Especialización en Analítica Estratégica de Datos

Modelo para la detección de Fraude de las IPS en la facturación de Bienes y Servicios a la EPS

Andrés Felipe Murillo Avendaño

Bogotá D.C., 2021

Tabla de contenido

Problema

Durante el año 2019 las EPS del país tuvieron unas pérdidas netas por 1,9 billones de pesos, mientras que el sector de las IPS tuvo una utilidad por 2,1 billones de pesos, según quinto informe de la de la Superintendencia Nacional de Salud[1]. Una de las posibles causas en esta diferencia de los actores, es el fraude por parte de las IPS a través de la facturación de productos y servicios que nunca se brindaron o elevando los costos de estos, entre otras causales.

Objetivo General

Reducir el fraude a través de un modelo basado en técnicas de minería de datos, para la prevención / detección del riesgo, a partir de la facturación presentada por las IPS ante las EPS.

Objetivos Específicos

1. Determinar patrones para la detección y prevención de fraude.
2. Crear un sistema de indicadores, para la detección de actividades sospechosas
3. Implementar los algoritmos de detección de patrones, para el aprendizaje del sistema

Justificación

El proyecto de análisis modelo para la detección de fraude de las Instituciones Prestadoras de Servicios de Salud - IPS en la facturación de cuentas médicas a la Entidad Promotora de Salud - EPS pretende identificar la mayor cantidad de escenarios en los que se pueda presentar un fraude al Sistema General de Seguridad Social en Salud (SGSSS). Según PricewaterhouseCoopers - PWC "el 39% de los entrevistados colombianos indicaron que sus empresas han sido víctimas de algún delito económico, lo cual se encuentra siete puntos por encima de los reportado en la Encuesta de 2016 (32%). Las cifras de los encuestados globales muestran un incremento de 13 puntos, pasando de 36% en 2016 a 49% en 2018"[2]. Ahora bien, en el sector salud se presentan casos de fraude por parte de los proveedores de la EPS los cuales van desde la presentación de facturas por pacientes inexistentes, suministro de medicamentos o prestación de servicios los cuales nunca fueron entregados a los pacientes o en una menor cuantía de lo que es recobrado en la factura, exámenes falsos o innecesarios, precios elevados de medicamentos frente a su valor de mercado, entre otras prácticas.

Por tal motivo este proyecto busca reconocer características propias del negocio de la EPS con las IPS, mediante análisis previos que identifican situaciones en las que se puede inferir presuntos fraudes, casos históricos detectados en sistemas de salud de otros países, creando de una serie de indicadores y junto con el apoyo de técnicas de minería de datos, crear algoritmos de detección de patrones de actividades sospechosas, como insumo para la EPS al momento de tomar decisiones frente a su política de prevención y detección del fraude.

---

[1] Forbes - En 2019 las pérdidas netas de las EPS aumentaron y llegaron a $1,9 billones, https://forbes.co/2020/07/22/actualidad/

[2] PWC - Fraude al descubierto Encuesta Global Crimen Económico 2018 Colombia – año 2018

Marco Teórico

El fraude se entiende como la intención deliberada por parte de un individuo a sabiendas de la falsedad o que no lo reconoce como verdadero y realiza una acción que le traerá algún tipo de beneficio con el engaño para él u otra persona. (Thorton, Meuller, Schoutsen, Hillersberg, 2013). Por tal motivo el desarrollo de un modelo que permita detectar las situación o comportamientos que se puedan encasillar como fraude dentro de la base de datos de facturación por parte de las IPS a las EPS, mediante técnicas de minería de datos, es necesario, porque a pesar de existir una auditoría a las facturas radicadas para su pago, sólo se puede verificar una muestra mínima debido al elevado número de facturas radicadas por las IPS, esto hace que la tarea manual de revisión sea casi imposible y demanda excesivos recursos de la EPS para este control. Por tal motivo basados en la literatura existente se procederá a desarrollar este proyecto de investigación adaptándolo a las características del Sistema General de Seguridad Social en Salud (SGSSS). Con este trabajo se pretende dar el primer paso hacia el siguiente salto, puesto que ya es posible procesar el aluvión de datos y analizarlo con las herramientas actuales.

De la literatura (Mesa, Ranieri, Maturana, Kaempffer - 2009) desarrollaron un modelo a través de regresión logística binomial, en el cual, de tan solo cuatro variables identificadas de los formularios de incapacidad médica lograron determinar en un 99.71% de los casos observados y posteriormente revisados cuales incapacidades eran realmente fraudulentas. Este modelo permite discriminar de una manera rápida y efectiva el fraude en el sistema de salud chileno, reduciendo el costo de auditorías mediante muestreo para la determinación de la validez de las incapacidades. También se tomó en cuenta para este proyecto (Thorton, Meuller, Schoutsen, Hillersberg, 2013) el cual desarrolla un modelo en el que se discrimina los diferentes niveles y actores en el sistema de salud norteamericano mediante el cual determinan que existen 7 niveles por los que debe atravesar una factura en la reclamación de pagos por gastos médicos, y contrastan esta información de cada nivel contra los 6 casos más comunes de fraudes detectados, como lo son la facturación fantasma, facturación duplicada, codificación, desagregación, excesivos o innecesarios tratamientos y sobornos. Este documento contribuye a entender que se pueden desarrollar modelos multidimensionales en los cuales se puede segmentar la información y de acuerdo con cada nivel se debe desarrollar el análisis más pertinente por cada nivel para la detección del fraude.

Palabras clave: Detección de Fraude, Análisis Predictivo, Facturación,

HISTORIA DE LOS DATOS

La información fue tomada de una base de datos de radicación de cuentas médicas de la EPS que nos suministró la información para poder realizar este proyecto de analítica, comprende el período de prestación de servicios del año 2020, este conjunto de datos consta de 123 columnas y 8.561.061 registros. Se descargó la base de datos e inicialmente se realizó un perfilamiento de datos para identificar las variables más representativas, posteriormente se tomó una muestra del 20 % del conjunto de datos original para un mejor manejo debido al volumen de información, obteniendo una base de 80 columnas y 1.730.042 registros.

Perfil de distribución de valores de columnas

Mediante la herramienta Microsoft Visual Studio se creó un proyecto de Integration Services, a través de la tarea de generación de perfiles de datos, con el fin de realizar un análisis, previo a la revisión a fondo del conjunto de datos, para identificar el comportamiento de la información (Campos nulos, cantidad de caracteres por columna, cantidad de registros únicos por columna, etc.). Se configura la tarea de generación de perfiles de datos, se especifica la fuente del conjunto de datos y los perfiles que se quieren aplicar para cada campo.

Este perfilamiento permite identificar la cantidad de valores únicos por columna o por campo del conjunto de datos, para el set de datos validado se evidencian columnas con cero o un único dato distintivo.

Ilustración 2 – Perfil de distribución por columna

Perfiles de distribución de valores de columna - [RVG].[setProyecto]

Columna	Número de valores distintos
proEsRecobrable	0
esUltimoBD	0
idCohorte	0
cmdgeneradaseven	1
OrigenBDUA	1
excluir	1
dxhomIDTipDiag	1
Serv_Padre	2
prsVistoBueno	2
prsgrabada	2
dprEsRecobrable	2
CuentaEvento	2
bd	2
Cobertura	2
FAC_ESTA	2
idCobertura	2
esContratoEspecial	2
NombreCobertuta	2

Fuente: Elaboración propia

Perfil de proporción de columnas sin información

Identifica el porcentaje de valores nulos o NA que posee una columna, para el set de datos validado se evidencian columnas con el 100% de valores nulos.

Ilustración 4 – Perfil proporción por columna

Perfiles de proporción de columnas NULL - [RVG].[setProyecto]

Columna	Recuento de NULL	Porcentaje de NULL
idCohorte	8561061	100.0000 %
proEsRecobrable	8561061	100.0000 %
esUltimoBD	8561061	100.0000 %
excluir	8559874	99.9861 %
cafRadOtroOperador	8508028	99.3805 %
idConceptoGralEspecifico	8403414	98.1586 %
ValorTotalGlosadoModulo	6592635	77.0072 %
ValorPorConciliar	6592635	77.0072 %
ValorAceptadoIPS	6592635	77.0072 %
ValorAceptadoEPS	6592635	77.0072 %

Fuente: Elaboración propia

Perfil de distribución de longitud de columnas

Identifica las columnas más extensas en cuanto a la cantidad de caracteres, para el set de datos validado se evidencia que son las descripciones de servicios o medicamentos.

