FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2017/2018

Trabajo de Fin de Grado

DEPURACIÓN DE UNA BASE DE DATOS.

***Alumno*: INÉS RUZAFA ESTEVE**

***Tutor*: JUANA Mª ALONSO REVENGA**

Junio de 2018

UNIVERSIDAD COMPLUTENSE
MADRID

ÍNDICE

1. INTRODUCCIÓN.

Eran las 10:00 horas del día 24 de marzo de 2014 en el aeropuerto de El Prat (Barcelona) y Patrick Sondenheimer, piloto comercial, se puso a los mandos del aparato del que era comandante. Por delante tenía un corto vuelo hacia Düsseldorf (Alemania), trayecto que había realizado cientos de veces anteriormente. Tras las rutinarias comprobaciones antes del despegue, pusieron rumbo a su destino, un destino tan inesperado para Sondenheimer como para el resto del mundo.

Nadie habría sido remotamente capaz de sospechar, hasta ese momento, lo que estaba a punto de suceder. El comandante, al poco de despegar, solicitó a su segundo que tomara el control de la aeronave para poder ir al baño sin ser consciente de que esa decisión le iba a costar muy cara.

En el momento que Sondenheimer salió de la cabina, alrededor de las 10:27 horas, el copiloto cerró la puerta de la misma, sellando así el destino de las 150 almas que iban a bordo. Pese a los fallidos intentos de convencer al copiloto para que abriera la puerta e incluso desesperados intentos de tirarla abajo a hachazos, el vuelo GWI 9525 con destino a Düsseldorf, se estrelló en los Alpes suizos. Fallecieron todos los pasajeros y la tripulación, 150 personas en total. A partir de este fatídico suceso, se endureció la normativa aérea. [1]

Este ejemplo, sirve para que se entienda la importancia de detectar cualquier factor anómalo en cualquier empresa sobre sus trabajadores. Es importante que se tenga la mayor cantidad de datos sobre los trabajadores de la misma para que se pueda entender a la perfección la situación en la que se encuentra la organización y también para posibles actuaciones de mejora sobre la misma.

Estos datos sobre la situación de la empresa y las actuaciones de mejora son recogidos por el departamento de Prevención de Riesgos Laborales (PRL) de la empresa. Según el blog de quirónprevención, en un artículo publicado el 17 de marzo de 2015: la prevención de riesgos laborales “es la disciplina que busca promover la seguridad y salud de los trabajadores mediante la identificación, evaluación y control de los peligros y

riesgos asociados a un entorno laboral, además de fomentar el desarrollo de actividades y medidas necesarias para prevenir los riesgos derivados del trabajo". [2]

En la prevención de riesgos laborales, siempre se ha partido de informes descriptivos. Es decir, su actuación en el futuro ha dependido siempre de sucesos del pasado. Según un artículo de Predictive Solutions Corporation, normalmente, los directivos de las empresas esperan a que uno de sus empleados, por ejemplo, se corte un brazo, para prevenir que otro de ellos no le ocurra la misma fatalidad. Pero mientras tanto, que ese primer empleado tenga cuidado y tome precauciones. [3]

Por ello, se ha empezado a aplicar aplicaciones estadísticas como la predicción en estos tipos de estudios. "Sus soluciones de software ayudan a rastrear tendencias y analizar datos relacionados con la seguridad. También emplean modelos patentados que predicen la probabilidad, frecuencia y localización de lesiones en el lugar de trabajo usando los datos de observación de seguridad de sus clientes"(Predictive Solutions Corporation) [3]. También, en el artículo de Sobhan Sarkar, Prediction of occupational accidents in a steel plant based on Text mining, se crea una red bayesiana que establece relaciones entre los factores que influyen en la variable objetivo de estudio y todo ello aplicado a accidentes laborales. [4]

1.1. Problemas encontrados.

La base de datos con la que se va a trabajar en este documento es de los eventos recogidos por el departamento de PRL de una determinada empresa.

Los tipos de evento ocurridos en una empresa pueden clasificarse como accidentes o incidentes. Un accidente es igual de inesperado que un incidente. La principal diferencia es que en un incidente nada ni nadie sufre ningún tipo de daño, suelen servir para alertar al departamento de PRL de que algo ha fallado. Por otro lado, el accidente es un evento que da lugar a consecuencias que afectan de forma negativa a algo o alguien y puede existir varios grados de gravedad sobre este. [5]

Uno de los problemas en la obtención de bases de datos de los eventos de una empresa es que la mayoría de los partes de accidente/incidente recolectados son rellenados por

técnicos o incluso por el mismo trabajador. Es decir, personas que no se dedican a evaluar luego esos datos, sino que solo saben que están obligados a informar de cualquier evento anómalo durante la jornada de trabajo, y eso hacen. Todavía en la mayoría de las empresas no está implantada una buena recogida de la información de todos los tipos de eventos que ocurren en ella. Por lo que, la mayoría de los informes datan de muchísimos valores perdidos o valores que no son correctos en los eventos recogidos por los empleados, ya sea porque piensan que tal información no es relevante, por prisa para irse a casa, porque no saben con qué formato se rellena tal variable, etcétera.

"Muchas empresas fallan en no dar prioridad a la gestión de calidad de los datos y no tener registros de la última vez que se realizó el control de la calidad de los datos" [6].

1.2. Planteamiento del trabajo.

Todos estos problemas son por una mala depuración de la base de datos. En el fichero que se obtuvo para un posible estudio se encontraron muchos valores perdidos, así como una mala introducción de los datos y categorización. De aquí surgió el objetivo de este trabajo: la depuración de una base de datos.

Se hace de vital importancia la depuración de la base de datos para poder realizar cualquier análisis de predicción y poder evaluar relaciones entre los factores recogidos a la hora de tener un accidente.

El primer paso será dejar la base de datos de forma que se pueda manejar fácilmente, recategorizando variables y cambiando el formato de ellas. Luego se manejarán los valores perdidos de ella mediante imputación múltiple.

"… debe evitarse el análisis completo de casos y el uso del método del indicador de ausencia, incluso cuando los datos faltan completamente al azar. Se sabe que los métodos de imputación múltiple son superiores a los métodos de imputación única…" [7]. Además en otro artículo se menciona que "los métodos de imputación únicos no tienen en cuenta la incertidumbre asociada a la falta de datos y tratar los valores de imputación individuales como si fueran reales en la fase de análisis puede dar lugar a

una subestimación de los errores típicos asociados con las estimaciones de diversas estadísticas calculadas a partir de datos" [8].

Una vez que se obtuvo el fichero de datos completo con una correcta depuración de datos, como se usa imputación múltiple para los datos faltantes, debido a su mejora en los resultados que el resto de análisis, se estudia una regresión logística para cada una de las imputaciones realizadas antes y poder combinarlas en una sola. De esta manera, se estudiarán relaciones de las covariables del análisis con la variable objetivo de estudio mediante los odds ratio.

Todo ello, se realizará con el programa estadístico SAS y se explicará paso a paso toda la sintaxis para una correcta depuración de datos y estudio posterior.

2. METODOLOGÍA PARA LA DEPURACIÓN DE DATOS.

En el inicio del estudio de este trabajo, se encontró un problema grave a la hora del análisis de regresión logística. Lo primero que hay que hacer antes de cualquier análisis es dejar la base de datos que se va a estudiar de forma adecuada. Cuando se seleccionaron las variables y eventos de forma que fuera lo más representativa posible nos encontramos con dos problemas. Uno de ellos era la inadecuada recogida de la información en algunas de las variables seleccionadas y otro fue la gran cantidad de datos perdidos.

Esto complicaba la realización de cualquier análisis debido a que las variables no estaban bien estructuradas y después que la muestra disminuía de forma significativa debido a que el programa SAS solo realizaría el análisis con los casos completos, se perdía mucha información. A partir de aquí, surgió la idea en este trabajo de realizar una especie de guía para la depuración de una base de datos.

2.1. Depuración de bases de datos.

La depuración de datos, también conocida como limpieza de datos, consiste en un proceso de detección y corrección de una base de datos. Esto se utiliza principalmente cuando se tienen datos incorrectos, incompletos, inexactos o irrelevantes. Con esta técnica es posibles corregir, completar, cambiar o eliminar esos datos. Las ventajas de esta técnica pueden ser la mejora en la adquisición, en este caso, de los tipos de evento ocurridos en la empresa. Con la depuración de datos se pueden eliminar observaciones que son incorrectas, que de esta forma se crea una información de la empresa más eficiente. Otra ventaja es la mejora en la toma de decisiones. Una correcta y progresiva depuración de la base de datos nos lleva a tomar mejores decisiones e incluso reducir costes. [6]

Dentro de la depuración de datos se encuentran técnicas que también son necesarias para el correcto procesamiento de los datos como codificación (cifrado) de los datos (darle valores numéricos a las variables categóricas para la facilidad del tratamiento de estas, o también por si la información es delicada, como nombres de personas o direcciones para que no se conozca información personal por la ley de protección de datos), transformación de los datos, imputación de valores perdidos, … [9]

Todo ello se aplicará a la base de datos proporcionada por el departamento de PRL de una empresa, ya que, como se ha mencionado, se han encontrado muchos problemas a la hora de realizar cualquier análisis.

El paso principal para una base de datos fiable es la correcta integridad de los datos. La integridad se define como la exactitud y fiabilidad de los datos. Por ejemplo, uno de los técnicos podría ingresar accidentalmente la fecha del evento en un campo de edad. Si el sistema creado aplica integridad a los datos estos errores se podrían evitar. Esto significa asegurarse de que los datos se mantengan sin cambios a lo largo de toda su vida. [9]

Si no tenemos un control de calidad de los datos, se pasa a la depuración de la base de datos. Para ello, se tiene que mirar los formatos de las variables (si tienen formato carácter o numérico) y que concuerden con lo que realmente se define en esa variable,

es decir, que no haya observaciones equivocadas. Para las variables con formato categórico, se tendrá que estudiar si sus categorías son correctas y que estén bien definidas. En caso de que no haya un número suficiente para la representación de determinadas categorías, recategorizar la variable para que en la variable objetivo de estudio (tipo de evento) se encuentre bien representado en todas las categorías de la variable. Y, por último, para las variables que tengan valores perdidos, en el caso de que se pierda mucha información al eliminar estas, imputar esos valores.

