INFORME DE ANÁLISIS DE ANOMALÍAS EN DATOS PAGO DE IMPUESTOS TRIMESTRE ENERO-MARZO, 2020

OBJETIVO

El presente informe tiene como objetivo realizar un análisis estadístico de los datos asociados al pago de impuestos por concepto de actividades económicas en la ciudad de Maracaibo, durante los meses de enero, febrero y marzo del presente año, para la detección de datos atípicos o anomalías que pudieran existir en dichos datos.

PLANTEAMIENTO DEL PROBLEMA Y ABORDAJE DE LA SOLUCIÓN

Todo proceso de optimización de gestión pasa por la integridad y control de los datos. De allí que, para realizar un mejor seguimiento y optimizar la gestión de recaudación, se plantea analizar el comportamiento de los contribuyentes a través de los datos e información disponibles, específicamente, el monto declarado y pagado relacionado al impuesto por sus actividades económicas. El universo de contribuyentes es de cinco mil novecientos cincuenta y dos (5952) personas, entre Jurídicas, gubernamentales y naturales.

Un valor atípico es una observación con al menos una variable que tiene un valor inusual. Procederemos inicialmente con el análisis estadístico-descriptivo, para luego realizar un análisis más profundo de los datos atípicos utilizando técnicas de aprendizaje automático de detección de anomalías.

ANÁLISIS DESCRIPTIVO

El análisis estadístico básico a los montos de pago de impuestos (en millones de Bs), arrojó los siguientes resultados:

Mínimo: 0.00591

Máximo: 15497.78

Varianza: 97273.7

Desviación estándar: 311.89

Media: 3

Promedio: 27.3

Asimetría (Skewness): 41.07

Curtosis (Kurtosis): 1936.08

La distribución de los montos se puede observar en el siguiente gráfico (Fig. 1).

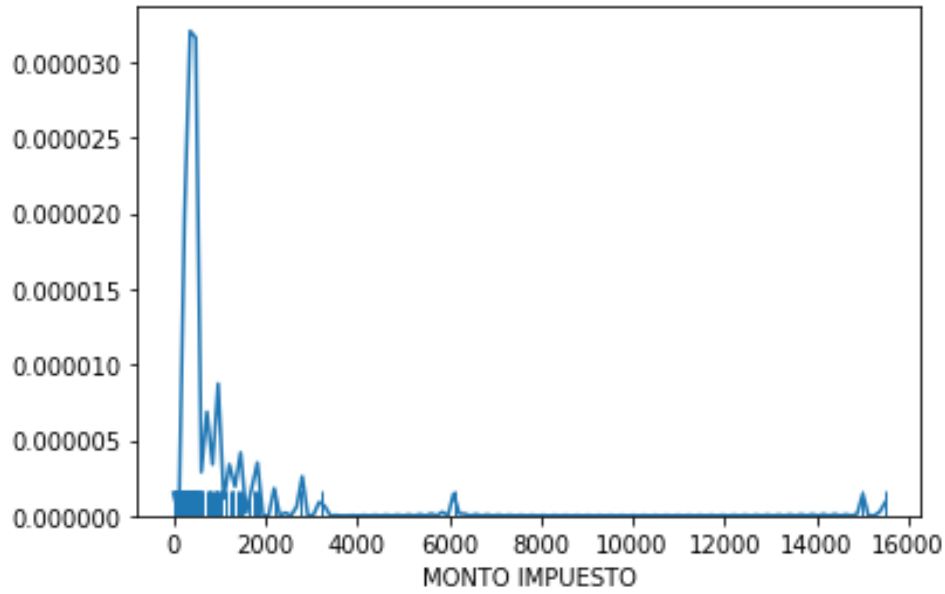

Fig. 1. Gráfico de distribución de los montos pagados en MMBs

Como puede observarse, la distribución de los datos no es normal (Gaussiana), tiene una asimetría acentuadamente positiva, por lo que la mayoría de los datos se encuentran en el rango de cero hasta dos mil (2000). Otra forma de visualizar la distribución de los datos es a través de la "función de distribución acumulativa empírica" (ECDF en inglés), mostrada en la Fig. 2, con los percentiles 2.5%, 25%, 50%, 75%, 97.5% (diamantes rojos).

