Análisis Exploratorio de Datos
Bootcamp Data Science
Rafael Ortega

Memoria

1. Análisis del perfil de las pacientes y del comportamiento de las citas en una consulta privada de ginecología, obstetricia y reproducción.
2. Análisis de variables obtenidas durante el control gestacional, con respecto a resultados perinatales.

Introducción

En una consulta privada de ginecología, obstetricia y reproducción ubicada en la ciudad de Rafaela, Prov. de Santa Fe, Argentina, se implementó desde el año 2014 un sistema informático de historias clínicas que, además de tener funcionalidades específicas para la especialidad, aloja todas la informacion en una base de datos Filemaker Pro.
Hay campos de texto, fechas, números enteros y con decimales. Hay campos de edición libre, y otros datos se completan a través de un listado predefinido, o a través de checkmarks.

A lo largo de los años, pareciera que la principal utilidad del sistema es justamente registrar información para ser utilizada cuando la paciente vuelve, y poder acceder fácilmente a su historial con un par de clicks. Sin embargo, sabemos que se han ido acumulando cientos y miles de datos con el uso cotidiano, pero están escondidos en la maquinaria de la base de datos y nadie los mira.

Seguramente esos datos nos pueden llegar a decir algunas cosas que no sabemos sobre la performance de la consulta en términos de productividad, conocer que tipo de pacientes son los que se acercan (de que aseguradoras, de que edad, adonde viven). Por otro lado, desde el punto de vista estrictamente médico, se almacenan datos que podrían demostrar alguna correlación entre variables, que hoy desconocemos, o simplemente quisiéramos corroborar.

Por este motivo, vamos a realizar un análisis exploratorio de datos sobre 3 tablas de la base de datos: datos de los pacientes, datos de las citas, datos de los embarazos.

Hipótesis principales:

1. Cual es el perfil de pacientes que solicitan cita en la consulta? Edad, aseguradora, localidad de residencia
2. ¿Cuál es el comportamiento de las citas en el tiempo según el tipo de cita (ginecológica, obstétrica, infertilidad)? ¿Qué proporción de pacientes regresan luego de una primera cita?
3. ¿Cómo se relaciona el incremento de peso materno diferencial entre las semanas 28 y 32 con el peso del RN?

Recolección de datos:

Los datasets son archivos .csv y/o excel que se exportan desde la interfaz del sistema de Filemaker Pro, a partir de las tablas de la base de datos relacional.

Al revisar las opciones de exportación de Filemaker, descubrí que al exportar las tablas con .csv, no se exporta el nombre de la columna. Entonces me vi obligado a descargarlos de otra manera (a los nombres de columnas), y ver si se puede crear una lista con ellos, y posteriormente agregarlos a cada columna. O bien, aprovechando que la cantidad de registros no superan los 9000, utilizar la opción de exportar directamente a Excel, que sí me pone los nombres de los campos como nombres de columnas. De esta manera puedo cargar el archivo con el comando 'read_excel' de pandas.

¿En qué consisten los datos (datasets) que vamos a analizar?

A. Datos personales de pacientes.
B. Datos de citas de estos pacientes.
C. Datos de embarazos de estas pacientes.

¿Qué preguntas podríamos responder con esta información?

De cada dataset (A, B y/o C), las preguntas principales que vamos a responder son:

A:

1. ¿Qué edad tenían los pacientes cuando solicitaron cita por primera vez?

HIPÓTESIS: El promedio de edad de las pacientes en su primera cita es de 35 años.

2. ¿De qué localidad son?

HIPÓTESIS: La mitad de las pacientes son de Rafaela

3. ¿Cuál es la distribución de cada aseguradora?

HIPÓTESIS: La aseguradora más frecuente es Prevención Salud

4. ¿Cuantas y cuales son las aseguradoras que reúnen el 80% del volumen?

HIPÓTESIS: Las aseguradoras que reúnen el 80% del volumen son 20.

5. ¿Cuántas pacientes nuevas hay por un periodo mensual/anual?

HIPÓTESIS: Cada mes hay 5 pacientes nuevos. Cada año hay 60 pacientes nuevos.

B:

1. ¿Cómo se distribuyen estos 3 tipos de consulta? En su totalidad y su variación por periodos mensuales/anuales.

HIPÓTESIS: Las consultas de Obstetricia son las más frecuentes.

HIPÓTESIS: Las consultas de infertilidad aumentan con el tiempo.

3. Definir si existió segunda cita. Tasa de repetición.
HIPÓTESIS: La tasa de repitencia es de 0,8.

4. Definir cuánto tiempo pasó entre la primera y segunda cita.
HIPÓTESIS: El tiempo promedio entre la primera y la segunda cita es de 30 días.