En cuanto a la buena calidad de los datos, los formatos de las variables y observaciones erróneas, explicarlo resulta más complicado ya que no hay una forma exacta. Por lo que se verá en el apartado del tratamiento de las variables aplicándolo directamente a la base de datos que se ha obtenido. Y en el siguiente apartado se hará de forma general una introducción a la imputación de valores perdidos para que el lector se ponga en situación y entienda todos los pasos posteriores aplicados a la base de datos seleccionada.

2.2. Metodología para el tratamiento de datos faltantes (missing).

"Los datos missing o faltantes son un grave problema a la hora de realizar cualquier estudio, ya que la mayoría de los análisis en estadísticas dan por hecho que la información está completa para todas las variables del análisis" [10]. La importancia de la imputación de los datos faltantes es porque en caso de que no se haga, la ausencia de algunas de las observaciones puede afectar de manera contundente a la muestra. Algunos procedimientos de SAS, utilizan toda la información en los datos para el cálculo de las estimaciones de los parámetros, como por ejemplo el PROC CORR. Aunque esto sea mejor que solo usar los datos completos puede llevar también a errores [11]. Por lo que sigue siendo una buena opción la imputación de los valores faltantes.

Cuando existe la falta de información, es necesario saber si estos datos están distribuidos aleatoriamente o si se pueden identificar algunas pautas. La evaluación de los datos ausentes nos puede evitar reducir la muestra o incluso tener sesgos potenciales.

"Hay que realizar una distinción entre mecanismos y patrones de perdida de datos. Los patrones describen los datos que son observados y los que están perdidos, mientras que los mecanismos describen el proceso mediante el cual se han dado esos valores perdidos" [12].

En casi todas las referencias sobre datos perdidos se distinguen tres mecanismos: [11]

- MCAR (Perdidos completamente al azar): Los datos missing de la variable que se estudie se dice que son completamente aleatorios si la presencia del dato perdido no depende del valor que tome dicha variable ni de ninguna otra variable que se tenga en el conjunto de datos. Este mecanismo es el más fácil de tratar. Un ejemplo es por ejemplo que se estén recogiendo datos de una encuesta sobre algunos aspectos de la cultura preventiva en la empresa en papel con la posterior perdida de los papeles donde se hicieron estas anotaciones.
- MAR (Perdidos al azar): Los datos se dicen que son perdidos al azar cuando la probabilidad de que aparezca un dato perdido en la variable Y no depende de dicha variable, pero si puede depender de una o más variables que estén en el fichero de datos.
- MNAR (Datos no perdidos aleatoriamente): En este mecanismo la probabilidad de pérdida de un dato puede depender tanto de la misma variable como de los valores observados de las demás variables.

En cuanto a los patrones de datos perdidos, no existe un acuerdo de tipos de patrones. Por lo que se hará una definición general de algunos tipos de patrones.

Un posible patrón es que falten valores únicamente en una variable. Otro patrón es el monótono, este tipo de patrón que es el que se va a buscar en el estudio de este trabajo, es aquel en el que los valores faltantes se encuentran ubicados en una parte específica de los datos que suele ser en la última parte. [12]

Una vez que se ha estudiado el grado de aleatoriedad y sus pautas en los datos faltantes se tiene que tomar la decisión de qué hacer con ellos. Existen dos opciones: eliminarlos o imputarlos.

La primera posibilidad, es la que se usa en la mayoría de los paquetes estadísticos y es eliminar todas aquellas observaciones que tengan valores perdidos. No es aconsejable su aplicación a no ser que el número de missing sea muy pequeño. También existen otros métodos como la eliminación según pareja (solo se trabaja con aquellas observaciones que tengan valores válidos para cada par de variables) o la eliminación de las observaciones o variables que contengan mayor número de valores perdidos. [11]

2.3. Métodos que imputan los datos faltantes.

Unas de las metodologías para el tratamiento de los datos faltantes es la imputación de estos a partir de las otras variables del fichero. Se distinguen dos tipos de imputación: la simple y la múltiple.

Según el tipo de aleatoriedad de los datos se deberían aplicar distintas técnicas de imputación (tabla 1). [11]

Aleatoriedad de los datos missing	**Tipo de variable**	**Método recomendado**
Monótono	Continua	Regresión
		Predicted Mean Matching
		Propensity Score
Monótono	Ordinal	Regresión Logística
Monótono	Nominal	Función Discriminante
MCAR	Continua	MCMC

***Tabla 1**: Grado de aleatoriedad junto con las técnicas de imputación*

2.3.1. Imputación simple.

Los métodos de imputación simple reemplazan los datos faltantes por un único valor. Los métodos más sencillos son aquellos que, por ejemplo, sustituyen el valor faltante por la media de la variable o, si se trata de una serie temporal, utilizando los valores adyacentes a él en el tiempo, etcétera.

Luego existen métodos de imputación multivariante, que normalmente se usan para la imputación múltiple, pero también se puede asignar un único valor mediante estos métodos. Que se explicará en el siguiente apartado.

2.3.2. Imputación múltiple.

La imputación múltiple es una metodología introducida por Rubin (1987) para el análisis de datos para el que algunos datos que se preveía recoger son valores perdidos.

Los métodos de imputación múltiple consisten en sustituir cada valor faltante por m conjuntos de valores, obteniéndose así m conjuntos de datos, lo que da lugar a m estimaciones con sus respectivas varianzas y errores estándares.

"Estos métodos permiten minimizar el sesgo o la pérdida de potencia estadística causada por la pérdida de datos con datos MCAR o MAR" [12].

El procedimiento de SAS PROC MI realiza métodos de regresión y MCMC (Markov Chain Monte Carlo) que aunque parten de la hipótesis de se tiene una distribución normal multivariante , son métodos robustos si esta hipótesis no se cumple cuando se tiene una estructura monótona. Que será el caso de la base de datos que se va a usar, ya que ninguna de las variables sigue una distribución normal en sus valores. [11]

Afortunadamente, la imputación múltiple puede utilizarse no sólo para variables continuas, sino también para variables binarias y categóricas.

Para variable continuas, existen varias elecciones de métodos de imputación sin importar los patrones de omisión, mientras que, para las variables categóricas, su imputación se ha limitado en mayor parte a la falta de datos faltantes monótonos en el

pasado. Sin embargo, en versiones anteriores a la 9.3 de SAS existe una manera para tratar los datos categóricos que faltan no monótonos, utilizando el método MCMC (Markov Chain Monte Carlo) para la imputación parcial de registros perdidos no monótonos. [8]

"Esto no es un óptimo, pero a menudo es aceptable porque, la mayoría de las veces, la cantidad de datos faltantes que no son monótonos es muy pequeña, y el impacto global de esta etapa de imputación parcial en el análisis en el punto temporal final del estudio será pequeño" [13].

El análisis con imputación múltiple se lleva a cabo generalmente en tres pasos:

1) Imputación: se generan los m conjuntos de datos imputados. Este paso se lleva a cabo en SAS utilizando el PROC MI.
2) Análisis: Cada uno de los m conjuntos de datos imputados se analiza por separado mediante el método que ha sido escogido, en este caso la regresión logística. Este paso puede ser implementado usando cualquier procedimiento analítico en SAS.
3) Agrupación: los resultados de los análisis realizados en el paso 2 se combinan en un resultado global. Este paso se realiza mediante el procedimiento de SAS PROC MIANALYZE.

En el paso 1, se van a encontrar varios pasos de imputación en el fichero. La imputación múltiple vendrá en el paso para hacer la estructura de datos perdidos monótona, y a partir de esta, se imputarán por un lado las variables continuas mediante el algoritmo EM y, por otro lado, las variables categóricas mediante regresión logística.

En este caso, se necesitaba tener los datos missing con estructura monótona y a eso es lo que le hemos dado varias imputaciones, pero si nuestros datos ya fueran monótonos, se podría hacer la imputación múltiple en el paso de la imputación de las variables continuas.

Se pueden usar varias técnicas de imputación de variables continuas en SAS. En este caso, se va a usar el algoritmo EM para la imputación de los datos perdidos de variables continuas, aunque también existe la técnica de regresión que es la que hace por defecto el PROC MI. Este método de imputación mediante el algoritmo EM es de gran

importancia a la hora de estimar valores ausentes en problemas multivariantes, sobre todo porque es menos exigente con la aleatoriedad de los datos missing. Se trata de un proceso iterativo donde se realizan sucesivamente dos pasos: Un paso de predicción y un paso de estimación, que se suceden de forma iterativa.

En cuanto a las variables de formato categórica, la imputación que se realizará será mediante regresión logística.

Todo ello, se hará mediante el procedimiento PROC MI.

En el paso 2, se ha optado por el análisis de una regresión logística. Donde la variable repuesta o dependiente es el tipo de evento y el resto de variables independientes. Este análisis se hará mediante el procedimiento PROC LOGISTIC, del cual se obtendrá una regresión logística por cada conjunto de datos imputados en el paso 1.

Por último, en el paso 3, se combinarán todos los análisis de regresión del paso 2 en uno solo mediante el procedimiento PROC MIANALYZE.

3. ANÁLISIS DEL FICHERO.

Normalmente para pruebas que son binarias se usan típicamente la regresión logística para modelar la variable dependiente. Para este apartado, se evalúa la relación de las covariables con la variable objetivo de estudio. [14]

Los modelos de regresión logística están basados en probabilidades y concretamente en logaritmos. Para saber si las variables independientes están relacionadas con la variable dependiente se calculan primero los coeficientes β que son los logit . Otra manera de expresar los logits o coeficientes es mediante los OR (odds ratio) y para ello se eleva al número exponencial a los logits o coeficientes. [14]

$$OR_i = \exp(\beta_i)$$

La regresión recurre a los odds ratio porque son medidas estandarizadas que permiten comparar el nivel de influencia o fortaleza de las variables independientes sobre la variable dependiente.