Perfiles de distribución de longitud de columnas - [RVG].[setProyecto]

Columna	Longitud mínima	Longitud máxima	Omitir espacios iniciales	Omitir espacios finales
DescripcionHeonServicio	4	1461	☐	☑
Descripcion_Cups_Cum	4	955	☐	☑
dxDescripcion	3	215	☐	☑
Agrupador2NTP	5	128	☐	☑
Cohorte	6	111	☐	☑
SubCohorte	2	106	☐	☑
Seccion	2	95	☐	☑
FAC_DESC	40	64	☐	☑
Especialidad	8	51	☐	☑
Servicio	7	28	☐	☑

Fuente: Elaboración propia

Proceso de Calidad de Datos

Se recibieron los datos de la EPS y se procedió a importar a SQL Server Express, visualizando el contenido de la información para verificar su consistencia y compararla contra otros criterios detectados durante el proceso de examinación preliminar. Por ejemplo, un tipo de dato para una columna puede contener valores nulos, pero esto estaría permitido ya que los valores nulos son permitidos durante fases iniciales de captura de los datos; una columna podría contener datos referentes a otros contenidos en otra tabla; en este caso no podría contener valores que no estarían contenidos o contemplados dentro de esa tabla foránea. En los anexos se detalla el proceso de depuración y de calidad de los datos después de la importación.

Exploración de los Datos

Se validan la cantidad de datos en la base

```
[ ]  # Observamos la cantidad de datos que tiene la base
     df.count()
```

```
EPS                   1080328
cmdcuentamedica       1080328
prsIdAmbito           1080328
Cobertura             1080328
prsFchEntrada         1080328
                       ...
NombreTipoProducto    1057134
idEstadoCuenta        1080328
NombreEstadoCuenta    1080328
excluir                     8
idNit                 1080328
Length: 123, dtype: int64
```

Fuente: Elaboración propia

Se identifican los valores Nulos

```
dfSelect.isnull().sum()
```

```
ValorTotalDetalle        0
Edad                 12468
Genero               12468
idTipoProducto       23194
idEstadoCuenta           0
idAmbito                 0
prsTipAfiliado       17478
dprcantproducto          0
prsgrabada               0
prsVistoBueno            0
EPS                      0
dtype: int64
```

Fuente: Elaboración propia

Estadísticas Descriptivas:

Se comparan las cantidades de acuerdo con el producto

```
condition_max_rating = dfSelect['dprcantproducto'] == max_dfSelect
dfSelect_max = dfSelect[condition_max_rating]
dfSelect_max
```

	ValorTotalDetalle	Edad	Genero	idTipoProducto	idEstadoCuenta	idAmbito	prsTipAfiliado	dprcantproducto	prsgrabada	prsVistoBueno	EPS
99522	24928780	66.0	2.0	1.0	102	3	1.0	1083860	1	1	2

Fuente: Elaboración propia

A partir de lo anterior, es interesante mirar que hay un producto que tiene una cantidad extremadamente alta y esto eleva el valor del procedimiento a casi 25 millones.

```
dfSelect.groupby('idAmbito').mean()
```

idAmbito	ValorTotalDetalle	Edad	Genero	idTipoProducto	idEstadoCuenta	prsTipAfiliado	dprcantproducto	prsgrabada	prsVistoBueno	EPS
1	237420.330229	50.149436	1.526205	4.346970	89.327498	2.335840	19.967850	0.872454	0.873096	2.0
2	27540.274084	39.991464	1.531252	3.385614	85.762783	2.456796	5.992722	0.838317	0.838439	2.0
3	179978.859419	49.476980	1.499620	3.416618	81.741080	2.386611	163.078749	0.798458	0.798522	2.0
4	231379.871376	48.283813	1.570820	4.192440	93.534528	2.100373	9.213759	0.916183	0.916183	2.0

Fuente: Elaboración propia

Promedio de si el ámbito de la cuenta es ambulatorio, urgencia, hospitalario o domiciliario. Aquí se podría pensar que los ámbitos Hospitalarios y domiciliarios sean mayores porque implican cargos en hospitalización, medicamentos, y algunos cargos de transporte y atención domiciliaria podemos observar que el servicio Ambulatorio es más costoso que el Hospitalario, habría que observar con más detenimiento si se presentan datos fraudulentos en algún cobro.

```
dfSelect.groupby('prsgrabada').mean()
```

prsgrabada	ValorTotalDetalle	Edad	Genero	idTipoProducto	idEstadoCuenta	idAmbito	prsTipAfiliado	dprcantproducto	prsVistoBueno	EPS
0	194205.055200	49.596955	1.453804	3.161394	1.571899	2.673313	2.238884	200.974959	0.000685	2.0
1	137729.939178	46.714138	1.522130	3.557659	101.998979	2.572991	2.433201	92.245412	1.000000	2.0

Fuente: Elaboración propia

Promedio de la cuenta si fue o no grabada para el paso al sistema contable. Podrían ser interesantes las cuentas que no fueron grabadas y si estas tienen un mayor valor del servicio.

```
dfSelect.groupby('prsVistoBueno').mean()
```

prsVistoBueno	ValorTotalDetalle	Edad	Genero	idTipoProducto	idEstadoCuenta	idAmbito	prsTipAfiliado	dprcantproducto	prsgrabada	EPS
0	194238.329592	49.609185	1.453784	3.160478	1.52329	2.673839	2.238515	201.111721	0.000000	2.0
1	137731.170873	46.711898	1.522124	3.557782	101.99439	2.572887	2.433258	92.231273	0.999844	2.0

Fuente: Elaboración propia

Promedio de la cuenta médica si tiene o no visto bueno en la auditoría. Por ende, las cuentas que no tienen el visto bueno podrían tener facturas que son fraudulentas.

Promedio por género

```
#promedio por genero
dfSelect.groupby('Genero_1.0').mean()
```

Genero_1.0	ValorTotalDetalle	Edad	dxIDDiagnostico	idTipoProducto	idEstadoCuenta	Genero_2.0	Genero_3.0
0	238146.075621	47.693564	10155.279757	3.128232	85.126078	0.953544	0.000007
1	142596.842228	52.017119	9702.320873	3.142856	85.395455	0.000000	0.000000

Fuente: Elaboración propia

Promedio por municipio prestador

```
#promedio   por departamento
dfSelect.groupby('MunicipioPrestador').mean()
```

MunicipioPrestador	ValorTotalDetalle	Edad	dxIDDiagnostico	idTipoProducto	idEstadoCuenta	ID	Genero_1.0	Genero_2.0	Genero_3.0
Acevedo	2.381777e+04	35.385753	11743.748353	3.643148	75.652174	5.496206e+06	0.504611	0.475626	0.0
Agrado	2.239070e+04	22.333333	12143.789474	3.446429	94.912281	5.669801e+06	0.421053	0.578947	0.0
Aguadas	1.034776e+06	48.571622	10720.078249	1.918367	67.840049	5.434267e+06	0.521368	0.382173	0.0
Aipe	1.938195e+04	23.428571	14961.238095	3.476190	102.000000	5.786158e+06	0.000000	1.000000	0.0
Alban	3.198550e+06	NaN	17131.000000	2.000000	102.000000	NaN	0.000000	0.000000	0.0
...	...	...	...	...	...	...	...	...	...
Yopal	1.114313e+05	35.800851	9890.144004	2.999002	14.256846	5.236244e+06	0.310670	0.244098	0.0
Yumbo	2.070743e+05	57.014179	9604.834138	2.723280	102.000000	5.424765e+06	0.483163	0.497804	0.0
Zapatoca	2.241893e+05	39.452434	11465.094875	3.159912	62.789531	5.338889e+06	0.428571	0.557252	0.0
Zarzal	9.432135e+04	22.478632	11233.435897	3.820513	102.000000	5.605266e+06	0.564103	0.435897	0.0
Zipaquira	2.442525e+05	34.496774	10386.588235	4.152941	93.088235	5.871757e+06	0.570588	0.341176	0.0

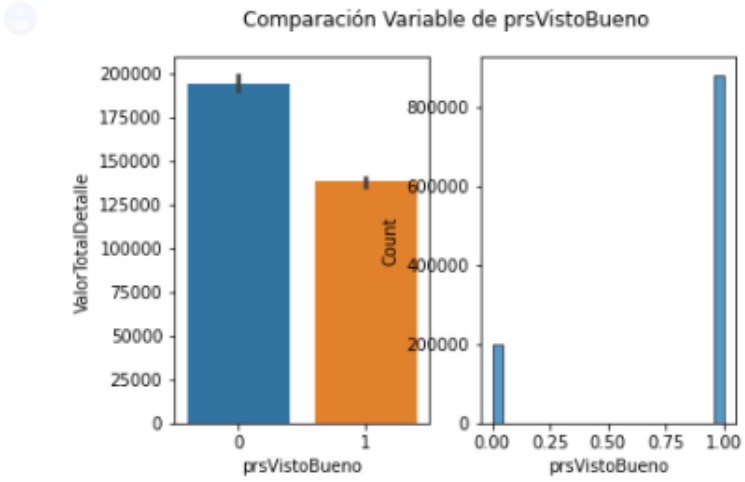
Fuente: Elaboración propia

Gráficas Descriptivas

```
[ ]  ax1 = plot.subplot(121)
     sns.barplot(x=dfSelect['prsVistoBueno'], y=dfSelect['ValorTotalDetalle'])

     ax2 = plot.subplot(122)
     sns.histplot(dfSelect['prsVistoBueno'])

     plot.suptitle('Comparación Variable de prsVistoBueno')
     plot.show()
```



Fuente: Elaboración propia

¿Por qué esta variable?