C:

1. Identificar el peso en la semana 30 de gestación +/- 2 sem.
2. Relacionarlo con el peso fetal al nacimiento.

HIPÓTESIS: El incremento de peso materno entre la semana 28 y 32 del embarazo modifica el peso del recién nacido al momento del nacimiento.

Limpieza de 1er data set: Datos de pacientes.xlsx

- Al ejecutar un .info(), me doy cuenta que los datos no están completos. Una de las variables más importantes del análisis, y que por lo tanto actúa como dato limitante, es la fecha de nacimiento (FN) de la paciente. La FN en este caso sirve para calcular la edad, que ya aparece en su propia columna y podría utilizarla. Pero el sistema de origen calcula la edad a la fecha de la exportación (actual) pero si quiero calcular la edad en otra fecha, necesito la FN como tal.
- Otra inquietud que me surge es si hay algún registro duplicado, porque los DNI deben ser únicos.

Identifique las columnas que no serán útiles: Comentarios, correo electronico, telefono fijo, telefono movil, plan de obra social, número de afiliado y dirección.

- Al ver los tipos de datos, me doy cuenta que los DNI están interpretados como Float. No los voy a necesitar para el análisis propiamente dicho, solo para asegurar que son registros únicos. Así que los dejó como float.
- La columna edad es Object, porque tiene la palabra "años".
- La fecha de nacimiento tiene la hora:min:seg.
- La primera fila es un dato de prueba.
- Obra Social es un término regional que puede dar confusión si un extranjero lee los datos.
- Fecha de creación del registro es un nombre de columna muy largo.

Asigno el data frame limpio a la variable `pac processed`.

En base a estas observaciones, la limpieza consistirá en conservar las filas que tengan al menos: DNI único + ID único.

- Se eliminarán las columnas que no son útiles.
- Se eliminará la primera fila.
- Si hay registros con DNI duplicado, verificar el resto de los datos de las filas, si son diferentes, entonces es un error de tipeo y no se trata de un registro doble.Solo utilizo este parámetro para asegurarme de que todos los registros tienen datos de personas distintas.

No me es útil como dato de análisis. Si son iguales, entonces elimino filas hasta dejar uno solo.

- Se eliminará el string " años" de la edad.
- Se modificará el nombre de la columna "Fecha de creación del registro" por un nombre corto "fecha_creacion".
- Se modificará el nombre de la columna "Obra Social" por el nombre "Aseguradora".
- Se eliminarán las horas:min:seg de la columna Fecha de Nacimiento.

Guardo el dataframe principal limpio en formato .csv

Limpieza de 2do data set: Citas de pacientes.xlsx

Quitamos las filas que no tengan ID paciente, que no tengan fecha de cita y que no tengan definido el tipo de cita.

Veo que tenemos 50 columnas, intentaremos quedarnos con las columnas que nos interesan. 'ID Paciente',
'Fecha de Cita',
'Año de la cita',
'Mes de la cita',
'Tipo de citas estadística',
'Consulta estadistica',
'Práctica estadística',
'Consulta y practica estadistica'

También modifique los nombres de columnas por nombres más cortos.

Durante el trabajo sobre la base de datos se habían hecho algunas agrupaciones sobre los tipos de cita, que eran muy detallados. Por ejemplo: Infertilidad Monitoreo Folicular - Infertilidad Eco basal - Infertilidad primera vez - etc. En el dataset ya deberían aparecer unificados en Infertilidad - Ginecológicas - Obstétricas.
Cuando reviso los valores únicos, me doy cuenta que además de los 3 tipos de citas “estadísticos”, hay algunos más que quedaron fuera de la agrupación. Hago una transformación de los datos de esa columna para quedarme con solo 4 tipos de cita: ginecológica, obstetricia , infertilidad y otros (en su mayoría consultas burocráticas de recetas y certificados).

'Ginecológica': 'Ginecología',
'Suelo pélvico': 'Ginecología',
'Ecografia': 'Obstetricia',
'Obstétrica': 'Obstetricia',
'Infertilidad': 'Infertilidad',
'Burocrática': 'Otro',
'Entrevista prenatal': 'Obstetricia',
'Infertilidad Resultado de Ciclo': 'Infertilidad',
'No': 'Otro',
'Eco basal': 'Obstetricia',
'MF - Eco 2': 'Obstetricia',

'MF - Eco 3': 'Obstetricia',
'MF - Eco 4': 'Obstetricia',
'Infertilidad - BHCG': 'Infertilidad',
'Encuentro Pre parto': 'Obstetricia',
'Monitor Fetal': 'Obstetricia',
'Infertilidad Estudios': 'Infertilidad',
'Monitor fetal': 'Obstetricia',
'burocrática': 'Otro'

También está la fecha de cita, y aparecen las columnas 'Año de la cita' y 'Mes de la cita' Nos quedamos con ellas.