El OR (Odds Ratio) es una medida de asociación entre dos variables que indica la fortaleza de relación entre dos variables. En este caso, se relaciona cada una de las variables independientes del modelo con la variable objetivo de estudio, que es el tipo de evento. Cuando el OR es 1 indica que no hay asociación entre ambas variables.

Para la interpretación del OR se tienen en cuenta las siguientes cuestiones: [14]

- Los exponenciales de β son odds ratio y pueden compararse entre sí para saber qué variables tienen más influencia o está asociada de manera más fuerte.
- Cuando el exp(β) es mayor que 1 señala que un aumento de la variable independiente, aumenta los odds de que ocurra el evento (la variable dependiente tipo_r). Cuando el exp(β) es menor que 1 indica que un aumento de la variable inpendiente reduce los odds de que ocurra el evento (variable dependiente).
- Cuanto más se aleja de 1, más fuerte es la relación entre las dos variables.
- Cuando el odds ratio es menor de 1 es conveniente calcular su inversa para no equivocarse y poder comparar todo.

4. APLICACIÓN A UNA BASE DE DATOS.

4.1. Obtención de la BBDD.

La base de datos que se ha escogido ha sido proporcionada por una empresa que va a ser anónima en el presente trabajo, pero el fichero recoge todos los eventos ocurridos en tal empresa sobre accidentes e incidentes desde que se empezaron a gestionar todos los eventos en tal organización. En este conjunto de datos el número de eventos es adecuado para hacer cualquier análisis, sin embargo, existen eventos repetidos en la BBDD. Por lo tanto, para obtener una muestra lo más representativa posible de los eventos en la empresa, se han seleccionado solo aquellos eventos que no tengan ninguna vez repetida la variable identificadora, que en este caso se denomina código. Por otro lado, en el gráfico 1 se representa la evolución de los accidentes e incidentes a lo largo del tiempo del fichero obtenido. Se observa claramente, como en los años

anteriores a 2014, este inclusive, no se recogieron eventos de tipo incidente. Además, de que en el año 2018 el número de eventos recogidos no es representativo a la realidad, ya que el año aún no ha acabado y seguramente el número de eventos en este año sea mucho más alto.

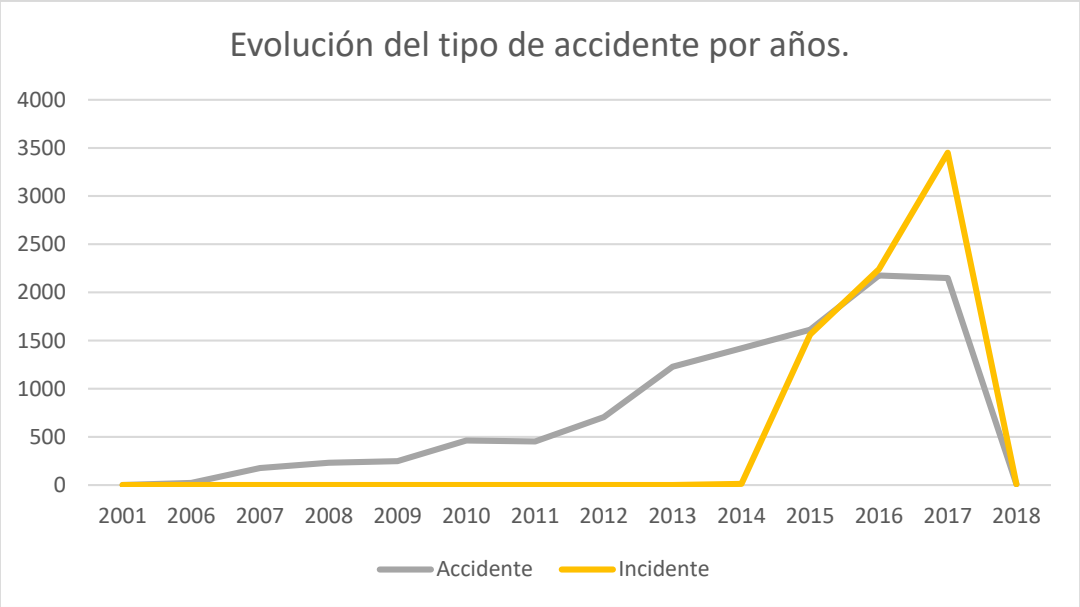

Ilustración 1: *Gráfico de la evolución según tipo de evento.*

Por lo tanto, se van a seleccionar solo las observaciones de los años 2015, 2016 y 2017. Hay 13192 eventos (5938 accidentes y 7254 incidentes) y 237 variables, la gran mayoría de carácter categórico, que suelen ser definiciones del evento. Como bien se ha mencionado en la teoría de la depuración de datos, una de las opciones para tener un fichero más eficiente es la eliminación de variables que tengan un número de valores faltantes muy elevado. Por ello, se han eliminado las variables que tengan un número de faltantes muy grande. Además, a juicio del investigador, tampoco se seleccionaron aquellas variables que definan solamente las observaciones de tipo accidente, para que las variables independientes en el estudio sean anteriores a un posible evento para poder predecirlo. Por lo tanto, de las 237 variables que se tiene, solo se han seleccionado 13 de ellas (incluida la variable objetivo de estudio y la variable identificadora) por ser las más representativas a la hora de predecir un posible evento. A continuación, se analizan de forma detallada cada una de ellas.

4.2. Tratamiento de la BBDD.

Antes de comenzar cualquier tipo de análisis es de vital importancia conocer a fondo la base de datos que se va a estudiar. De las variables seleccionadas para el estudio se va a realizar un análisis de cada una de ellas y su tratamiento.

- Tipo: Esta variable es una variable categórica dicotómica y será la variable objetivo de estudio. Es en la que se representa si el evento ha sido un accidente o un incidente.

  En el gráfico 1 ubicado en el apartado anterior, se observa como los accidentes han incrementado de un año a otro y en el siguiente se han mantenido. Y en cuanto al número de incidentes, estos han ido en aumento a lo largo de los 3 años seleccionados.

  Además, esta variable no tiene ningún valor perdido en el conjunto de datos. Hay, como bien se ha mencionado, 5938 accidentes y 7254 incidentes. Se codificará, para facilitar el estudio posterior, en una nueva variable llamada tipo_r, donde tomará el valor 1 si es un Accidente y el valor 0 si es un Incidente.
- Riesgo: Es una variable continua que mide el grado de riesgo de que ocurra tal evento.

  Se representa, a continuación, el histograma de esta variable ajustada a una distribución exponencial (gráfico 2). Se observa como el histograma sigue una forma muy parecida a una exponencial, por lo tanto, se podría decir que esta variable sigue una distribución de probabilidad exponencial. La mayor frecuencia de esta variable se encuentra en el valor 0, pero como se puede observar su rango se extiende hasta el valor 300, pero en menor medida.

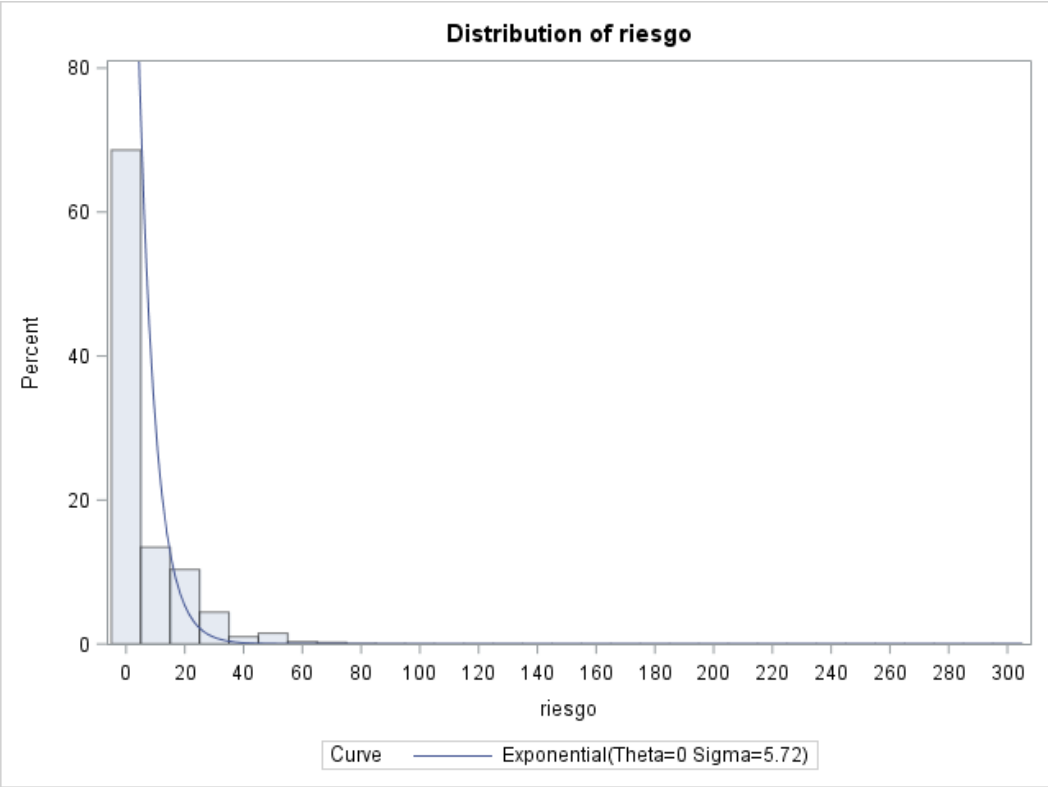

Ilustración 2: *Histograma de la variable riesgo.*

- Puntuación: Es una variable continua que mide la puntuación de las medidas preventivas en la empresa.

  Para conocer más esta variable, se van a representar (tabla 2), en este caso, unos estadísticos descriptivos básicos sobre ella.

