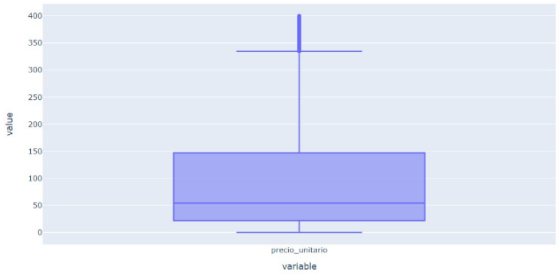
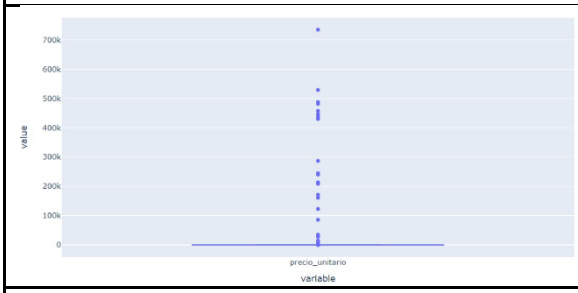
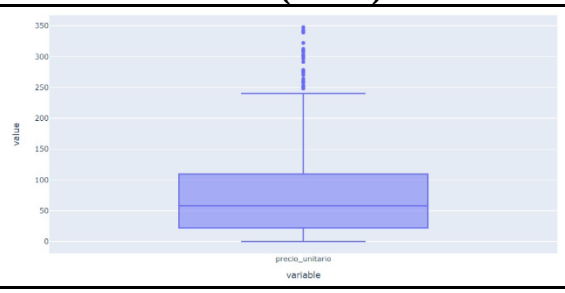
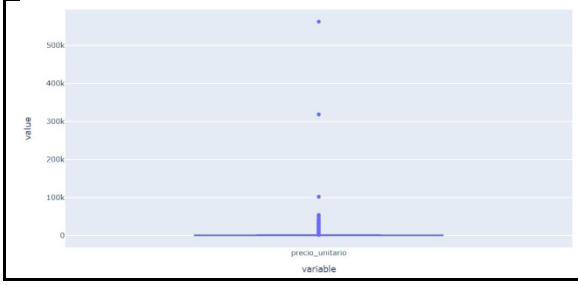
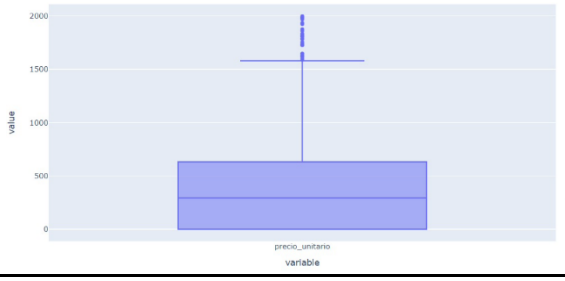

Dendrograma de descripcion - Agnes
Height
0.001
0.002
0.003
0.004
0.005
fibras
sinteticas
poliester
algodon
punto
para
pantalon
dama
camiseta
hombre
mujer
tejido
caballero
niña
bebe
niño



Análisis del precio unitario

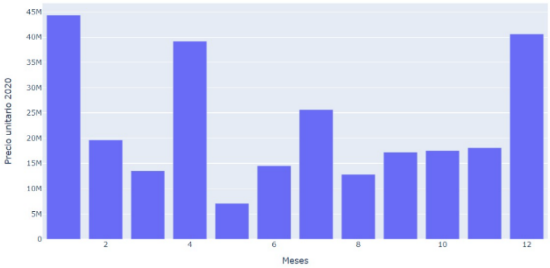
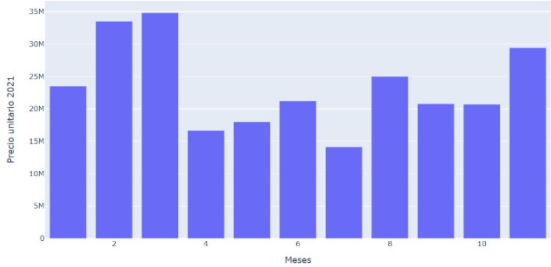
Finalmente, se realizaron diagramas de caja y bigotes para poder visualizar el precio unitario y su distribución para cada país de los que consideramos más importantes. Como nota, la primera gráfica muestra todos los precios, mientras que la segunda hace un zum para poder visualizar de mejor manera su distribución.