Se podría pensar que si las cuentas médicas no tienen el visto bueno en la auditoría es porque tienen inconsistencias en sus valores. Por ende, se podría buscar facturas fraudulentas que justamente no hayan tenido validez de la auditoría.

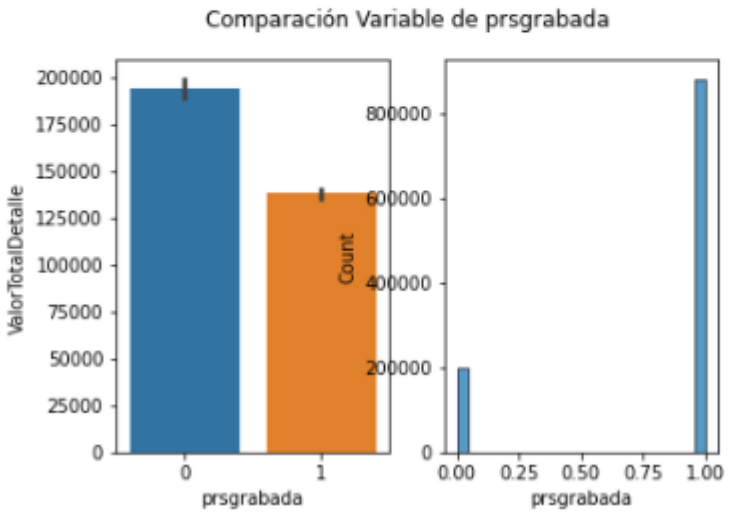
Interpretación

Observando las gráficas, se aprecia que en la de la derecha, hay una menor cantidad de facturas que no tienen visto bueno de la auditoría, sin embargo, estas cuentas tienen valores de pago del servicio mucho más alto que las que tienen visto bueno (gráfica izquierda). Entonces, se podría pensar que justamente estas facturas que no tienen visto bueno de la auditoría es porque son de una gran suma de dinero y, por ende, inconsistentes en de dónde salieron valores tan altos de la factura, lo que podrían ser facturas fraudulentas con una sobreestimación de los precios del servicio.

```
[ ]  ax1 = plot.subplot(121)
     sns.barplot(x=dfSelect['prsgrabada'], y=dfSelect['ValorTotalDetalle'])

     ax2 = plot.subplot(122)
     sns.histplot(dfSelect['prsgrabada'])

     plot.suptitle('Comparación Variable de prsgrabada')
     plot.show()
```



Fuente: Elaboración propia

¿Por qué esta variable?

Se podría pensar que si las cuentas médicas no son grabadas es porque están evadiendo estos pagos y hay cosas ocultas detrás de estas cuentas. Por ende, se podría buscar facturas fraudulentas que justamente no hayan sido grabadas.

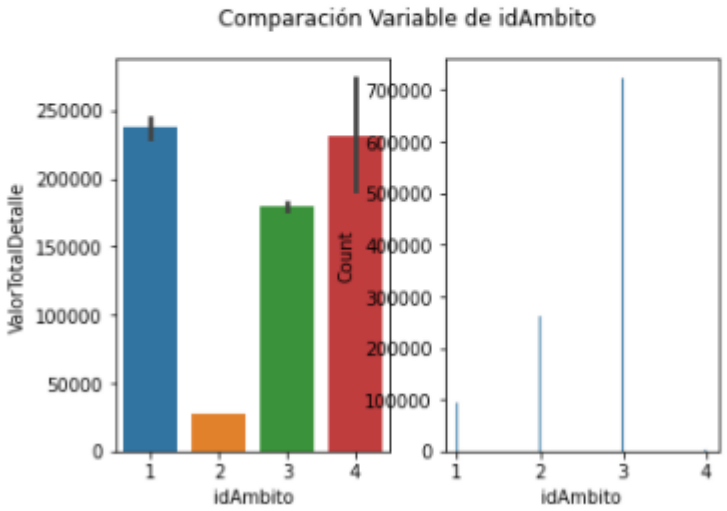
Interpretación

Observando las gráficas, se aprecia que en la de la derecha, hay una menor cantidad de facturas que no son gravadas, sin embargo, estas cuentas tiene valores de pago del servicio mucho más alto que las que sí son grabadas (gráfica izquierda). Entonces, se podría pensar que justamente estas facturas que no son grabadas esconden información en grandes sumas de dinero cobradas en sus facturas y, por ende, no graban estas para ocultar esta información, lo que podrían ser facturas fraudulentas con una sobreestimación de los precios del servicio

```
[ ]  ax1 = plot.subplot(121)
     sns.barplot(x=dfSelect['idAmbito'], y=dfSelect['ValorTotalDetalle'])

     ax2 = plot.subplot(122)
     sns.histplot(dfSelect['idAmbito'])

     plot.suptitle('Comparación Variable de idAmbito')
     plot.show()
```



Fuente: Elaboración propia

Se tiene el número de cuentas médicas que fueron ambulatorias (1), urgencias (2), hospitalaria (3) y domiciliaria (4).

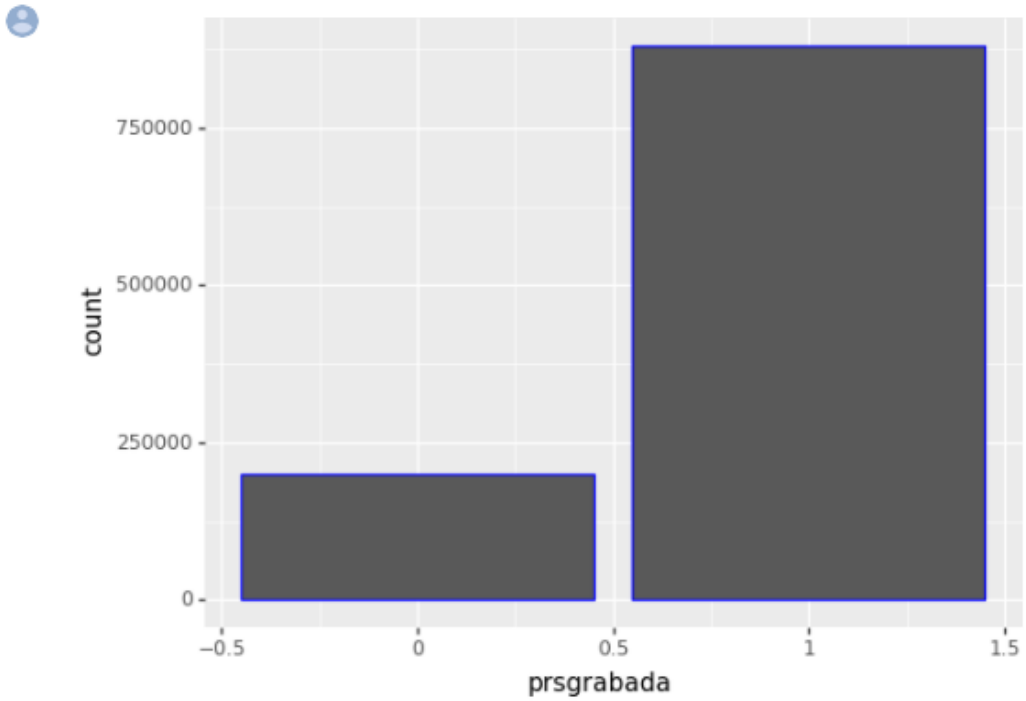
¿Por qué esta variable?

Se podría pensar que los tipos de cuentas que impliquen movilización a la casa o gastos dentro de varios días en el hospital como hospitalaria y domiciliaria, tienen valores más altos por los costos de movilizar los implementos, personal, alimentación durante la hospitalización etc. Por ende, se podría buscar en estos datos si efectivamente esto ocurre o si podrían haber facturas fraudulentas si se presentan valores grandes de otros tipos de cuenta.