Basic Statistical Measures			
Location		**Variability**	
Mean	73.03336	**Std Deviation**	31.60917
Median	70.00000	**Variance**	999.13993
Mode	50.00000	**Range**	385.00000
		Interquartile Range	40.00000

Tabla 2: *Estadísticos descriptivos de la variable puntuación.*

La media de la puntuación de las medidas preventivas está en una puntuación de 73 puntos, no es un valor muy alto teniendo en cuenta que el máximo valor de esta es de 385 puntos en medidas preventivas. La puntuación más repetida en la base de datos es de 50 puntos, es decir, su moda.

En el siguiente histograma se representa dicha variable, que ayudará a comprender su distribución.

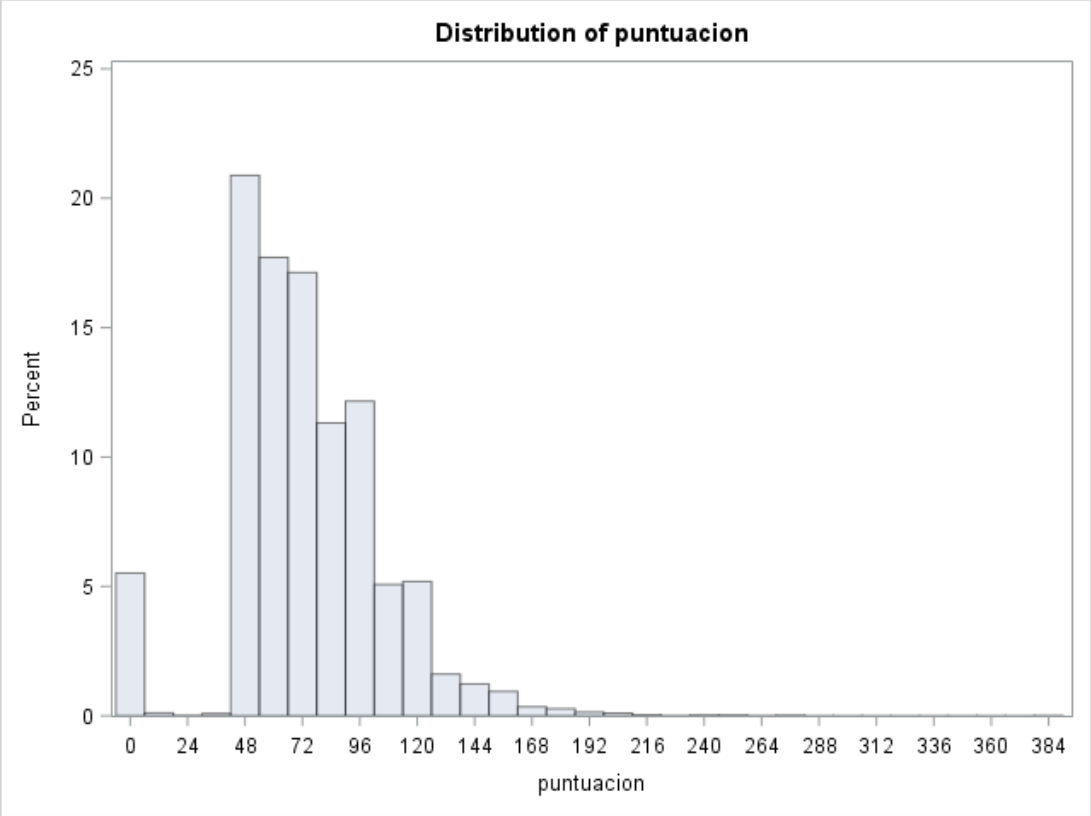

***Ilustración 3**: Histograma de la variable puntuacion.*

- Sexo: Esta variable es categórica y representa el sexo del que ha sufrido dicho evento.

  A continuación, se representa una tabla de frecuencia con los distintos valores que toma la variable (tabla 3).

Sexo				
Sexo	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	406	7.34	406	7.34
Hombre	4385	79.31	4791	86.65
Mujer	738	13.35	5529	100.00
Frequency Missing = 7663				

***Tabla 3:** Tabla de frecuencias de la variable sexo.*

Se observa que hay una categoría de la variable que es 0, que se va a tomar como un valor perdido, ya que no se sabe si pertenecería a la categoría 'Hombre' o a la categoría 'Mujer'. Por lo tanto, se va a recategorizar esta variable, dándole el valor missing a la categoría cero. Además, para un mejor tratamiento de las variables, se van a codificar las categorías 'Hombre' y 'Mujer'.

De esta manera, la variable sexo va a pasar a llamarse sexo_r, donde tomará el valor 1 si es Mujer y el valor 0 si es Hombre. Quedaría de la siguiente forma:

sexo_r	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4385	85.59	4385	85.59
1	738	14.41	5123	100.00
Frequency Missing = 8069				

Tabla 4: *Tabla de frecuencias de la variable sexo_r.*

Se observa como el número de hombres y mujeres es el mismo, pero el número de missing ha aumentado, ya que se le han añadido los 406 ceros. En total se tienen 8069 valores perdidos en esta variable. Teniendo en cuenta que el total de las observaciones es de 13192, más de la mitad de las observaciones de esta variable no tiene valor.

- Diasemana: Es una variable categórica que indica el día de la semana en el que tuvo lugar el evento.

  A continuación, se muestra la tabla 5 de frecuencias de los distintos días de la semana de dicha variable:

DiaSemana				
DiaSemana	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
domingo	423	3.21	423	3.21
jueves	2599	19.70	3022	22.91
lunes	2242	17.00	5264	39.90
martes	2532	19.19	7796	59.10
miércoles	2549	19.32	10345	78.42
sábado	801	6.07	11146	84.49
viernes	2046	15.51	13192	100.00

Tabla 5: *Tabla de frecuencias de la variable diasemana.*

Para la facilidad del tratamiento de la variable, se va a codificar. Es decir, a cada día de la semana se va a sustituir por un valor numérico. Donde el 1 será el Lunes y el 7 el Domingo, en orden de los días de la semana, llamada diasemana_r.

En el siguiente gráfico (ilustración 4), se observa como la mayoría de los incidentes ocurren entre semana, de lunes a viernes, teniendo mayor frecuencia de accidentes los jueves y de incidentes los martes. En el fin de semana no se observan tantos eventos, ya que son los días que los empleados no suelen trabajar, por lo que es lógico.

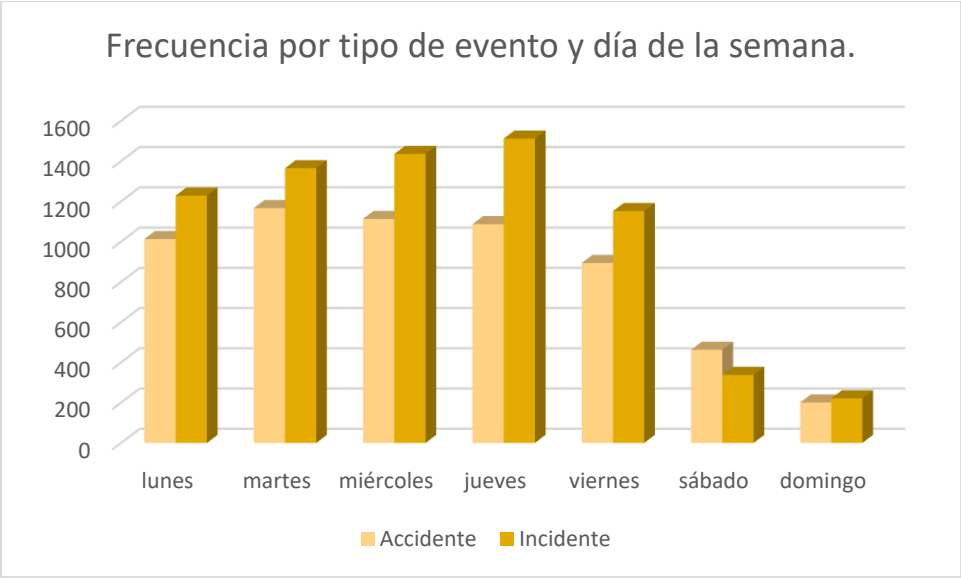

Ilustración 4: *Histograma de la variable diasemana según el tipo de evento.*

Además, como se puede observar, esta variable no tiene ningún valor perdido.

- Circunstancia: mide el grado de la circunstancia asociada al evento, es decir, es el conjunto de circunstancias que se producen alrededor del evento dado. Esta variable es difícil de tratar. El formato que tiene en el fichero de datos es categórico, sin embargo, los valores de esta son números. Mediante un estudio de frecuencias de los valores de esta, se obtiene que uno de los eventos en esta variable se encuentra definido como 'Menos Grave' y otra de las observaciones en esta variable se encuentra definida como '#N/A'.

  Por lo tanto, se va a transformar esta variable, de manera que si toma el valor categórico se va a pasar a un valor missing y el resto de valores se dejarán tal cual, cambiando el formato de esta a numérico. La cual se pasará a llamar circunstancia_r.

  A continuación, se muestra la tabla 6 de los estadísticos básicos de esta nueva variable.

Basic Statistical Measures			
Location		**Variability**	
Mean	66.72885	**Std Deviation**	25.38130
Median	65.00000	**Variance**	644.21026
Mode	50.00000	**Range**	230.00000
		Interquartile Range	30.00000

Tabla 6: *Estadísticos descriptivos de la variable circunstancia_r.*

Se observa como el número de circunstancias asociadas al evento más frecuente es de 50. Por otro lado, en media se producen más de 66 circunstancias asociadas. El número máximo de circunstancias asociadas es de 230.

En la representación del histograma siguiente de la variable se observa de forma gráfica los valores obtenidos en la tabla.