China	China (Zoom)
	
USA	**USA (Zoom)**
	
México	**México (Zoom)**

	
Brasil	**Brasil (Zoom)**
	
UK	**UK (Zoom)**
	

De igual manera, obtuvimos el precio promedio de algunos productos y las exportaciones totales registradas por año. Por ejemplo, el precio promedio de hilo de poliéster es de 159 y el precio promedio de batas de poliéster es de 501, con estos promedios podríamos ver si existen valores atípicos. Las exportaciones del 2020 fueron 512,294 y del 2021 fueron 508,788.

Finalmente, obtuvimos el movimiento de dinero por mes para ambos años, es decir, sumamos todos los precios unitarios del año para así obtener un total mensual, estos fueron resultados.

2020	2021
	

Estas gráficas anteriores nos muestran que los meses en los que hay más movimiento de mercancía son principalmente los meses de inicio y de fin de año, durante este periodo de tiempo es cuando hay más movimiento de dinero. De igual manera, hay que considerar que el 2020 fue un año atípico debido al inicio de la pandemia, puede que esto cause que las gráficas difieran un poco entre sí.

Conclusiones

La base de datos necesitó diferentes modelos de limpieza y con herramientas distintas, desde librerías de Python hasta Excel, con todas estas herramientas logramos obtener una descripción más corta y concisa. Ya con la base un poco más limpia, pudimos hacer los análisis mostrados anteriormente, como los de los países, graficando con diagrama de caja y bigotes su precio unitario total (para visualizar la masa económica de exportaciones en general). Al analizar las palabras más frecuentes desde la base original (sin limpieza), podemos observar que algunas de ellas fueron pantalón, pantalones, calcetines, etc., esto nos puede servir tener una idea de las palabras que podemos relacionar en el diccionario (como pantalón y pantalones) así como ver que productos se mueven más. Pudimos observar mensualmente la masa monetaria comercializada en el 2020 y 2021, y el periodo de tiempo en el que se mueve más dinero en ambos años, son finales e inicios de ambos años, recordando que el 2020 fue un año especialmente atípico en el tema comercial, de manera global a causa de la pandemia. Con la base de datos más limpia, podremos realizar de mejor manera los análisis y procesos futuros en este proyecto.

Entrega 3

Introducción

En esta entrega, utilizando el diccionario realizado la entrega pasada, creamos nuevas variables (combinaciones de palabras a partir de las palabras originales) para analizarlas con el precio y poder obtener conclusiones que enriquezcan el proyecto.