En la base de datos original también se trabajó con los criterios para definir una cita médica concreta. Debe reunir varios requisitos: Que la cita esté realizada (no cancelación, no posponer, etc), que se haga pagado (vía seguro o privado), que no sea parte de una cita de seguimiento de repro (en sistema se registra todo el ciclo de tratamiento con "citas", pero no en todas se hacen actuaciones). Vamos a quedarnos con esas columnas de 'Citas estadísticas', 'Prácticas estadísticas' y 'Citas y prácticas estadísticas'. Hay una tercera posibilidad y se da cuando la consulta médica y la práctica se hacen simultáneamente, eso se considera como una sola cita a fines estadísticos, y es excluyente de las citas solas, y las prácticas solas, porque sino se duplicarían las observaciones.

Consulta y practica estadistica considera como 1 a las citas en las que hubo una consulta Y una práctica
Las citas estadísticas consideran aquellas fechas en las que hubo una cita, independiente de si hubo práctica.
Las prácticas estadísticas consideran aquellas fechas en las que hubo una práctica, independiente de si hubo una consulta simultánea.
Para definir si existió segunda cita y qué porcentaje de mujeres regresaron a una segunda cita (Tasa de repetición), y para saber cuánto tiempo pasó entre la primera y segunda cita, hay que hacer algunos cálculos adicionales sobre el dataframe.

Para esto me quedaré con las columnas ID, fecha de cita, y tipo de cita estadística. Tengo que poner la fecha de cita en el index, y ordenar según fecha de cita ascendente. Restar el primer registro al segundo registro y obtener el tiempo entre 1era y 2da. Pero antes debo reconocer cuántas tienen la 2da cita. Quitar las que vinieron una sola vez.

En el manejo de los missing values, decidí que los NaN de consultas y de prácticas deben valer 0. Entonces codifico 1 si la cita existió, y 0 si no existió.
Al conocer el origen de los datos, no tengo dudas que los Nan equivalen a campos vacíos voluntariamente (= a cero) y no a omisiones. El criterio utilizado en la base de datos para definir una consulta (o práctica) es que esté realizada y facturada, entonces es casi imposible que haya un campo vacío si no corresponde.

Cambio el tipo de datos de las columnas float a integer.

Guardo el dataframe principal limpio como Citas - Processed.csv

Limpieza de 3er data set: Embarazos.xlsx

Es un dataframe con muchas columnas (229)

Me doy cuenta que las observaciones (registros) no tienen índice. Creó una columna con el ID correlativo ascendente.

Tengo que comprobar si el número de EG definitivo se corresponde con edad gestacional.
En la base de datos este campo se completa con un cálculo automático que es un f string entre texto y número. "30 sem y 4 días".
Compruebo en un caso al azar a través de la búsqueda en la base de datos por el nombre de bebe "Federica" que es registro único, que la exportación se hace en float, cuyos dos primeros dígitos corresponden a la semana, y el último corresponde al día.
Me doy cuenta que no necesito hacer ninguna transformación, aparte de pasarlo a int, porque puedo trabajar con ese dato perfectamente

Hay un problema grande ahora, porque los pesos maternos están registrados en campos que hacen referencia a una cita determinada, y no relacionados a una edad gestacional determinada.
Entonces, una paciente que va a su 6ta cita, puede estar de más o de menos semanas que otra. Tengo que tener el dato de edad gestacional por fuera del orden de las citas.
Entonces se me ocurre seleccionar las columnas de edad gestacional y peso para cada cita, y concatenarlas verticalmente de manera que los datos edad gestacional, e incremento de peso materno me queden todos alineados uno encima de otro, cada uno con su ID paciente.

Ahora si hago un filtro con una máscara que me conserve las edades gestacionales entre 28 y 32 sem. Si una paciente tuvo dos observaciones en esa ventana, saco el promedio de las mismas de modo que quede una observación única por paciente en esa edad gestacional.

Elimino ID pacientes duplicados en el data frame concatenado y entonces ya tendría los datos perfectos para el análisis.

El análisis en este caso se hace con comprobación de normalidad de Shapiro-Wilk, con variables cuantitativas y el ‘n’ de observaciones bajo.
Las dos variables no muestran distribución normal. Entonces las comparo con un test de correlación de Spearman