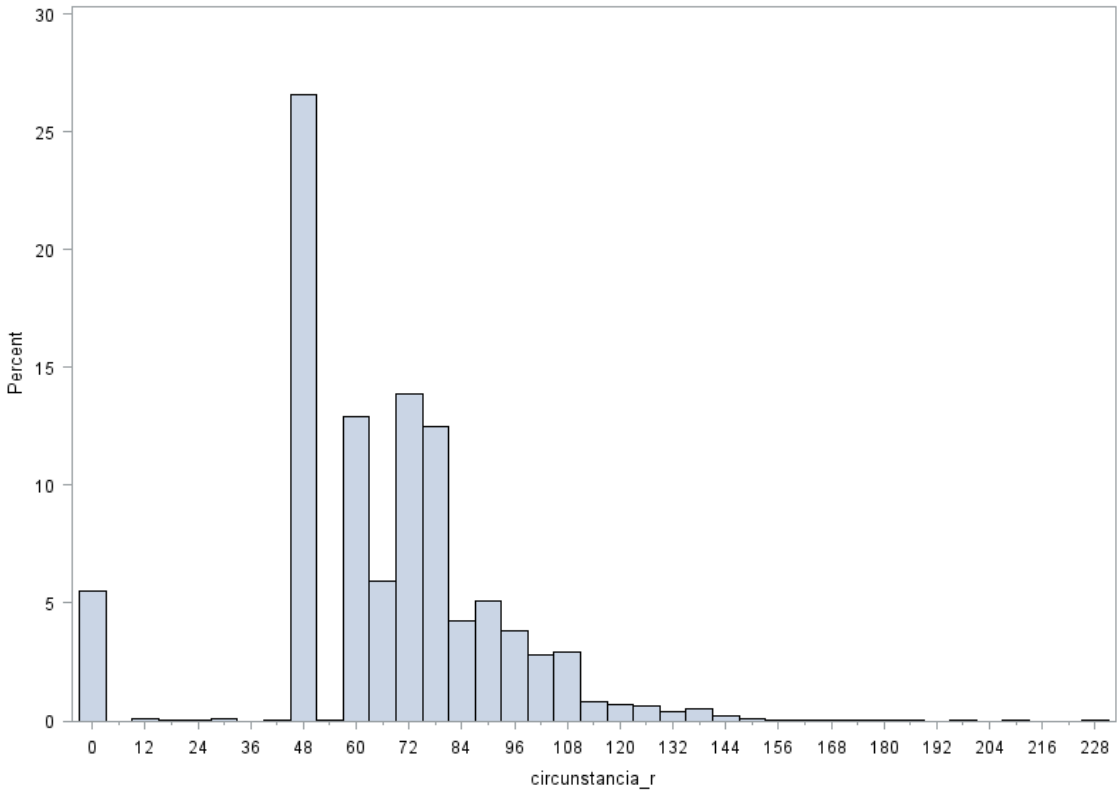

Ilustración 5: *Histograma de la variable circunstancia_r.*

- Clasificación: Esta variable es de formato categórico y mide la gravedad del evento.

  A continuación, se muestra la tabla 7 de frecuencias de la variable en cuestión.

clasificacion	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#N/A	1	0.01	1	0.01
Grave	1516	11.51	1517	11.52
Leve	3856	29.27	5373	40.78
Menos Grave	370	2.81	5743	43.59
Menos grave	7293	55.36	13036	98.95
Muy Grave	30	0.23	13066	99.18
Muy grave	108	0.82	13174	100.00
Frequency Missing = 18				

Tabla 7: *Tabla de frecuencias de la variable clasificacion.*

Se observa como las categorías que toma esta variable no se encuentran definidas correctamente, ya que categorías iguales se toman como distintas por estar escritas en minúscula o mayúscula. Por otro lado, también se tiene el valor '#N/A' que no es una categoría como tal.

Por todo ello, se va a transformar la variable, de tal manera que tome un valor numérico según la gravedad del evento y la categoría que no mide el grado se transforma a valor perdido.

Primero se tiene que pasar las categorías de la variable todo a minúsculas o a mayúsculas. En este caso, se pasarán a minúsculas con la sentencia lowcase() de SAS. Una vez que se tiene todo de la misma forma, se transforma la variable, que pasará a denominarse clasificación_r.

Esta nueva variable tomará el valor 1 si el evento ha sido clasificado como leve, el valor 2 si ha sido clasificado como menos grave, el valor 3 si ha sido clasificado como grave y, por último, el valor 4 si ha sido clasificado como muy grave. Y habrá un valor perdido más que antes, esto es porque se le ha añadido la observación que tomaba el valor '#N/A'.

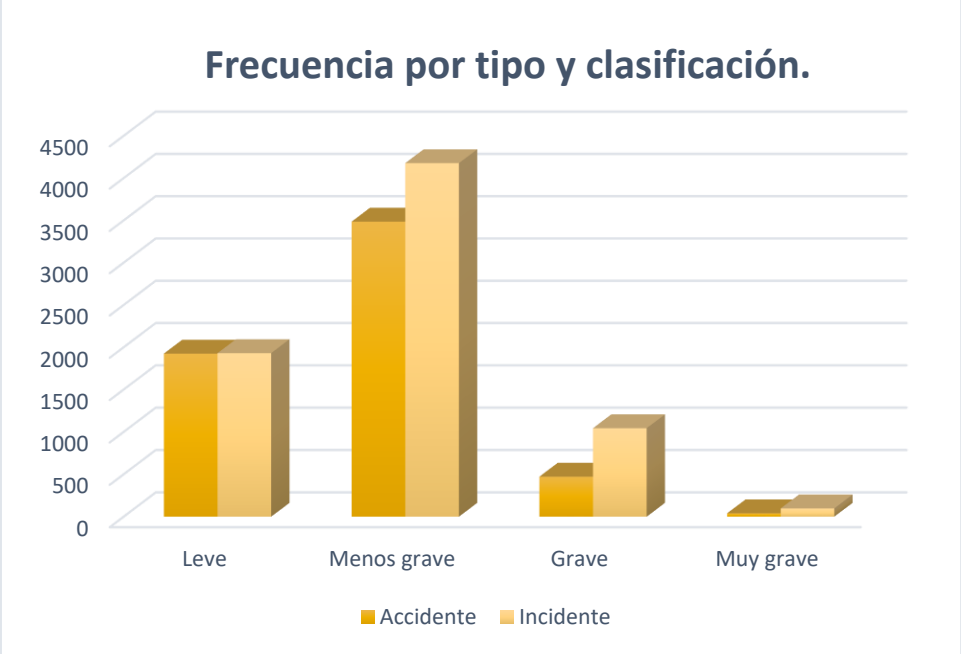

Ilustración 6: *Histograma de la variable clasificacion_r según tipo de evento.*

En la ilustración 6, se observa cómo se distribuye esta variable nueva en función del tipo de evento. En mayor medida se encuentran clasificadas como menos graves tanto los incidentes como accidentes. Y existen más incidentes clasificados con las categorías más graves.

- Edad: Esta variable representa la edad del empleado que ha sufrido tal evento. Aquí se vuelve a encontrar un problema, que no están bien recogidos los datos de la variable, tiene tanto observaciones categóricas como numéricas. Es decir, se han encontrado en esta variable eventos en los que su edad se encuentra definida en un intervalo de edad, como por ejemplo 'De 18 a 25 años', etc,.. y otras en el que se encuentra su edad exacta. Por este motivo, para que toda la variable se encuentre en el mismo formato, los eventos que tienen edades definidas como un intervalo se van a transformar en numéricas dándole el valor intermedio del intervalo de edad.

  Se crea la variable edad_r donde van a tomar valores numéricos todas las observaciones.

  Aunque se ha transformado la variable para que esta sea continua, el siguiente gráfico (ilustración 7) se va a representar en intervalos de edad, para una mejor interpretación.

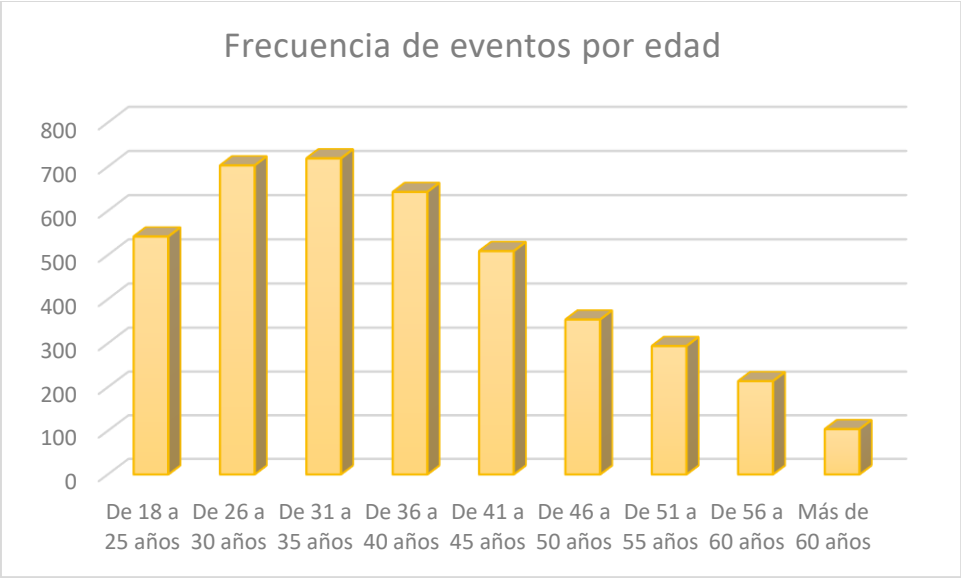

Ilustración 7: *Histograma de la variable edad por intervalos.*

Se observa como personas comprendidas entre los 18 y 40 años son los que sufren más eventos en la empresa. Siendo los empleados de mayor edad los que menos eventos sufren.

Además, el número de valores perdidos en esta variable es de los más altos, 9110 missing. La mayor parte de esta variable tiene valor perdido. Indagando más en ella, la mayor parte de los perdidos se encuentra cuando el tipo de evento es un incidente. Por lo tanto, en el gráfico anterior puede ser también la representación de la edad de los empleados que han sufrido solo accidentes.

- Antigüedad: Esta variable representa, en años, el tiempo que el empleado que sufrió un evento lleva trabajando en la empresa.

  Con esta variable se vuelve a encontrar el mismo problema que con la variable edad. Algunas observaciones se encuentran definidas como un intervalo, como por ejemplo 'Menos de 1 año', 'Entre 1 y 5 años', etc… y otras con la antigüedad exacta.