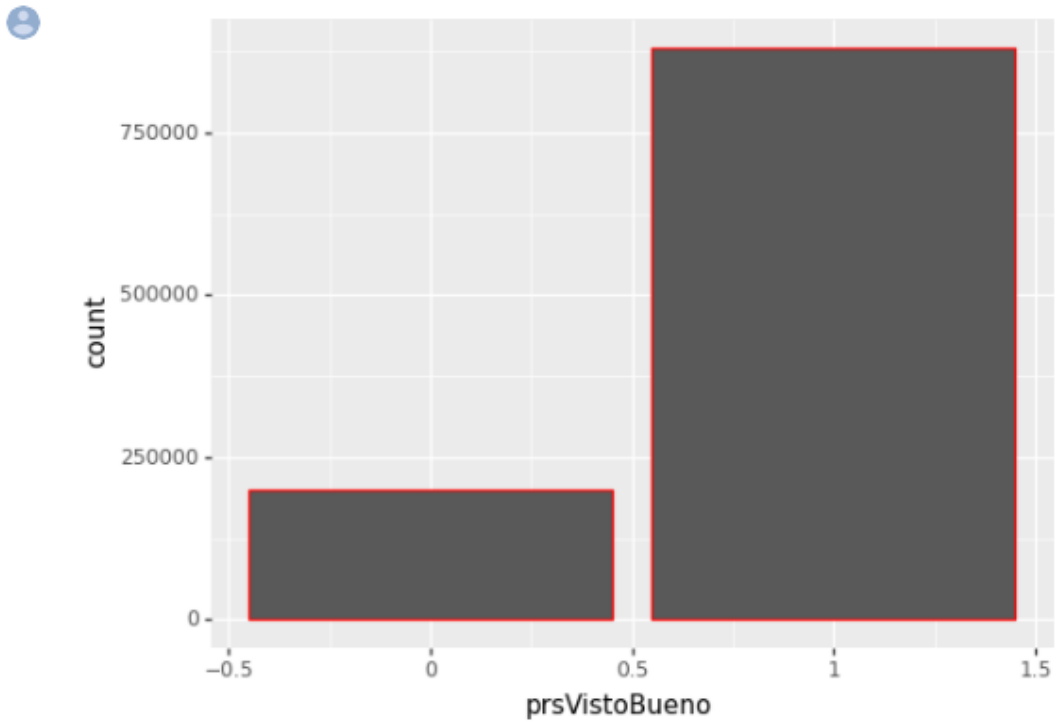
Interpretación

Observando las gráficas, se aprecia que en la de la derecha, hay una cantidad mínima de facturas que son de tipo domiciliaria, y un gran número de facturas de tipo hospitalario, lo cual hace mucho sentido porque las personas se movilizan más seguido a los centros médicos. Sin embargo, también se podría pensar que se sobreestiman esos gastos de movilización a las casas para generar más caja en los centros médicos. Lo anterior, podría resultar en facturas fraudulentas ya que al haber un gran costo en facturas tipo 4 domiciliario y baja cantidad de estas, entonces se podría pensar que hay una sobreestimación de los precios del servicio, de igual forma con las facturas de tipo ambulatorio que son más costosas que las de tipo hospitalario , habría que analizar más a fondo si esto se debe a algún tipo de medicamento o enfermedad.

PRSGRABADA VS PRSVISTOBUENO



Fuente: Elaboración propia



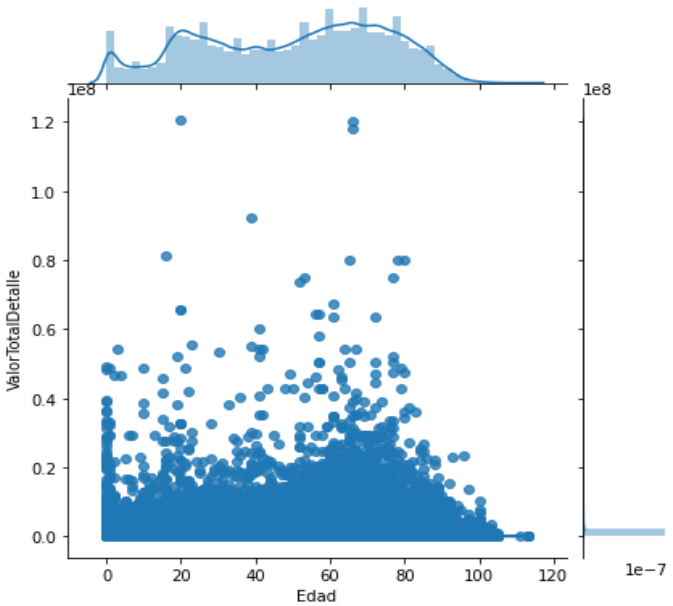

```
<ggplot: (8745751009141)>
```

Fuente: Elaboración propia

En los gráficos aparentemente se evidencia un comportamiento correcto entre las facturas Con Visto Bueno vs las Grabadas.

```
sns.jointplot(x='Edad', y='ValorTotalDetalle', data=dfSelect, kind='reg')
```

```
<seaborn.axisgrid.JointGrid at 0x22fb19129a0>
```

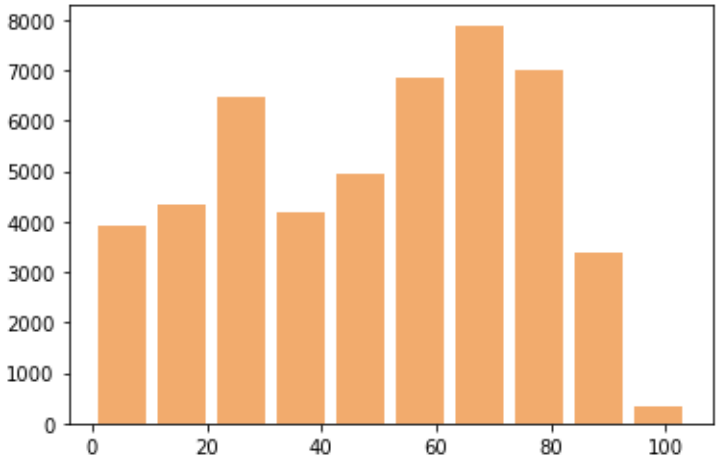


Fuente: Elaboración propia

Histograma por edades

```
plot.hist(x=dfSelect['Edad'], color='#F2AB6D', rwidth=0.8)
```

```
(array([3918., 4340., 6471., 4193., 4935., 6869., 7895., 7008., 3373.,
        346.]),
 array([  0. ,  10.4,  20.8,  31.2,  41.6,  52. ,  62.4,  72.8,  83.2,
        93.6, 104. ]),
 <a list of 10 Patch objects>)
```



Fuente: Elaboración propia

PRUEBAS PARA LA DETECCIÓN DEL FRAUDE

Para la detección de fraude en la base de datos de facturación se usarán 3 tipos de test que son utilizados en los procesos de auditoría forense, la cual "es la ciencia de recopilación y presentación de información financiera en una forma adecuada para una corte de jurisprudencia contra los autores de delitos económicos. De la integración de la contabilidad, la auditoría y las técnicas de investigación se obtiene la especialidad conocida como la contabilidad forense, que se centra en la detección o prevención de fraudes contables" (Vergara, 2017)[3], por tal motivo utilizaremos técnicas que permiten discernir sobre la veracidad de los datos presentados por las IPS a la EPS, para este caso se usarán las siguientes:

1. Test IGUAL-IGUAL-IGUAL

El propósito de este test es el de identificar duplicados dentro de la base de datos, los cuales se pueden analizar a fondo para comprobar si estos corresponden a errores o casos de fraude. Para este análisis es importante identificar las columnas dentro de la base de datos que permitirán encontrar si existen duplicados con los parámetros seleccionados, y con base en

[3] Detección de Fraudes en Bodegas de Datos basado en los niveles de agregación – Karen Vergara - 2017

los hallazgos. (Nigrini, 2011)[4]. Es un test bastante sencillo que busca encontrar duplicados exactos dentro de la base, no tiene un límite de campos a establecer para identificar la duplicidad.

Para nuestro caso de análisis se se buscaron tres campos, "idNit" este campo corresponde al numero de identificacion de la IPS, "PrNumFactu1" es el numero de la factura emitida por la IPS, "cmdTotalFactura" corresponde al valor monetario de la factura. Inicialmente se construyó un query para identificar duplicados dentro del conjunto de datos que cumplan con los tres campos antes mencionados. Se realizaron dos test, el primero se denominó SSS1, por sus siglas en inglés (SAME-SAME-SAME) el cual arrojó un total de 408 facturas sospechosas, por un valor total de $876.977.622,00 pesos M/CTE. De una revisión más detallada al conjunto de datos se pudo determinar la existencia de facturas a las cuales en la columna "cmdNumeroFactura", correspondiente al número de la factura, existían campos con los prefijos en letras, o que estos campos contaban con ceros por delante del número de la factura, por lo que se debió corregir estas inconsistencias, separando los prefijos de las facturas, colocando el prefijo en su campo correspondiente y borrando los ceros que anteceden los números de las facturas. Se realizó nuevamente el test de IGUAL-IGUAL-IGUAL el cual se denominó como SSS2, y se encontraron 680 facturas sospechosas y el valor total de las facturas es de $1.693.940.668,00 pesos M/CTE.