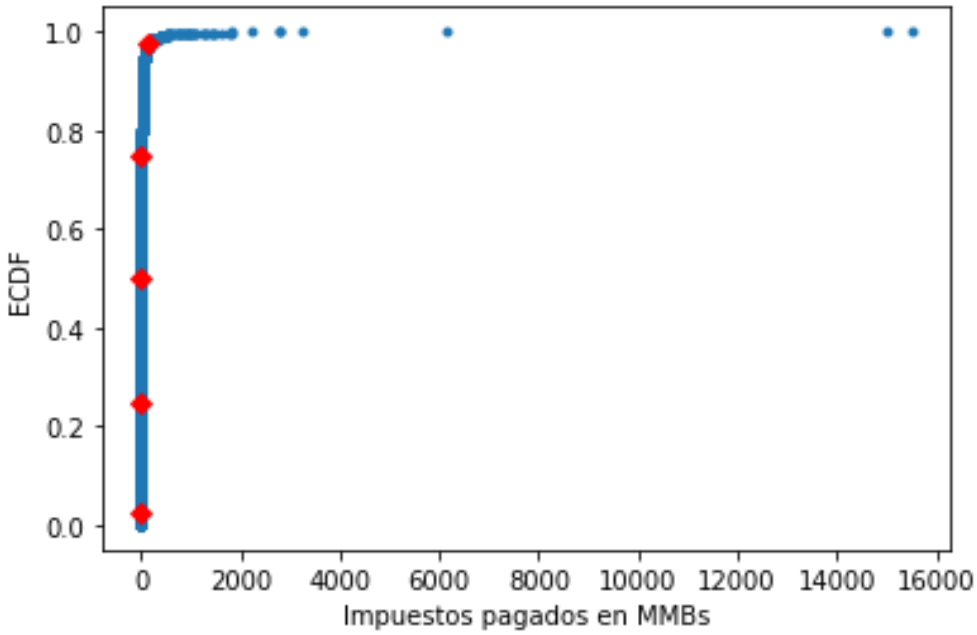

Fig. 2. Función de distribución acumulativa empírica

En la Fig 2 se pueden observar fácilmente los primeros y más evidentes datos atípicos. Estos corresponden a los contribuyentes con una marcada diferencia, hacia arriba, en el monto declarado y pagado, acentuándose desde dos mil millones en adelante.

Seguidamente, y con la finalidad de evaluar otra variable, observaremos la variación de los datos respecto a la desviación estándar, para ello utilizaremos un gráfico de dispersión (Fig. 3).