  Por lo que se volverá a realizar el mismo procedimiento que antes. Se pasará la variable a numérica, transformando las observaciones definidas en intervalos al valor intermedio de tal intervalo, que pasará a llamarse antigüedad_r.

  A continuación, se muestra el gráfico (ilustración 8) por intervalos de antigüedad en la empresa para todos los tipos de evento.

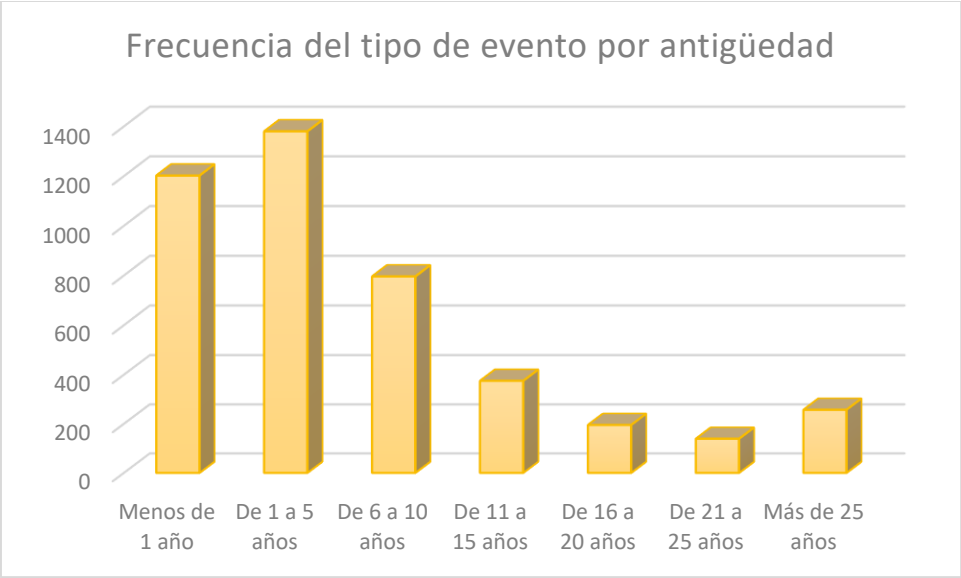

***Ilustración 8**: Histograma de la antiguedad_r por intervalos.*

Se observa como los que sufren más eventos son los que llevan menos tiempo en la empresa, debido a la inexperiencia en su trabajo. Por otro lado, los que llevan más de 25 años en la empresa, en comparación con el resto, sufren más de 200 evento, pero este es debido a despistes en su trabajo por ser más mayores en edad (ya que a mayor antigüedad más edad tienes).

Realizando un estudio más exhaustivo, según el tipo de evento, y se vuelve a notar que en los incidentes tampoco se definen la antigüedad.

El número de perdidos es de 8848, una vez más, más de la mitad de los valores de esta variable.

- Repetición: Esta variable es categórica y representa la probabilidad de repetición del evento.

  A continuación, se muestra la tabla 8 de frecuencias de esta variable.

repeticion	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Alta	935	7.85	935	7.85
Baja	4813	40.42	5748	48.28
Media	2481	20.84	8229	69.12
Muy Baja	3360	28.22	11589	97.34
Muy alta	317	2.66	11906	100.00
Frequency Missing = 1286				

***Tabla 8:** Tabla de frecuencias de la variable repeticion.*

Se observa como las probabilidades de repetición que más se repiten en el fichero es Baja y muy baja, es decir, normalmente los eventos no suelen repetirse.

Para la mejora del tratamiento, se va a transformar la variable a formato numérico, dándole el valor 1 a la categoría muy baja y el valor 5 a la categoría muy alta, en orden de probabilidad de repetición, llamada repeticion_r.

En esta variable se tienen 1286 valores perdidos.

En el siguiente gráfico (ilustración 9), en el que se representa la frecuencia del tipo de evento distribuido según la repetición, se observa como los dos tipos de evento son en mayor medida de repeticiones de probabilidad muy baja o baja.

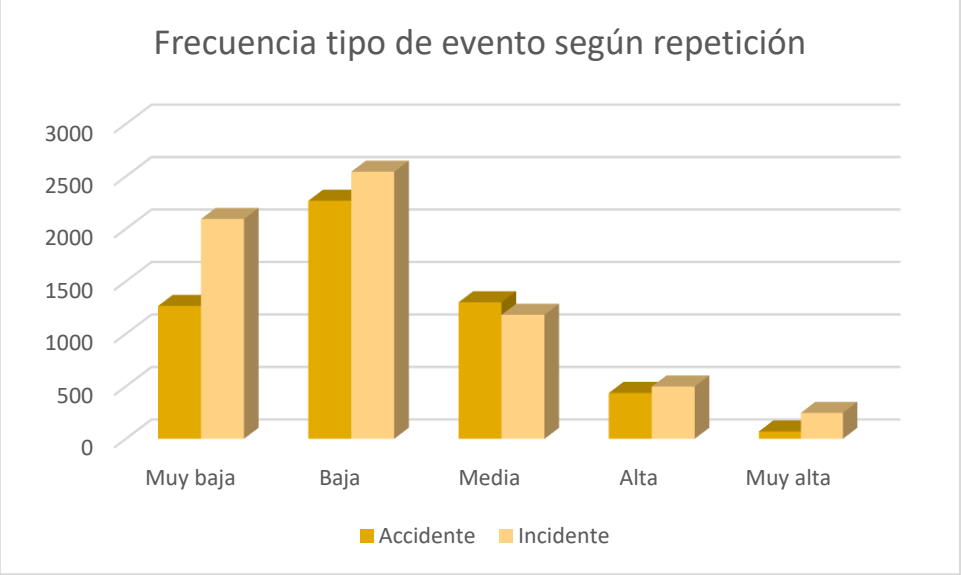

Ilustración 9: *Histograma de la variable repeticion_r según tipo de evento.*

Por otro lado, casi 300 incidentes tienen una probabilidad de repetición muy alta y con probabilidad alta tanto incidentes como accidentes tienen casi 500.

- Accidentado: Esta variable es categórica y representa el tipo de empleado que ha sufrido el evento.

  Las categorías de accidentado se representan en la siguiente tabla de frecuencias.

Accidentado				
Accidentado	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Ambiental	185	1.40	185	1.40
Contratista	8377	63.50	8562	64.90
Histórico	719	5.45	9281	70.35
Propio	3474	26.33	12755	96.69
Propio/Contratista (mixto)	304	2.30	13059	98.99
Tercero	133	1.01	13192	100.00

Tabla 9: *Tabla de frecuencias de la variable accidentado.*

Las categorías son el tipo de contrato que tiene el empleado en el momento del evento. Se observa que el número que aparece en cada categoría es muy distinto para cada una, por lo que se va a proceder, tanto a codificar la variable para la facilidad del tratamiento como a recategorizarla.

Se va a recategorizar la variable de manera que el accidentado 'Tercero', 'Ambiental', 'Histórico' y 'Propio/Contratista (mixto)' se agrupan en una misma categoría que se va a denominar 'Otros o mixto'. El resto de categorías se va a quedar igual, ya que se encuentran bien representadas.

Una vez recategorizada, se va a codificar la variable, se le da el valor 1 a tipo de accidentado 'Contratista', que son aquellos contratados por la empresa para un determinado trabajo; el valor 2 si es 'Propio', que son aquellos empleados dentro de la empresa y el valor 3 si es 'Otro o mixto'. Por lo tanto, quedaría de la siguiente manera, creando la variable accidentado_r.

accidentado_r	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Contratista	8377	63.50	8377	63.50
Propio	3474	26.33	11851	89.83
Otro o mixto	1341	10.17	13192	100.00

Tabla 10: *Tabla de frecuencias de la variable accidentado_*

- Año: Esta variable se va a tomar como continua y representa el año donde tuvo lugar el accidente, que se llama a_o.

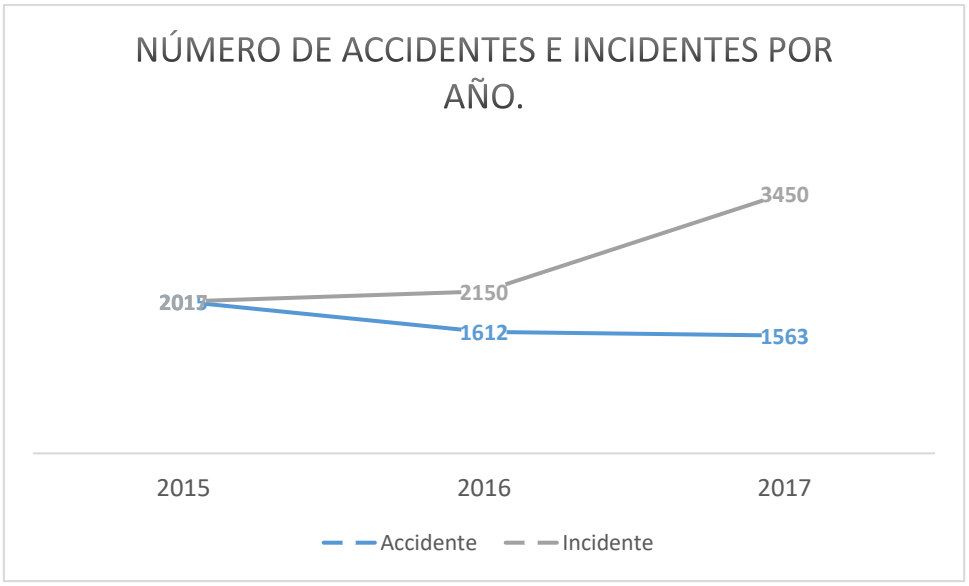

Ilustración 10: *Gráfico de la evolución de los eventos en los 3 años.*

Se observa como los incidentes en la empresa van en crecimiento y cada vez hay más, y que los accidentes disminuyen con el tiempo, aunque de forma pausada.

4.3. Imputación de los valores faltantes.

Una vez que se ha "limpiado" la base de datos y estudiado sus valores perdidos, se van a aplicar los 3 pasos que normalmente se siguen para la imputación múltiple explicados anteriormente.