2. Test IGUAL-IGUAL-DIFERENTE

Para este test al igual que el test de IGUAL-IGUAL-IGUAL se determinan los parámetros que se van a analizar, este test permite que se seleccionen hasta 8 campos para obtener resultados iguales y uno adicional que no concuerde con los demás. Mark Nigrini dice "este test es una herramienta poderosa para encontrar errores y fraude, pues en su experiencia este detecta los errores en las cuentas por pagar, cuando mayor es el periodo de tiempo analizado, mayores son las probabilidades del test IGUAL-IGUAL-DIFERENTE de encontrar errores." (Nigrini, 2011)[5]

Este test lo que busca es encontrar duplicados casi exactos, pero con una diferencia al test IGUAL-IGUAL-IGUAL y es que esta vez buscamos campos iguales, pero con la condición que un campo adicional sea diferente, para esto construimos un nuevo query en el cual al igual que en nuestro test anterior buscara 2 campos iguales como lo son "idNit" este campo corresponde al número de identificación de la IPS, "PrNumFactu1" es el número de la factura emitida por la IPS, pero con la diferencia que esta vez el campo diferenciador sea el campo "EPS" el cual identifica si la factura presentada está siendo radicada por régimen subsidiado o régimen contributivo. El resultado de este test SSD1 por sus siglas en inglés (SAME-SAME-DIFFERENT), fue 432 facturas por un valor total de $1.830.860.874,00 pesos M/CTE. Al igual que en el test anterior en la columna "cmdNumeroFactura", correspondiente al número de la factura, existían campos con los prefijos en letras, o que estos campos

[4] Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations, Mark Nigrini, Capítulo 12, 2011, pag. 234-235

[5] Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations, Mark Nigrini, Capítulo 12, 2011, pag. 235-236

contaban con ceros por delante del número de la factura, por lo que se debió corregir estas inconsistencias. Por lo que se creó un nuevo test denominado SSD2 y se encontraron 710 facturas por un valor de $2.677.555.382,00 pesos M/CTE.

3. Rangos Intercuartílicos

Un diagrama de cajas y bigotes permite identificar la mediana, los cuartiles de los datos que se estén analizando y por ende el rango intercuartílico (RIC) dado por la fórmula Q3–Q1, así como los valores atípicos inferiores y superiores que corresponden a Q1–1.5·RIC y Q3+1.5·RIC respectivamente.

Tomando estas reglas como referencia, se realizó el cálculo de rangos intercuartílicos para el campo valor unitario por código de servicio, identificando los límites inferiores y superiores para cada servicio y de esta forma identificar los valores que estén fuera de este rango como valores atípicos.

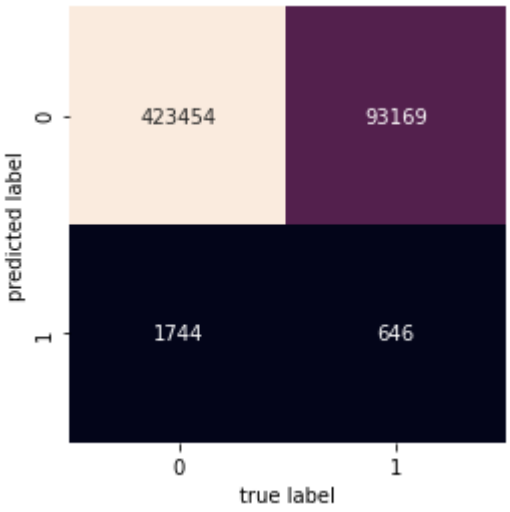
Se identificaron 312.966 registros con valores atípicos en el campo de valor de producto y se marcaron en una nueva columna denominada “ValUnitAt“, con la creación y marcación de esta variable se tiene un punto de partida para la aplicación de modelos de clasificación supervisados.

4. Regresión Logística

Se utilizó el algoritmo de regresión logística, como algoritmo de clasificación, para poder predecir la probabilidad que este identifique la variable categórica en este caso “ValUnitAt”, dado que esta variable se codificó con cero (0) para las facturas consideradas como normales y con uno (1) las facturas consideradas sospechosas

Este tipo de análisis de regresión es de carácter predictivo, con él, se quiso probar si existía algún tipo de relación entre las variables del set de datos con la variable “ValUnitAt” que permitan explicar la variable dependiente. Este modelo se usó como un primer intento para clasificar correctamente las facturas con sospecha de fraude, siendo este tipo de algoritmo bastante sencillo de ejecutar pues no requiere de muchos recursos computacionales.

Los resultados del algoritmo regresión logística nos arrojó un accuracy score del 82%, lo cual en una primera mirada parece un muy buen resultado. Pero si se mira más detalladamente con las métricas del reporte de clasificación en primer lugar obtenemos un resultado de precisión del 100% para las facturas consideradas como normales (codificadas con el 0), mientras que las facturas consideradas como sospechosas de fraude el 1% (codificadas con el 1). La metrica de recall tiene una métrica de 82% para las facturas con cero, mientras que las facturas codificadas con 1 obtuvo un 27%, y por último el f1-score clasificó el 90% de las facturas normales y las facturas que se consideran sospechosas solo logró identificar el 1%. Por tal motivo este algoritmo de clasificación no tuvo buenos resultados, dado que no logró capturar las características de las facturas catalogadas con el número 1, consideradas como sospechosas y se descarta como un modelo que pueda a futuro servir para predecir el fraude en esta base de datos.



Fuente: Elaboración propia

	precision	recall	f1-score	support
0	1.00	0.82	0.90	516623
1	0.01	0.27	0.01	2390
accuracy			0.82	519013
macro avg	0.50	0.54	0.46	519013
weighted avg	0.99	0.82	0.90	519013

Fuente: Elaboración propia

5. Regresión Logística y uso de Ridge y Lasso

¿Cuándo es efectiva Lasso (L1)?

"Lasso nos va a servir de ayuda cuando sospechamos que varios de los atributos de entrada (features) sean irrelevantes. Al usar Lasso, estamos fomentando que la solución sea un poco densa. Es decir, favorecemos que algunos de los coeficientes acaban valiendo 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice mejor. Lasso nos puede ayudar, en este sentido, a hacer la selección de atributos de entrada. Lasso funciona mejor cuando los atributos no están muy correlados entre ellos."[6]

[6] Regularización Lasso L1, Ridge L2 y ElasticNet - https://www.iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/

Regularización Ridge (L2)

"En la regularización Ridge, también llamada L2, la complejidad C se mide como la media del cuadrado de los coeficientes del modelo. Al igual que ocurría en Lasso, la regularización Ridge se puede aplicar a varias técnicas de aprendizaje automático."[7]

```
[ ] # Matriz de confusión
    pd.crosstab(y_test, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)
```

Predicted	0	1	All
True			
0	424909	116	425025
1	92949	1039	93988
All	517858	1155	519013

```
[ ] # Accuracy o precisión
    print("model score: %.8f" % grid_search.score(x_test, y_test))

    model score: 0.82068850
```

```
[ ] ##accuracy, precisión, recall
    print(classification_report(y_test, y_pred))
```

```
                  precision    recall  f1-score   support

               0       0.82      1.00      0.90    425025
               1       0.90      0.01      0.02     93988

        accuracy                           0.82    519013
       macro avg       0.86      0.51      0.46    519013
    weighted avg       0.83      0.82      0.74    519013
```

Fuente: Elaboración propia

Regularización LASSO

"Es un método que combina un modelo de regresión con un procedimiento de contracción de algunos parámetros hacia cero y selección de variables, imponiendo una restricción o una penalización sobre los coeficientes de regresión. Lasso nos va a servir de ayuda cuando sospechamos que varios de los atributos de entrada (features) son irrelevantes. Al usar Lasso, estamos fomentando que la solución sea un poco densa. Es decir, favorecemos que algunos de los coeficientes acaban valiendo 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice

[7] Regularización Lasso L1, Ridge L2 y ElasticNet - https://www.iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/

mejor. Lasso nos puede ayudar, en este sentido, a hacer la selección de atributos de entrada. Lasso funciona mejor cuando los atributos no están muy correlados entre ellos."[8]

```
[ ] # Matriz de confusión
    pd.crosstab(y_test, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)
```

Predicted	0	1	All
True			
0	424909	116	425025
1	92950	1038	93988
All	517859	1154	519013

```
# Accuracy o precisión
print("model score: %.8f" % grid_search.score(x_test, y_test))
```

```
model score: 0.82068657
```

+ Código + Texto

```
[ ] ##accuracy, precisión, recall
    print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.82      1.00      0.90    425025
           1       0.90      0.01      0.02     93988

    accuracy                           0.82    519013
   macro avg       0.86      0.51      0.46    519013
weighted avg       0.83      0.82      0.74    519013
```

Fuente: Elaboración propia

6. K NEIGHBORS

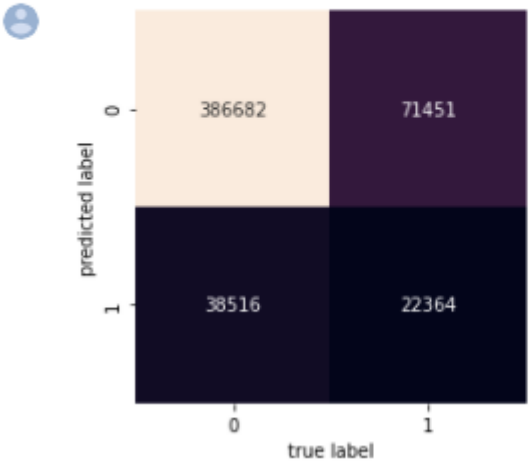
```
from sklearn.neighbors import KNeighborsClassifier
knn= KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)
Y_pred=knn.predict(X_test)
print('Precision vecinos mas cercanos')
print(knn.score(X_train,y_train))
```

```
Precision vecinos mas cercanos
0.8753440256178836
```