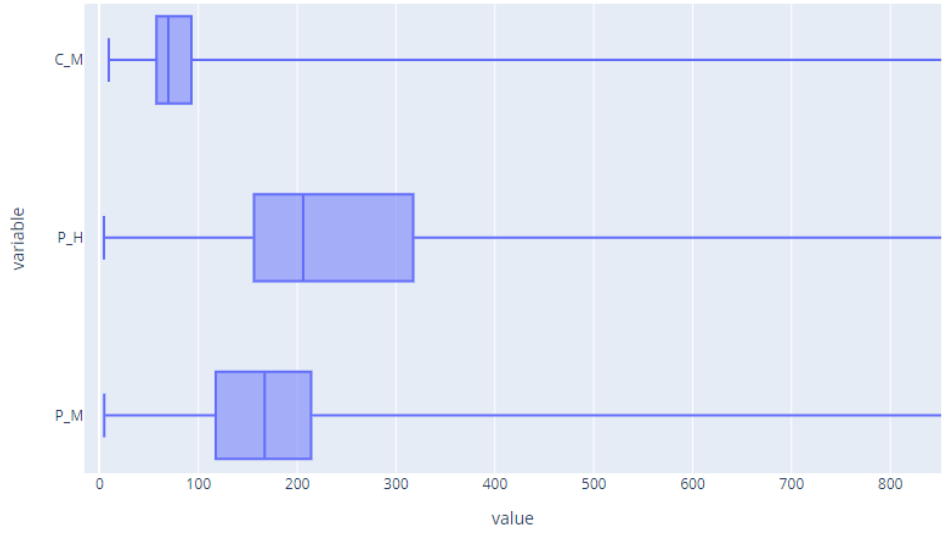
Procedimiento

Teniendo el análisis de pares de palabras y el dendograma, realizado en la entrega pasada, seleccionamos las palabras que nos parecieron de mayor relevancia en cuanto a frecuencia y relación con ayuda de dichos resultados. Con esto se creó una especie de filtro en el que aplicamos el diccionario, de las palabras seleccionadas para que en una sola palabra podamos realizar un análisis con la menor pérdida de información posible, por ejemplo, pantalón abarca: pantalón, pantalones, patalón, etc. Dicha palabra la relacionamos con dama y hombre para observar cómo se comporta el precio en individual y en conjunto. Realizamos el mismo procedimiento con camiseta y mujer, abarcando un total de 3 relaciones o variables creadas.

Resultados

Palabras en individual

Realizamos el análisis de precio en relación con las palabras en individual seleccionadas, para después realizar comparaciones con los pares de palabras creados.

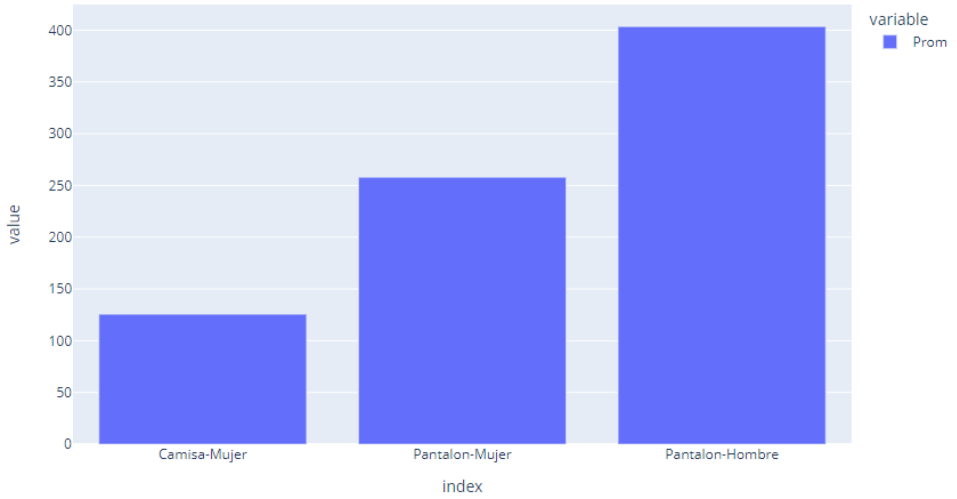


Análisis en pareja de palabras

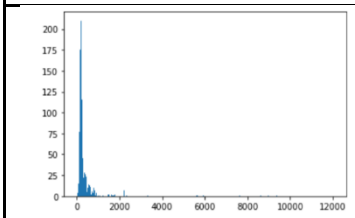
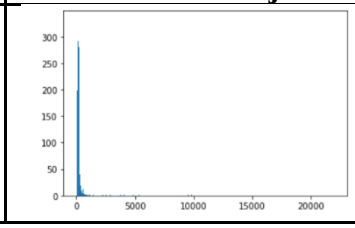
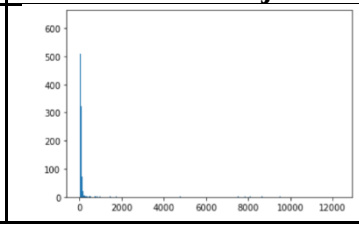
Posteriormente, realizamos el análisis del precio en comparación con los pares de palabras, como por ejemplo pantalón dama, para visualizar la información y sacar conclusiones. Los pares de palabras elegidos fueron:

- Camisa mujer
- Pantalón mujer
- Pantalón hombre

La siguiente gráfica muestra los precios promedio de estos pares de palabras.