Paso 1: Imputación.

El procedimiento de SAS PROC MI, ofrece varios métodos de imputación tanto de variable categóricas como continuas. La elección del método a utilizar depende de la monotonía de los datos faltantes. [15]

Al coger la base de datos, se vio que sus datos faltantes no seguían un patrón monótono. Para analizar esto se usó la siguiente sintaxis:

```
proc mi data=tfg.acc_f seed=33 nimpute=0;
var tipo a_o codigo_modificado riesgo puntuacion sexo_r diasemana_r circunstancia_r
clasificacion_r edad_r antiguedad_r repeticion_r accidentado_r;
run;
```

Igualando la sentencia nimpute a 0, se hace un estudio de la monotonía de los datos perdidos, sin realizar ninguna imputación. En la sentencia var se han ordenado las variables, poniendo primero las que no tienen ningún missing y después las que si lo tienen.

La tabla que se obtiene podría decirse que se divide en dos. La primera parte se muestran las variables en el orden que se le ha indicado en la sentencia var y en las filas se enumera distintos patrones de datos perdidos. En las celdas, pueden aparecer tres tipos distintos de signo. Si aparece una 'X' es que la variable se observa en el grupo correspondiente, si se observa un '.' es que la variable falta y si se observa un 'O' significa que la variable falta y no se imputará. En la segunda parte de la tabla se representan las medias de las variables específicas del grupo, que en este caso no se van a mostrar en el documento.

Modelos de datos ausentes														
Grupo	**tipo_r**	**a_o**	**diasemana_r**	**accidentado_r**	**Riesgo**	**puntuacion**	**sexo_r**	**circunstancia_r**	**clasificacion_r**	**edad_r**	**antiguedad_r**	**repeticion_r**	**Frec**	**Porcentaje**
1	X	X	X	X	X	X	X	X	X	X	X	X	3691	27.98
2	X	X	X	X	X	X	X	X	X	X	X	.	81	0.61
3	X	X	X	X	X	X	X	X	X	X	.	X	160	1.21
4	X	X	X	X	X	X	X	X	X	X	.	.	7	0.05
5	X	X	X	X	X	X	X	X	X	.	X	X	355	2.69
6	X	X	X	X	X	X	X	X	X	.	X	.	22	0.17
7	X	X	X	X	X	X	X	X	X	.	.	X	647	4.90
8	X	X	X	X	X	X	X	X	X	.	.	.	23	0.17
9	X	X	X	X	X	X	X	X	.	.	X	X	1	0.01
10	X	X	X	X	X	X	X	.	X	X	X	X	2	0.02
11	X	X	X	X	X	X	.	X	X	X	X	X	15	0.11
12	X	X	X	X	X	X	.	X	X	X	X	.	1	0.01
13	X	X	X	X	X	X	.	X	X	.	X	X	41	0.31
14	X	X	X	X	X	X	.	X	X	.	X	.	10	0.08
15	X	X	X	X	X	X	.	X	X	.	.	X	6855	51.96
16	X	X	X	X	X	X	.	X	X	.	.	.	1109	8.41
17	X	X	X	X	X	.	X	.	.	.	.	X	1	0.01
18	X	X	X	X	X	.	.	.	.	.	.	.	1	0.01
19	X	X	X	X	.	X	X	X	X	X	X	X	104	0.79
20	X	X	X	X	.	X	X	X	X	X	X	.	4	0.03
21	X	X	X	X	.	X	X	X	X	X	.	.	2	0.02

22	X	X	X	X	.	X	X	X	X	.	X	X	2	0.02
23	X	X	X	X	.	X	X	X	X	.	.	X	7	0.05
24	X	X	X	X	.	X	X	X	X	.	.	.	1	0.01
25	X	X	X	X	.	X	.	X	X	X	X	X	1	0.01
26	X	X	X	X	.	X	.	X	X	.	X	X	1	0.01
27	X	X	X	X	.	X	.	X	X	.	.	X	10	0.08
28	X	X	X	X	.	X	.	X	X	.	.	.	8	0.06
29	X	X	X	X	.	.	X	X	X	X	.	X	1	0.01
30	X	X	X	X	.	.	X	.	X	X	X	X	9	0.07
31	X	X	X	X	.	.	X	.	X	X	X	.	2	0.02
32	X	X	X	X	.	.	X	.	.	X	X	X	1	0.01
33	X	X	X	X	.	.	.	.	X	X	X	X	1	0.01
34	X	X	X	X	.	.	.	.	X	.	.	.	1	0.01
35	X	X	X	X	.	.	.	.	.	.	.	X	1	0.01
36	X	X	X	X	.	.	.	.	.	.	.	.	14	0.11

Tabla 11: *Patrón de perdidos.*

Se observa en la salida (la tabla 11) que incluso con el reordenamiento de las variables no sería posible producir un patrón monótono de datos perdidos. Las variables con todas las observaciones completas son tipo_r, a_o, diasemana_r y accidentado_r. Estas variables serán las que se coloquen en primera posición en la sentencia var al realizar las imputaciones, por ello el reordenamiento.

En este caso, con el fichero de datos utilizado, se han obtenido 36 patrones de datos perdidos. En ningún caso aparece el valor 'O' en ninguna casilla, por lo que todos los datos faltantes pueden ser imputados. Sin embargo, se observa claramente que nuestro fichero de datos no tiene un patrón de datos perdidos monótono. Ya que cada patrón obtenido no sigue un orden de valores completos con los perdidos. Pero si podemos saber de cuantas observaciones completas disponemos que son 3691 observaciones. Por lo tanto, que nuestros datos no tengan un patrón monótono significa que los valores de datos faltantes de las variables no se concentran en una parte específica del conjunto de datos.

Como se ha visto, estas variables no siguen el patrón monótono. Y la cantidad de datos que no es monótona es un poco elevada, por lo que se hará una imputación múltiple para crear la monotonía de los datos faltantes. Este patrón monótono es necesario debido a la categórica naturaleza de las variables con datos faltantes con el fin de utilizar un método de imputación correcto para estas.

Para ello, primero se va a ver en la tabla 12, el número de datos perdidos, el número de datos completos, el valor mínimo, el valor máximo, la media y la desviación típica de cada una de las variables que se van a incluir en el estudio. Esto se obtiene mediante la siguiente sentencia.

```
proc means data = tfg.acc_f nmiss N min max mean std;
var _numeric_;
run;
```

Al lado del fichero que se quiere analizar tfg.acc_f se incluyen las sentencias que se quieren obtener de cada variable incluida en la sentencia var. En este caso, se ha puesto la opción _numeric_ en la sentencia var, ya que todas las variables tienen un formato numérico al haber variables continuas y las que son categóricas se han codificado en números.

Variable	Número de valores ausentes	N	Mínimo	Máximo	Media	Dev std
a_o	0	13192	2015.00	2017.00	2016.18	0.7946274
Riesgo	170	13022	0	300.0000000	5.7153279	11.9431869
Puntuación	32	13160	0	385.0000000	73.0333587	31.6091748
sexo_r	8069	5123	0	1.0000000	0.1440562	0.3511810
diasemana_r	0	13192	1.0000000	7.0000000	3.2857793	1.6331466
circunstancia_r	33	13159	0	230.0000000	66.7288548	25.3812974
clasificacion_r	19	13173	1.0000000	4.0000000	1.8433159	0.6520660
edad_r	9110	4082	18.0000000	70.0000000	37.5536502	10.9289518
antiguedad_r	8848	4344	0	60.0000000	7.2879834	8.8179940
repeticion_r	1286	11906	1.0000000	5.0000000	2.1631110	1.0088745
accidentado_r	0	13192	1.0000000	3.0000000	1.4666465	0.6724781
tipo_r	0	13192	0	1.0000000	0.4501213	0.4975248

Tabla 12: *Estadísticos descriptivos del fichero completo.*

De esta manera, se podrá conocer los valores de las variables para poder crear el patrón monótono.

Debido a la naturaleza categórica de algunas de las variables, en la siguiente sintaxis se imputan 5 conjuntos de datos con suficientes valores de datos para producir un patrón monótono de datos faltantes (mcmc impute=monotone). Los valores imputados se redondean y limitan durante la imputación, de manera que los valores imputados coincidan con el formato de los valores observados. Los valores de las sentencias round, min y max corresponden a las variables enumeradas en la sentencia var. De ahí la necesidad de sacar la Tabla 12.

```
proc mi nimpute=5 data=tfg.acc_f out=tfg.acctipo_monotone seed=33
 round=1   1   1  1     1  1    1  1
 min=  0   0   0  0     1  18   0  1
 max= 300 385 1  230   4  70   60  5;
 mcmc impute=monotone;
var riesgo puntuacion sexo_r circunstancia_r clasificacion_r edad_r antiguedad_r
repeticion_r ;
run ;
```

Una vez que se ha ejecutado la sintaxis anterior, podemos volver a ejecutar un PROC MI con la sentencia nimpute=0 del fichero de datos creado antes, en la que se obtiene la tabla 13 de patrón de los perdidos, con la siguiente sintaxis.

```
proc mi data=tfg.acctipo_monotone seed=33 nimpute=0;
var tipo_r a_o diasemana_r accidentado_r riesgo puntuacion sexo_r circunstancia_r
clasificacion_r edad_r antiguedad_r repeticion_r;
run;
```

Missing Data Patterns														
Group	tipo_r	a_o	diasemana_r	accidentado_r	riesgo	Puntuacion	sexo_r	circunstancia_r	clasificacion_r	edad_r	antiguedad_r	repeticion_r	Freq	Percent
1	X	X	X	X	X	X	X	X	X	X	X	X	59530	90.25
2	X	X	X	X	X	X	X	X	X	X	X	.	600	0.91
3	X	X	X	X	X	X	X	X	X	X	.	.	45	0.07
4	X	X	X	X	X	X	X	X	X	.	.	.	5710	8.66
5	X	X	X	X	X	.	.	.	.	.	.	.	5	0.01
6	X	X	X	X	.	.	.	.	.	.	.	.	70	0.11

Tabla 13: *Patrones perdidos con estructura monótona.*

Se observa en la tabla 13 como los datos faltantes ya se encuentran al final del estudio, por lo tanto, el patrón ya es monótono.