Fuente: Elaboración propia

[8] Regularización Lasso L1, Ridge L2 y ElasticNet - https://www.iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/

```
mat = confusion_matrix(y_test, Y_pred)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False)
plt.xlabel('true label')
plt.ylabel('predicted label');
```



Fuente: Elaboración propia

```
print(metrics.classification_report(Y_pred, y_test))
```

```
              precision    recall  f1-score   support

           0       0.91      0.84      0.88    458133
           1       0.24      0.37      0.29     60880

    accuracy                           0.79    519013
   macro avg       0.57      0.61      0.58    519013
weighted avg       0.83      0.79      0.81    519013
```

Fuente: Elaboración propia

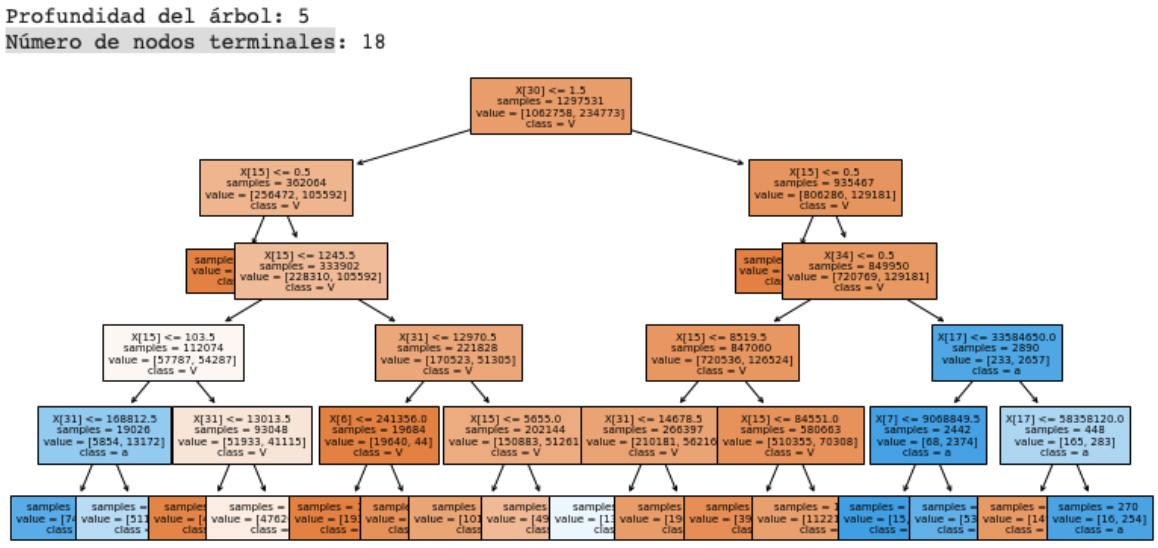
7. Árbol de Decisión

Este algoritmo se utiliza para realizar una clasificación supervisada, mediante la búsqueda de una variable dependiente concreta, en nuestro caso, la variable dependiente para este análisis se tomará de una nueva columna que se agrega al conjunto de datos luego de realizar los Rangos Intercuartílicos, IGUAL-IGUAL-IGUAL e IGUAL-IGUAL-DIFERENTE, con estos encontramos que el test SSD2 (por sus siglas en inglés SAME-SAME-DIFFERENT) contiene en sus 2663 registros, los resultados de los test SSS1, SSS2 y SSD1, y se les clasifica como “facturas sospechosas”, de igual manera se catalogan como “facturas sospechosas” los 310.303 encontrados después de este proceso, el resto de la base de datos se clasifican como “Ok”. Este árbol de decisión al tener como variable dependiente una variable cualitativa, se considera como un árbol de clasificación.

“La idea de un árbol de decisión es dividir el conjunto de datos original en dos o más subconjuntos en cada paso del algoritmo, para aislar mejor las clases deseadas. Cada paso produce una división en el conjunto de datos y cada división se puede representar gráficamente como un nodo. La secuencia de nodos, es decir, la secuencia de divisiones se puede visualizar como un árbol, cuyas ramas definen una ruta de regla para aislar las clases deseadas.” (https://www.knime.com/knime-introductory-course/chapter6/section3, 22 de mayo 2020)

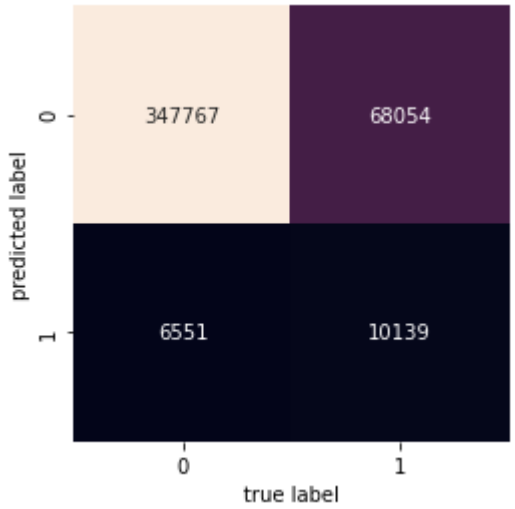
La partición se realizó 70 - 30, 70% para entrenamiento del algoritmo y un 30% para evaluar el muestreo estratificado sobre la columna denominada “ValUnitAt”. Después se agrega el nodo de aprendizaje del árbol de decisión el cual se configura mediante la medida de calidad Gini Index, se ejecuta el árbol de decisión con una profundidad de 5 y Número de nodos terminales de 18, por último se calcula la precisión del modelo, el cual arroja un nivel de precisión del 83%, pero este árbol solo identifica mediante el f1-score correctamente el 90% de las facturas consideradas normales las cuales fueron marcadas con cero “0”, en cambio en el f1-score para las “facturas sospechosas” solo es del 21%.

Ilustración 11 – Árbol de decisión



Fuente: Elaboración propia

Ilustración 12 – Matriz de Confusión Árbol de decisión



Fuente: Elaboración propia

El árbol de decisión resultante presenta el siguiente reporte de la clasificación realizada por el algoritmo.

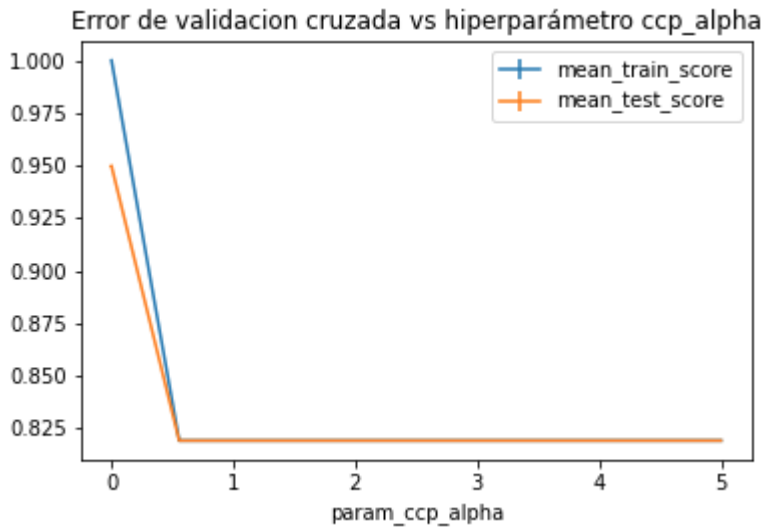
Ilustración 13 – Reporte de métricas

	precision	recall	f1-score	support
0	0.84	0.98	0.90	354318
1	0.61	0.13	0.21	78193
accuracy			0.83	432511
macro avg	0.72	0.56	0.56	432511
weighted avg	0.79	0.83	0.78	432511

Fuente: Elaboración propia

En el resultado anterior se utilizó un valor de max_depth=5, pero no sabemos si es el mejor valor. Para poder identificar la profundidad óptima que reduce la varianza y aumenta la capacidad predictiva del modelo, se realizó la poda del árbol mediante el proceso de pruning.

Ilustración 14 – Resultados de poda del árbol – 250625 variables



Fuente: Elaboración propia

La profundidad después de la poda del árbol fue de 48 niveles, el número de nodos terminales fue de 46.378 y El accuracy después de la poda aumenta al 95.2%. Con estos datos podemos decir que el árbol de decisión presentó un sobre ajuste en los datos y al intentar buscar el número óptimo de niveles y nodos del árbol presentó el problema de sobreajuste.

8. Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado que se basa en el aprendizaje por conjuntos. Se construyeron dos modelos de clasificación Random Forest para predecir la correcta clasificación de las facturas denominadas como sospechosas, uno con 10 árboles de decisión y otro con 100 árboles de decisión. Esto se hace con el fin de poder probar si al aumentar el número de aumenta con el número de árboles de decisión en el modelo.