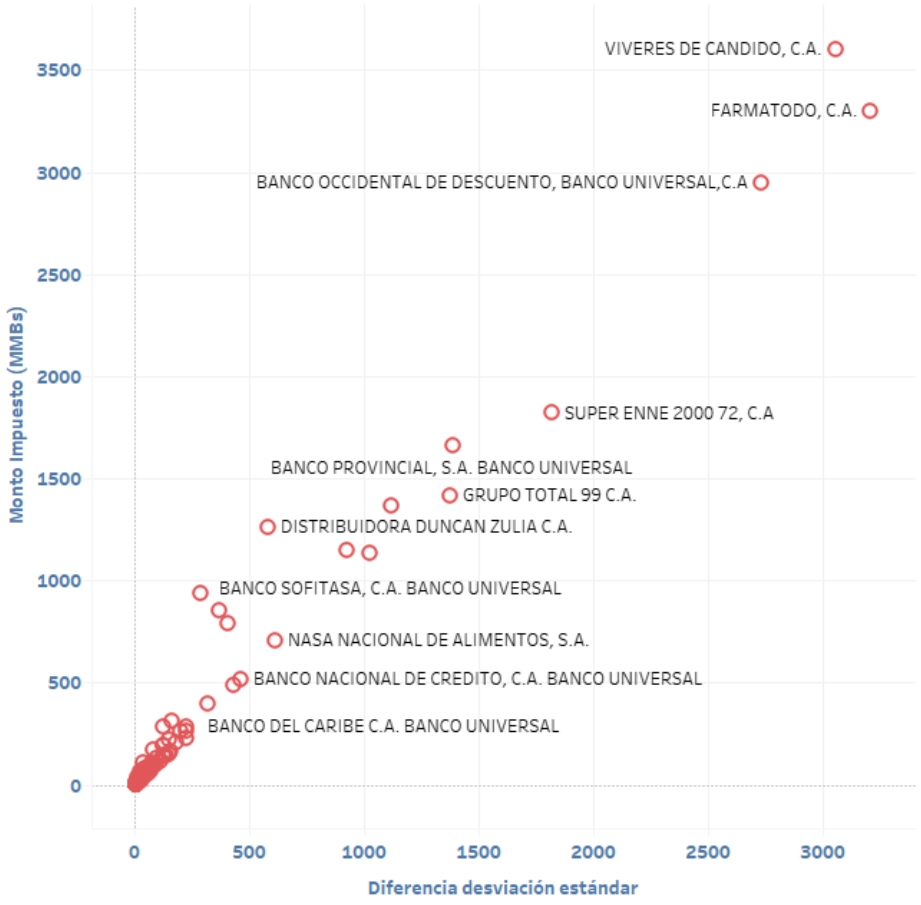

Fig. 3. Relación entre monto pagado y la desviación estándar

El siguiente análisis evaluará el comportamiento de los datos atípicos al comparar el monto pagado con la cantidad de sucursales o filiales. Esto con la finalidad de extraer algún patrón en caso de encontrarse. Obsérvese que se distinguen claramente cuatro (4) grupos o clúster, resaltados en distintos colores, de acuerdo a las características comunes de estos contribuyentes. Se incluye la empresa “PEPSI-COLA VENEZUELA C.A”, aunque no es un “dato atípico” evidente, sin embargo, su bajo monto pagado debería ser considerado para siguientes evaluaciones (Ver Fig. 4).