Finalmente, graficamos un histograma con los precios de los pares de palabras seleccionadas, para realizar un último análisis de los productos seleccionados, estos fueron los resultados.

Pantalón hombre	Pantalón mujer	Camisa mujer
		

Conclusiones

Pudimos observar, que la prenda de hombre, en este caso el pantalón, es más costoso que los de mujer, lo que puede dar a entender es que, como estrategia de mercado, la ropa de hombre es más cara al ser una constante la que los hombres compran ropa con menor

frecuencia. Pudimos observar que los precios altos (*outliers*) están presentes en cada variable, lo que puede estorbar un poco en el análisis, por lo que los omitimos en los diagramas de caja y bigotes para una mejor visualización de la información.

Entrega 4

Introducción

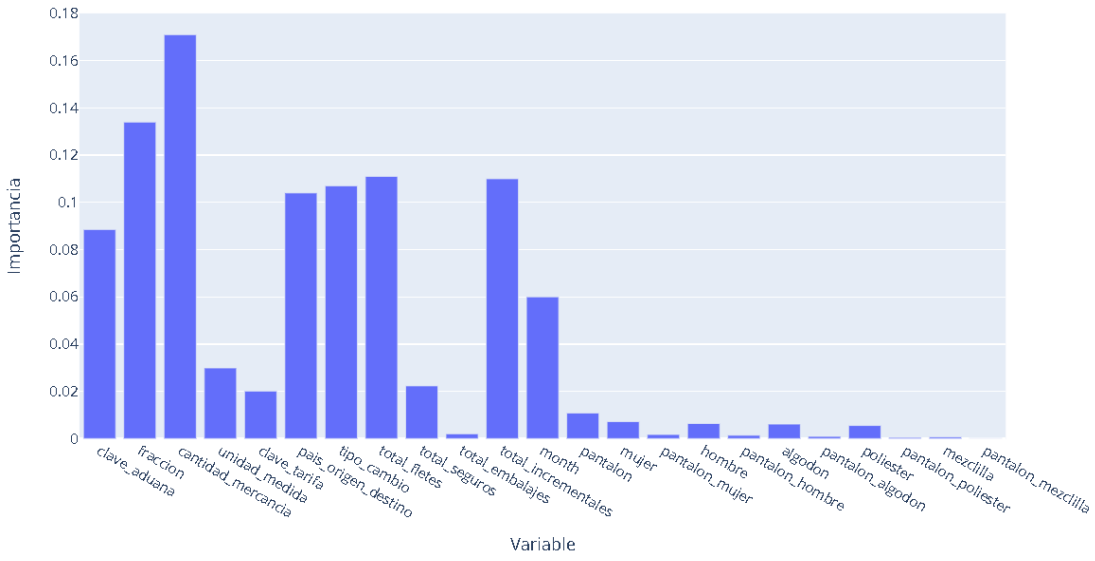
En esta semana lo que se hizo fue entrenar un modelo de clasificación por medio de la base de datos que trabajamos en entregas anteriores, esto para que en un futuro se pueda implementar. Se realizó el entrenamiento de un modelo clasificatorio con herramientas de ciencias de datos que podrá notificar en caso de precios atípicos en la mercancía.