Para el primer modelo con 10 árboles de decisión el cual se se ejecutó con los parámetros por defecto, configurados en la librería sklearn RandomForestClassifier, con este modelo se obtuvo un accuracy score con 10 árboles de decisión del 90.92%.

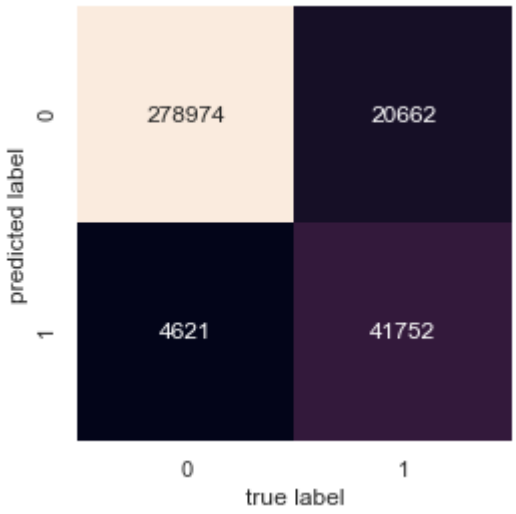
Posteriormente se realizó la prueba modificando los parámetros por defecto, cambiandolo por n estimators=100, sin embargo al aumentar el número de decisión se pretende aumentar la precisión del modelo random forest, sin embargo, el accuracy score con 100 árboles de decisión fue del 90.92%. Lo cual nos demuestra que a más árboles de decisión integrados al modelo esto no influye en la precisión obtenida por este algoritmo.

De igual manera se identificaron las variables más importantes para el modelo random forest, en el siguiente recuadro se muestran las 10 variables con mayor importancia para predecir la variable objetivo, en este caso “ValUnitAt”, la cual es la variable con la que queremos identificar si el algoritmo logra reconocer e identificar las facturas sospechosas de fraude.

Puntaje de las variables más importantes en el modelo Random Forest.

Nombre de la Columna	Puntaje
CodigoServicio	0.164817
dprvlrproducto	0.156324
ValorTotalDetalle	0.119212
cmdNumeroFactura	0.051139
dprIDDetallePrestacion	0.046818
idTipoProducto	0.039223
idNit	0.038798
cmdTotalFactura	0.035871
cmdcuentamedica	0.033795
dprcantproducto	0.031769

La matriz de confusión nos muestra con un 20% del set de datos que el algoritmo logra aumentar significativamente los resultados frente a los resultados de los otros algoritmos utilizados. Se puede observar cómo el algoritmo logró identificar correctamente 41.752 facturas consideradas sospechosas de fraude.



Fuente: Elaboración propia

Las métricas del modelo nos muestran un accuracy score del 93%, pero esta vez los puntajes precisión, recall y f1-score, logran identificar correctamente las facturas marcadas con el número 1 en la columna “ValUnitAt”, para la precisión, se obtuvo un 90%, recall tuvo un puntaje del 67% y el f1-score fue de 77%. Con lo anterior podemos concluir que el modelo de Random Forest, se pueden identificar facturas susceptibles de ser fraudulentas en la base de datos.

	precision	recall	f1-score	support
0	0.93	0.98	0.96	283595
1	0.90	0.67	0.77	62414
accuracy			0.93	346009
macro avg	0.92	0.83	0.86	346009
weighted avg	0.93	0.93	0.92	346009

Fuente: Elaboración propia

9 . Métricas comparativas de los Modelos

Modelo	Clase	Precisión	Recall	F1-score	Support	accuracy
Regresión Logistica	0	0.82	1.00	0.90	425.206	
	1	0.00	0.00	0.00	93.807	0,82
K Neighbors	0	0.91	0.84	0.88	305.049	
	1	0.24	0.38	0.30	40.960	0,79
Random Forest	0	0.93	0.98	0.96	283.595	
	1	0.90	0.67	0.77	62.414	0.93
Árbol de decisión	0	0.84	0.94	0.90	354.318	
	1	0.61	0.13	0.21	78.193	0,83

Conclusiones

- El proceso de exploración, depuración y de examinación en general de los datos del proyecto, es el primer paso y es sumamente necesario con el fin de garantizar o lograr alta calidad de los datos que van a ser usados en el proyecto de analitica.

- Los test IGUAL-IGUAL-IGUAL e IGUAL-IGUAL-DIFERENTE aplicados a la detección de facturas fraudulentas en un conjunto de datos, permite encontrar casos que nos son fáciles de detectar mediante los métodos tradicionales de auditoría, como por ejemplo la selección de una muestra aleatoria al conjunto de datos, o muestreo estratificado o dada la experiencia de quien revise la información a criterio personal. Estas pruebas facilitan la selección de casos para investigar a fondo y verificar si lo que estos test detectaron corresponde a un error o son efectivamente casos de fraude de parte de las IPS.

- Con la aplicación del método IGUAL - IGUAL - DIFERENTE se identificó que algunas IPS radican la misma factura en los dos regímenes dejándolas en estados radicado activo, lo que implica una mala práctica por parte de las IPS. Se sugiere realizar un análisis detallado sobre estos casos para identificar si se trata de fraudes reales. Estos casos verificados corresponden al período de prestación de servicios del año 2020 en el cual se pudieron identificar mediante el método IGUAL-IGUAL-IGUAL 680 facturas por un valor total de $1.693.940.668,00 pesos M/CTE, El test IGUAL-IGUAL-DIFERENTE identificó 710 facturas por un valor de $2.677.555.382,00 pesos M/CTE y cabe resaltar de este test que en la teoría del profesor Nigrini se hace más eficiente entre mayor es el periodo de evaluación, si se expande el periodo de análisis a un año, queda la inquietud de cuántos casos se pueden encontrar, y hacer un análisis de comportamiento de las regiones año a año, y sobre todo cuantificar cuántos de estos casos fueron identificados en el proceso de auditoría tradicional.

- Los rangos intercuartílicos permitieron identificar límites inferiores y superiores para cada servicio y de esta forma identificar los valores que estén fuera de este rango que se catalogaron como valores atípicos. Se identificaron 312.966 registros que junto con los test IGUAL-IGUAL-IGUAL e IGUAL-IGUAL-DIFERENTE fueron el punto de partida para la aplicación de modelos de clasificación supervisados.

- El algoritmo Random Forest tuvo la mayor precisión de todos los modelos utilizados, un 93%. Este modelo fue el que mejor pudo clasificar las facturas marcadas como sospechosas con un f1-score del 77% y del 96% a las facturas consideradas como normales. La explicación se debe a que este proyecto se basó en la identificación de datos atípicos dentro de la base de datos, y este tipo de algoritmo se caracteriza por realizar la clasificación basado en la búsqueda de datos atípicos. Con lo anterior podemos concluir que este modelo clasifica e identifica facturas sospechosas dentro de la base de datos de esta EPS. Si tenemos en cuenta la justificación de este proyecto que en el 2019 según reporte de la Supersalud las utilidades de las IPS aumentaron mientras que para el mismo periodo las pérdidas de las EPS se incrementaron. Este modelo puede funcionar como soporte al área de auditoría de la EPS que por volumen de facturación (8.561.061

registros que consta esta base para el 2020), les resulta casi imposible poder detectar y revisar estos errores en la facturación presentada por las IPS, siendo este modelo una alternativa para mejorar la identificación de fraudes en la facturación.

BIBLIOGRAFÍA

- Mesa FR, Ranieri A, Maturana S, Kaempffer AM, Fraudes a los Sistemas de Salud en Chile: Un Modelo para su Detección, Rev. Panam 2009, págs. 56-61
- Cinco Principios para la Gestión de Riesgo de Fraude – PWC – Julio 2019
- Técnicas de Minería de Datos para la Detección y Prevención del Lavado de Activos y la Financiación del Terrorismo (LA/FT) - Unidad de Información y Análisis Financiero (UIAF) – 2014
- Dallas Thorton, Ronald Mueller, Paulus Schoutsen, Jos Hillersberg - Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analisys Techniques for Fraud Detection - 2013
- Karen Vergara - Detección de Fraudes en Bodegas de Datos basado en los niveles de agregación – 2017
- 4 cosas que no sabes sobre la asombrosa Ley de Benford en Auditoría – Nahun Frett – 20-jun-2016 www.auditool.org
- Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations, Mark Nigrini, Capitulo 12, 2011

ANEXOS

Anexo 1 - Diccionario de Datos

A continuación, se muestra el diccionario de datos, donde se explica el significado de las columnas en el conjunto de datos denominado klCabecera.