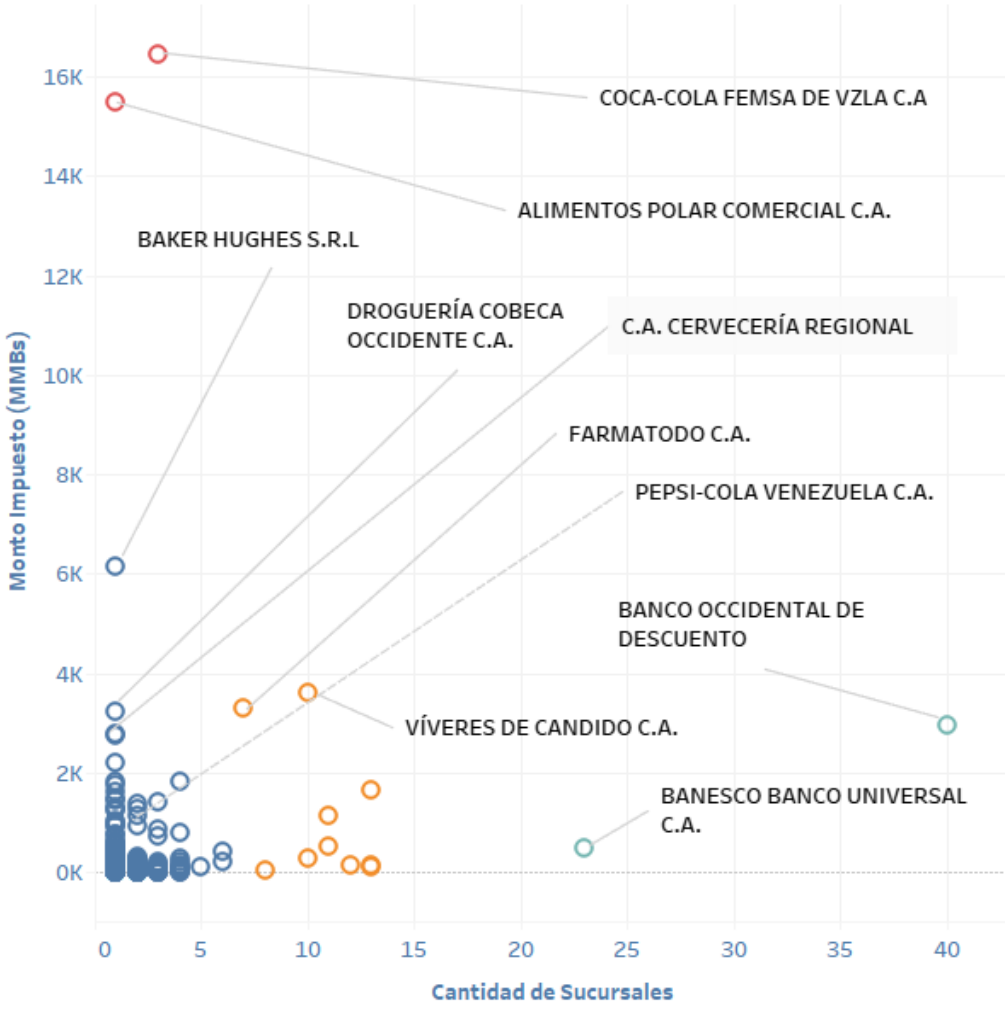

Fig. 4. Relación entre monto pagado y la cantidad de sucursales

Por último, realizaremos la estimación de potenciales datos anómalos utilizando la técnica de aprendizaje automático conocida como "metodología de envoltura elíptica" (Elliptic envelope methodology). Esta metodología asume que los datos provienen de distribuciones conocidas, tal como lo definimos anteriormente, en el análisis descriptivo. De esta manera, una vez definida la forma de los datos, podremos definir las observaciones periféricas como observaciones que se encuentran lo suficientemente lejos de la forma ajustada. Como resultado, obtendremos los "índices" de los registros potencialmente anómalos (desviados). Para nuestro caso, y dado que la metodología lo requiere, definimos (arbitrariamente) un porcentaje de contaminación (datos anómalos) del 5% de toda la data, con lo que, al final obtendremos 297 registros, tal como puede observarse en la siguiente previsualización de los datos (Fig. 5).

	RIF	RIM	CONTRIBUYENTE	ACTIVIDAD	MONTO LIQUIDADO PAGADO	MONTO IMPUESTO	ALICUOTA	ALICUOTA_PCT	MONTO INGRESOS
52	J-300619460	2000021932	BANCO OCCIDENTAL DE DESCUENTO, BANCO UNIVERSAL...	Bancos, empresas de seguros y reaseguros, casa...	8.795181e+07	87.951813	6.000000	0.060000	1.465864e+05
66	J-070558393	2000080707	SUPER ENNE 2000 72, C.A	Cadenas de supermercados, hipermercados, megat...	4.502377e+08	450.237744	1.500000	0.015000	3.001585e+06
70	J-000389233	2000080007	SEGUROS CARACAS DE LIBERTY MUTUAL C.A.	Bancos, empresas de seguros y reaseguros, casa...	2.207416e+09	2207.415541	6.000000	0.060000	3.679026e+06
165	J-406327840	2900050168	ALIMENTOS COMERCIAL LA GRAN FORTUNA, C.A	Abastos, bodegas y pequeños detales de víveres.	1.124796e+08	112.479598	2.000000	0.020000	5.623980e+05
181	J-308381233	2900013526	ZAPATERIA GASOLINA EXTRA C.A. (SUCURSAL)	Distribución y venta de calzados, carteras y o...	7.539444e+07	75.394438	3.000000	0.030000	2.513148e+05
...	...	...	...	...	...	...	...	...	...
5869	J-403194181	2900043577	DISTRIBUIDORA SANCHEZ ORDOÑEZ, C.A	Cosméticos, perfumes y artículos de tocador.	6.902885e+07	69.028849	3.000000	0.030000	2.300962e+05
5874	J-302938929	2000813225	DA VINCI BARRA RISTORANTE C.A.	Restaurantes, fuentes de soda, pizzerías, hela...	1.503091e+08	150.309071	2.000000	0.020000	7.515454e+05
5882	J-309233831	207P000709	MARIU INVERSIONES C.A.	Distribución de pinturas, lacas, barnices y ma...	1.050161e+08	105.016079	2.817895	0.028179	3.726757e+05
5899	J-301609573	2000813875	ABADIA DE LAS MERCEDES C.A	Agencias Funerarias y Capillas Velatorias.	2.174781e+08	217.478131	2.817895	0.028179	7.717752e+05
5910	J-411496812	7000000873	NST STORE MARACAIBO, C.A	Ventas de electrodomésticos.	1.541426e+08	154.142645	2.000000	0.020000	7.707132e+05

297 rows × 9 columns

Fig. 5. Previsualización del resultado de la predicción de datos anómalos con método Elliptic envelope

CONCLUSIONES

Con los resultados que hemos obtenido, se ha logrado visualizar de forma práctica aquellos casos que pueden ser considerados como “anomalías” y que, luego del debido análisis, ayudarán a orientar políticas y acciones que conlleven a la optimización del manejo de los datos, así como a una mejora en el seguimiento y control de gestión de recaudación. En vista del carácter arbitrario del método seleccionado para la predicción de anomalías, se entiende que estos resultados no deben ser tomados en absoluto como concluyentes, sino más bien como base para futuras evaluaciones.