Procedimiento

Partiendo de la base obtenida en entregas anteriores, lo que se hizo fue crear una nueva columna para nuestro DataFrame, dicha columna tiene el objetivo de informar si el producto analizado tiene un precio atípico o no, dicho análisis se realizó por fracción comparando productos de características similares, esto dado que no es posible comparar precios de productos distintos. Más adelante, se hizo una selección aleatoria de los datos para lograr un correcto balance y evitar que el modelo sea erróneo, para dicho análisis se tomaron 10,000 muestras de atípicos y 10,000 típicos. Posteriormente, se normalizó la base y se buscaron los hiperparámetros en ambos modelos (SVM y *Random forest*), al no tener éxito con SVM, optamos por usar el segundo modelo y agregar más variables de entrada que no se estaban considerando antes, a su vez dicho modelo ayudó para saber que columnas podían ser eliminadas, al no aportar información al modelo estas sólo metían ruido al mismo.

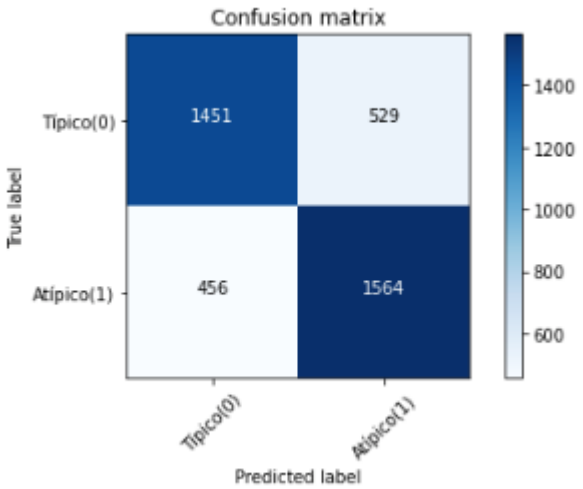
Resultados

Primero que nada, obtuvimos la siguiente importancia de los predictores seleccionados.



Teniendo en cuenta los resultados anteriores, decidimos hacer una comparación entre un modelo con todas las variables y un modelo eliminando las variables que no eran importantes o metían ruido al modelo. Al final obtuvimos que un modelo *Random Forest*, eliminando las variables no importantes, fue el mejor, esto fue lo obtenido al entrenar el modelo y considerando todo lo mencionado anteriormente.

Medidas datos de prueba *Random Forest*

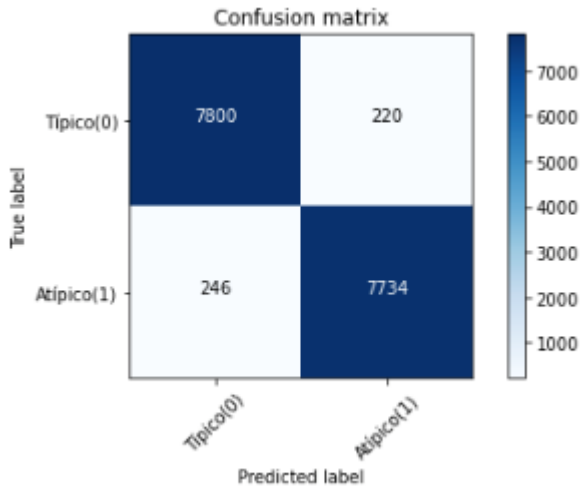


	precision	recall	f1-score	support
0	0.76	0.73	0.75	1980
1	0.75	0.77	0.76	2020
accuracy			0.75	4000
macro avg	0.75	0.75	0.75	4000
weighted avg	0.75	0.75	0.75	4000

Score_Random Forest test

Medidas	
Accuracy	0.753750
Precision	0.747253
Recall	0.774257
F1 score	0.753623

Medidas datos de entrenamiento *Random Forest*



	precision	recall	f1-score	support
0	0.97	0.97	0.97	8020
1	0.97	0.97	0.97	7980
accuracy			0.97	16000
macro avg	0.97	0.97	0.97	16000
weighted avg	0.97	0.97	0.97	16000

Medidas	Score_Random Forest train
Accuracy	0.970875
Precision	0.972341
Recall	0.969173
F1 score	0.970875

Conclusiones

Pudimos observar un modelo que clasificó satisfactoriamente los precios atípicos y típicos teniendo como resultado métricas lo suficientemente ajustadas en base a las iteraciones previas buscando el modelo indicado, cabe mencionar la importancia que tuvo el normalizar (escalar) los datos dado que al no hacer esto, simplemente no nos acercábamos a un buen modelo, a su vez eliminamos la mayoría de las columnas creadas previamente por medio del análisis del texto, esto diciéndonos que no fueron la mejor elección.