Column_name	Type	Observación	Valores
EPS	smallint	Identificador de red y regimen de la cuenta medica	1. Red - Contributivo 2. No Red - Contributivo 3. Red - Subsidiado 4. No Red - Subsidiado
cmdcuentamedica	int	Consecutivo de cuenta medica	
prsIdAmbito	smallint	Identificador de ambito de la cuenta medica	1. Ambulatorio 2. Urgencia 3. Hospitalario 4. Domiciliario
Cobertura	smallint	Identificador de cobertura de la cuenta medica	1. PBS 0. NoPBS
prsFchEntrada	datetime	Fecha inicio prestacion de servicio	
prsFchSalida	datetime	Fecha fin prestacion de servicio	
prsFechaVistoBueno	datetime	Fecha de visto bueno auditoria cuenta medica	
dprvlrproducto	money	Valor unitario del producto o servicio	
dprcantproducto	numeric	Cantidad producto o servicio facturado	
ValorTotalDetalle	numeric	Valor total del servicio o producto	

dprVlrCopago	numeric	Valor copago	
dprVlrModeradora	numeric	Valor cuota moderadora	
prsIDPrestacion	int	Consecutivo de prestación	
dprIDDetallePrestacion	int	Consecutivo de detalle de prestación	
prsIdPaciente	int	Identificador de afiliado por regimen	
Edad	tinyint	Edad del afiliado	
prsTipAfiliado	int	Identificador de tipo de afiliado	1. Cotizante 2. Beneficiario 3. Beneficiario Adicional 6. Cabeza Flia Subsidiado 7. Benef. Subsidiado
prsIdMunicPaciente	int	Identificador de municipio del afiliado	
prsIpsBasicaPaciente	int	Identificador de IPS basica del afiliado	
CodigoServicio	int	Codigo del servicio o producto	
CuentaEvento	tinyint	Identifica si la modalidad de contrato es o no evento	1. Es evento 2. No es evento

GrupoEtareo	smallint	Identificador de grupo de afiliados por edad y genero	1. Menores de un año 2. De 1 a 4 años 3. De 5 a 14 años 4. De 15 a 18 años (Hombres) 5. De 15 a 18 años (Mujeres) 6. De 19 a 44 años (Hombres) 7. De 19 a 44 años (Mujeres) 8. De 45 a 49 años 9. De 50 a 54 años 10. De 55 a 59 años 11. De 60 a 64 años 12. De 65 a 69 años 13. De 70 a 74 años 14. De 75 años y mayores
tipoproducto	smallint	Identificador del tipo de producto o servicio	0. No Homologado 1. Medicamento 2. Procedimiento 8. Insumo
prsgrabada	tinyint	Identifica si la cuenta fue o no grabada para el paso al sistema contable	0. No grabada 1. Grabada
prsVistoBueno	tinyint	Identifica si la cuenta medica tiene o no visto bueno en la auditoria	0. Sin visto bueno 1. Con visto bueno
prsTipoAtencion	int	Identificador de tipos de atencion	

dprEsRecobrable	tinyint	Identifica si el servicio es o no recobrable	0. No recobrable 1. Recobrable
ValorTotalGlosadoModulo	numeric	Valor glosado	
ValorAceptadoIPS	numeric	Valor aceptado por la IPS	
ValorAceptadoEPS	numeric	Valor aceptado por la EPS	
ValorPorConciliar	numeric	Valor pendiente por conciliar	
ddpIDDiagnostico	int	Identificador de diagnostico del afiliado	
cmdIDTipoCuenta	smallint	Identificador del tipo de contrato de la cuenta medica	1. Capitación 2. Por Evento 3. PyP Evento 4. Monto Fijo 5. PyP Capita 6. Pago Global Prospectivo 7. Pago por Actividad Final 8. Pago en Bloque 9. Tarjeta mas Efectiva 10. Proveedores Medicamentos
cmdPrefijoFactura	varchar	Prefijo de la factura	
cmdNumeroFactura	varchar	Numero de factura	
cmdFchEmisionFactura	datetime	Fecha de factura	
cmdfchradicacion	datetime	Fecha de radicacion ante la EPS	
cmdTotalFactura	money	Valor total de la factura	

ipsidips	int	Identificador de la IPS	
cmdIdestadocuenta	int	Identificador de estado de la cuenta medica	
PeriodoRadicacion	int	Periodo de radicación (Periodos de 21 a 20 de cada mes)	
FAC_CONT	int	Consecutivo de contabilidad	
cabVrModeradora	money	Valor de cuota moderadora	
cabVrCopago	money	Valor de copago	
cabVrDscto	money	Valor de descuento	
VlrGlosa	money	Valor glosado	
FAC_FECH	date	Fecha de causación	
FAC_DESC	varchar	Descripcion contable	
FAC_ESTA	varchar	Estado en sistema contable	
FAC_VATO	money	Valor con deduccion de impuestos	
FecPagoReciente	date	Ultima fecha de pago	
ValTotalPago	money	Valor total pagado	
PrNumFactu1	varchar	Prefijo y numero de factura	
PrNumFactu2	varchar	Prefijo y numero de factura limpio	
cabVrCausado	numeric	Valor que migra al sistema contable	
cabVrPagado	numeric	Valor Pagado	
cmdgeneradaseven	tinyin	Identifica si la factura migra al sistema contable	

	t		
cmdTotalIva	numeric	Valor del IVA	
cmdTotalReteFuente	numeric	Valor retefuente	
cmdTotalReteIca	numeric	Valor reteica	
cmdTotalICO	numeric	Valor total ICO	
cafRadOtroOperador	varchar	Radicado de otro operadpr	
cmdIdestadocuentacalc	tinyint	Identificados de estado de cuenta calculado	
bd	tinyint	Identificador de base de datos	
esrg	bit	Identificador de reembolso glosa	
esUltimo	smallint	Identifica si la cuenta medica es la ultima recepcionada	
esUltimoBD	bit	Identifica si la cuenta medica es la ultima recepcionada por regimen	
idAmbito	smallint	Identificador de ambito de la cuenta medica	
NombreAmbito	varchar	Nombre del ambito de la cuenta medica	
idTipoCuenta	smallint	Identificador del tipo de cuenta	
NombreTipoCuenta	varchar	Nombre del tipo de cuenta medica	
esContratoEspecial	bit	Identifica si es o no un contrato especial	
idCobertura	smallint	Identificador de cobertura de la cuenta medica	
NombreCobertuta	varchar	Nombre de la cobertura de la cuenta medica	

DepartamentoPrestador	varchar	Departamento del prestador	
MunicipioPrestador	varchar	Municipio del prestador	
CodigoDANEDeptoMunPrestador	varchar	Codigo DANE del prestador	
RegionalPrestador	varchar	Regional del prestador	
ipsIDMunicipio	smallint	Identificador del municipio del prestador	
idDimPrestador	int	Ientificador unico de prestador y sede	
ID	int	Identificador unico del afiliado	
FechaNacimiento	date	Fecha de nacimiento del afiliado	
Genero	tinyint	Genero del afiliado	
DANEAfiliado	varchar	Codigo DANE del afiliado	
FechaRadicacion	datetime	Fecha de radicacion de afiliacion	
FechaAfiliacion	datetime	Fecha de afiliacion	
FechaRetiro	datetime	Fecha de retiro del afiliado	
TipoAfiliado	tinyint	Tipo de afiliado	
EstadoAfiliacion	tinyint	Estado de afiliacion	
RazonEstadoAfiliacion	tinyint	Descripcion del estado de afiliacion	
IDIPS	int	Identificador de la IPS del afiliado	
GrupoEtario	tinyint	Grupo de edades a la que pertenece el afiliado	

OrigenBDUA	int	Identifica si el origen es BDUA	
EstadoBDUA	varchar	Estado en BDUA	
Cohorte	varchar	Cohorte o cohortes a las que pertenece el afiliado	
SubCohorte	varchar	Subcohorte o subcohortes a las que pertenece el afiliado	
CantidadCohortes	tinyint	Cantidad de cohortes del afiliado	
CodigoHeonServicio	int	Codigo interno del servicio	
DescripcionHeonServicio	varchar	Descripcion del servicio	
Tipo_Cod	varchar	Tipo de servicio prestado	
Codigo	varchar	Codigo normativo del servicio	
Descripcion_Cups_Cum	varchar	Descripcion normativa del servicio	
Seccion	varchar	Seccion a la que pertenece el servicio	
Servicio	varchar	Descripcion del servicio	
Agrupador2NTP	varchar	Agrupador del servicio	
Nivel	tinyint	Nivel del servicio	
Especialidad	varchar	Especialidad a la que pertenece el servicio	
Serv_Padre	tinyint	Servicio principal	
idGrupoEtareo	smallint	Identificador de grupo de edad del afiliado	
NombreGrupoEtareo	varch	Grupo de edades a la que pertenece el afiliado	

	ar		
dxIDDiagnostico	int	Identificador del diagnostico	
dxhomIDTipDiag	smallint	Identificador del tipo de diagnostico	
dxIDHomolDiag	varchar	Codigo CIE10 del diagnostico del afiliado	
dxDescripcion	varchar	Descripcion del diagnostico	
idServicio	int	Identificador del servicio	
idTipoProducto	smallint	Identificador del tipo de producto o servicio	
NombreTipoProducto	varchar	Nombre del tipo de producto o servicio	
idEstadoCuenta	smallint	Identificador del estado de cuenta medica	
NombreEstadoCuenta	varchar	Descripcion del estado de cuenta medica	