1.6. Valoración de productos, resultados e impactos

En esta experiencia se nos permitió generar una solución a un problema presentado, mediante la aplicación de conocimientos adquiridos a lo largo de la carrera. Como primer producto se trabajó en un análisis de bases que permitió tener un primer acercamiento de la base, lo que permitido dar el primer paso a generar algunos *insights*.

Posteriormente generamos una limpieza de datos y un diccionario para lograr obtener descripciones limpias, para su posterior análisis. Al desarrollar esta ingeniería de características, nos encontramos con muchas palabras que se tiene relación entre sí, por tanto, se generó este análisis para ver la relación de pares de palabras

A partir de esto con toda la información obtenida, se buscó generar un modelo que permitiera a partir de la entrada de texto y del precio de la importación se pudiera determinar si el precio es un dato atípico

1.7. Bibliografía y otros recursos

McKinney, W. (2012). Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

Duncan M. McGreggor. (2015). Mastering Matplotlib. Packt Publishing.

Stefanie Molin. (2021). Hands-On Data Analysis with Pandas : A Python Data Science Handbook for Data Collection, Wrangling, Analysis, and Visualization, 2nd Edition: Vol. Second edition. Packt Publishing.

1.8. Anexos generales

Anexo 1. Productos realizados en el canal de Gitlab en el siguiente enlace: https://gitlab.com/dpmontoy1/pap-mmd-v2022/-/issues.

2. Productos

Desde del inicio del proyecto se optó por usar la herramienta de Gitlab para hacer más sencillo el compartir los avances entre nosotros y nuestra asesora, dicha herramienta contiene todo lo elaborado por parte del equipo, a continuación, se muestra el enlace en el que se contienen todos los archivos y resúmenes, siendo los del equipo 2 los pertenecientes a nuestra autoría. Enlace en la sección 1.8 en anexos generales, anexo 1.

3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

3.1 Sensibilización ante las realidades

Iván Arellano:

El trabajo de un analista de datos prevé mucha manipulación de datos la cual se enfrenta a diferentes bases de datos dependiendo la manera de recolectarla de cada entidad. Por lo tanto, es esencial, para nosotros como analistas de datos saber manejar los diferentes tipos de datos para poder hacerse un espacio dentro de la industria.

Al trabajar en este PAP pude reflexionar y trabajar en equipo para continuar desarrollando mis habilidades de trabajo en equipo y solución de problemas

Andrés Ramírez:

Este PAP me hizo darme cuenta del poco avance que se ha tenido en las plataformas del gobierno para poder analizar sus bases de datos, dicha información debería estar lista para ser manipulada o incluso ser automático, una vez que el usuario introduce la información debería darse un informe detallado de lo que buscamos, si bien es entendible el que sea necesaria la intervención ocasional de algún programador, creo que hay ciertos aspectos que podrían mejorar, como tal esto fue frustrante dado que no sólo se ha visto este tipo de ineficiencia en el área que nos concierne sino en otras también como el área de la salud, estamos muy atrasados respecto a otros países que incluso llevan juicios en línea y lo hacen menos duración, por ejemplo.

Juan Pablo Rodríguez:

Me doy cuenta, que al requerir tanto trabajo las bases de datos para su limpieza, me hace sospechar que el gobierno no hace una limpieza apropiada sobre cada base de datos que se posee, por lo que realizan análisis que pueden ser deficientes, dando así una oportunidad laboral nueva para científicos de datos, porque requiere de ética de trabajo realizar una buena limpieza previo al modelaje por hacer.

Pablo Rivera:

Este PAP me hizo reflexionar acerca de los diferentes problemas que se pueden resolver mediante herramientas de *Machine Learning*, en este caso, aplicamos nuestros conocimientos como equipo para poder detectar precios atípicos y posiblemente actividades ilícitas en territorio mexicano. Anteriormente, no me imaginaba muy bien cómo aplicar herramientas de ML para resolver problemas reales, sin embargo, este ejemplo me sirvió para ver cómo aplicar estas herramientas para resolver problemas importantes en la sociedad y, gracias a lo que hemos aprendido en la carrera, poner de nuestra parte y buscar el beneficio de la mayor cantidad de personas posibles.

3.2 Aprendizajes logrados

Iván Arellano:

Como En este PAP se me permitió desarrollar mis aprendizajes de la carrera. En forma práctica y para un proyecto real. En el cual se me permitió trabajar en equipo para desarrollar un modelo de predicción de texto para observar precios atípicos.

El principal reto personal que viví dentro de este PAP fue que nunca había realizado un modelo que trabajara con texto, así que al inicio fue un proceso de mucha investigación y preguntas. Y poco a poco después de ir haciendo varias pruebas y con apoyo de nuestra consultora, se logró generar un modelo muy bueno.

Andrés Ramírez:

Durante el desarrollo del proyecto tuve la oportunidad de llevar a cabo diversos tipos de análisis, tanto numéricos como de texto, en dichos análisis logramos aplicar los conocimientos de la carrera en algo práctico y dejar de lado lo teórico. A su vez, logré trabajar en equipo con personas de la carrera, si bien ya los conocía creo que fue una experiencia muy enriquecedora. En cuanto a los problemas que se nos presentaron, si bien fueron pocos, hubo uno casi al final cuando diseñamos el modelo y que se solucionó simplemente llevando a cabo varios tipos de análisis y cambiando técnicas para la manipulación de la base que no habíamos usado. También veo importante el destacar que esta es mi primera experiencia

aplicando conocimientos de la carrera en algo externo a la universidad y se llevó a cabo con éxito.

Juan Pablo Rodríguez:

Este PAP me ayudó a practicar más las habilidades aprendidas a lo largo de la carrera, relacionadas con la programación y ciencia de datos, además de la organización para trabajar en equipo en un periodo de tiempo limitado. Representó un reto ya que había trabajado muy pocas veces en la limpieza de datos y de adaptación de modelos clasificatorios, que los vi para una sola clase en el quinto semestre de la carrera, por lo que requirió de mucha investigación en conjunto, prueba y error hasta lograr obtener resultados que nos convencieran. Me dio un panorama de las oportunidades laborales que puede brindar la ciencia de datos aplicados a prácticamente cualquier base datos, desde el gobierno como era el caso de esta base, a claramente las empresas que necesiten información relevante sacada de sus datos.

Pablo Rivera:

Al ya haber tomado este PAP anteriormente, logré reforzar mis conocimientos en cuanto al análisis de una base de datos sucia, desde la limpieza de los datos hasta la creación del modelo de IA final. Algo que diferenció este PAP al anterior fue el hecho de trabajar con texto, es decir, todo el análisis y modelado fue con base a una descripción (texto) lo cual fue algo nuevo para mí y me aportó nuevas herramientas para trabajar y manipular dicho texto. Asimismo, esto presentó la mayor parte de los problemas que encontré durante el PAP, usualmente al programar estoy acostumbrado a trabajar con datos numéricos, por lo que trabajar con texto requirió una mayor exigencia en cuanto a mis habilidades de programar y de buscar cómo obtener los resultados deseados a través de búsquedas en internet y apoyo con mis compañeros de equipo. En pocas palabras, reforcé habilidades que ya tenía, mejoré mi capacidad de trabajar en equipo y adquirí nuevas habilidades en programación y manipulación de texto.