Ahora se puede empezar con la imputación de los valores perdidos del fichero creado con estructura monótona.

- **Imputación de las variables continuas**

En este paso se muestra como ejecutar la segunda imputación utilizando el conjunto de datos producido anteriormente. En este caso, se va a realizar una imputación de las variables continuas en los 5 conjuntos creados antes, que se encuentran en el fichero llamado acctipo_monotone.

Como se ha estudiado, el mecanismo de patrones monótonos en la base de datos es importante para elegir el método de imputación.

En el procedimiento PROC MI, el método de imputación por defecto es el MCMC. Se sabe que cuando el conjunto de datos no tiene una estructura monótona de los valores faltantes, el método a usar para la imputación sería mediante el algoritmo EM. Aunque en este caso, se haya conseguido una estructura monótona, se va a utilizar este método para la imputación, ya que mediante este se obtienen mejores resultados en cuanto al rango de valores que debe tener la variable. [16]

En la siguiente sintaxis, se ha imputado el fichero de los 5 conjuntos de datos de la imputación anterior, pero una sola vez, mediante el método MCMC, con los valores iniciales estimados por el algoritmo EM. Esta imputación es solamente para las variables continuas, que se encuentran indicadas en la sentencia var.

```
proc mi data=tfg.acctipo_monotone seed=33 simple nimpute=1
OUT=acc_im;
em itprint out=tfg.acc_em;
var   riesgo puntuacion  circunstancia_r  edad_r antiguedad_r  ;
run;
```

La opción SIMPLE en el procedimiento calcula los estadísticos univariantes para las 5 variables con los datos no missing en cada una de ellas.

Univariate Statistics							
Variable	N	Mean	Std Dev	Minimum	Maximum	Missing Values	
						Count	Percent
Riesgo	65890	5.67927	11.88194	0	300.00000	70	0.11
puntuacion	65885	73.00964	31.61498	0	385.00000	75	0.11
circunstancia_r	65885	66.68950	25.39915	0	230.00000	75	0.11
edad_r	60175	39.06912	10.41775	18.00000	70.00000	5785	8.77
antiguedad_r	60130	9.28199	7.76710	0	60.00000	5830	8.84

Tabla 14: *Estadísticos univariantes sin tener en cuenta los missings.*

Y la matriz de correlaciones con el conjunto de pares completos.

Pairwise Correlations					
	Riesgo	puntuacion	circunstancia_r	edad_r	antiguedad_r
Riesgo	1.000000000	0.619444183	0.303415527	0.029428708	0.045900456
puntuacion	0.619444183	1.000000000	0.929820275	-0.053905435	-0.018610513
circunstancia_r	0.303415527	0.929820275	1.000000000	-0.081786786	-0.050702641
edad_r	0.029428708	-0.053905435	-0.081786786	1.000000000	0.391905881
antiguedad_r	0.045900456	-0.018610513	-0.050702641	0.391905881	1.000000000

Tabla 15: *Matriz de correlaciones con el conjunto de pares completos.*

Los siguientes valores son utilizados como valores iniciales para el algoritmo EM.

Initial Parameter Estimates for EM						
TYPE	_NAME_	Riesgo	Puntuacion	circunstancia_r	edad_r	antiguedad_r
MEAN		5.679268	73.009638	66.689504	39.069115	9.281989
COV	Riesgo	141.180546	0	0	0	0
COV	Puntuación	0	999.506661	0	0	0
COV	circunstancia_r	0	0	645.116647	0	0
COV	edad_r	0	0	0	108.529590	0
COV	antiguedad_r	0	0	0	0	60.327838

Tabla 16: *Valores iniciales para el algoritmo EM.*

La opción ITPRINT en la sentencia EM nos muestra las iteraciones de este algoritmo con el valor de la verosimilitud de cada iteración y las medias estimadas.

EM (MLE) Iteration History					
Iteration	-2 Log L	puntuacion	circunstancia_r	edad_r	antiguedad_r
0	2054004	73.009638	66.689504	39.069115	9.281989
1	1734376	73.009638	66.689504	39.069115	9.281989
2	1734105	73.018919	66.693157	39.166259	9.337549
3	1734094	73.018948	66.693168	39.205410	9.359219
4	1734092	73.018948	66.693168	39.220198	9.367289
5	1734092	73.018948	66.693168	39.225802	9.370329
6	1734092	73.018948	66.693168	39.227934	9.371483
7	1734092	73.018948	66.693168	39.228746	9.371923
8	1734092	73.018948	66.693168	39.229056	9.372091
9	1734092	73.018948	66.693168	39.229174	9.372155
10	1734092	73.018948	66.693168	39.229219	9.372179
11	1734092	73.018948	66.693168	39.229236	9.372188
12	1734092	73.018948	66.693168	39.229242	9.372192

Tabla 17: *Iteraciones del algoritmo EM.*

Los valores de la última iteración son entonces los valores estimados con el algoritmo EM.

El fichero obtenido en out=acc_im es el fichero que tiene los 5 conjuntos con los valores imputados.

Out=tfg.acc_em es el fichero que contiene los valores imputados mediante el algoritmo EM. Luego existe también la opción OUTEM=data que crea un fichero con los estadísticos estimados con el algoritmo EM: Medias y covarianzas.

- **Imputación de las variables categóricas.**

Este paso muestra la tercera imputación y última, utilizando el conjunto de datos último creado tfg.acc_em que genera el conjunto de datos imputado tfg.acc_full. Pasa lo mismo que con la imputación anterior, se sigue teniendo 5 conjuntos de datos distintos de la primera imputación realizada por eso aquí solo se va a hacer una única imputación para las variables categóricas para completar la información.

En la sintaxis, se ha usado monotone logistic requiere que el procedimiento PROC MI realice 3 regresiones logísticas para primero imputar la variable sexo_r mediante el resto, luego la imputación de la variable clasificacion_r utilizando sexo_r y, por último, imputar la variable repeticion_r usando las dos anteriores. En la sentencia class se tienen que incluir todas las variables categóricas, en este caso solo se incluyen aquellas que se van a imputar ya que el resto de categóricas tiene todos los valores completos.

```
proc mi data=tfg.acc_em nimpute=1 out=tfg.acc_full;
class sexo_r clasificacion_r repeticion_r;
var tipo_r a_o diasemana_r accidentado_r riesgo edad_r antiguedad_r puntuacion circunstancia_r sexo_r clasificacion_r repeticion_r ;
 monotone logistic (sexo_r = tipo_r a_o diasemana_r accidentado_r riesgo edad_r antiguedad_r puntuacion circunstancia_r );
 monotone logistic (clasificacion_r = tipo_r a_o diasemana_r accidentado_r riesgo edad_r antiguedad_r puntuacion circunstancia_r sexo_r);
 monotone logistic (repeticion_r =tipo_r a_o diasemana_r accidentado_r riesgo edad_r antiguedad_r puntuacion circunstancia_r sexo_r clasificacion_r );
 run;
```

Después de la ejecución de la sintaxis anterior, ya se tendría el conjunto de datos completo. La siguiente salida de SAS (tabla 18) indica el método que se ha usado para cada variable.

Monotone Model Specification	
Method	**Imputed Variables**
Regression	a_o diasemana_r accidentado_r riesgo edad_r antiguedad_r puntuacion circunstancia_r
Logistic Regression	sexo_r clasificacion_r repeticion_r

Tabla 18: *Métodos de imputación en cada variable.*

Para el método de regresión, las variables incluidas en él no han sido imputadas ya que se han imputado en pasos anteriores, y mediante el método de regresión logística se han imputado las variables tipo CLASS que se indicaban.

Paso 2: Análisis.

Aquí es importante que todos los pasos anteriores se hayan hecho correctamente, sobre todo el paso preliminar del tratamiento de las variables. Como lo que se explica en pasos anteriores de una correcta recategorización de las variables, para que todas las categorías se encuentren bien explicadas para cada categoría de la variable respuesta.

Con el procedimiento PROC LOGISTIC, si eso no se ha hecho, se encargará de producir una advertencia de una "separación casi completa". Si esto ocurre, el examinador tendría que modificar el modelo excluyendo o recategorizando ciertas covariables para evitar este problema. [17]

En la siguiente sintaxis se utiliza un modelo de regresión logística para estimar si el evento será un accidente o un incidente usando las covariables continuas y categóricas para predecir la variable tipo_r.

El conjunto de datos de salida de la última imputación es la que se usa para el procedimiento PROC LOGISTIC. Como esta salida contiene 5 copias imputadas del original, el procedimiento de análisis se ejecuta con la sentencia BY _imputation_, de manera que se realizará el análisis para cada conjunto de datos imputados.

El fichero de datos de salida de ODS OUTPUT PARAMETERESTIMATES es para capturar las estimaciones de los coeficientes de regresión, estimados por el modelo de análisis de cada uno de conjuntos de datos imputados. Estos conjuntos de datos también estarán identificados por la variable _imputation_.

```
ods output parameterestimates=tfg.parametros (where=(_Imputation_ ne .)) ;
 proc logistic data=tfg.acc_full ;
 class diasemana_r accidentado_r sexo_r clasificacion_r repeticion_r  ;
 model tipo_r (event='1') =  a_o diasemana_r accidentado_r riesgo edad_r
antiguedad_r puntuacion circunstancia_r sexo_r clasificacion_r repeticion_r
/ lackfit expb;
 by _imputation_;
run;
```

Una vez ejecutada la sintaxis, se obtiene el análisis de regresión logística para cada una de las imputaciones.

Cabe destacar, que no se ha seleccionado ningún método de selección de variables para el modelo, ya que como luego se van a combinar todos los análisis no sería posible porque en cada conjunto se seleccionarían unas variables determinadas.

Paso 3: Agrupación.

En este paso, se invoca un PROC MINALYZE para combinar los resultados del PROC LOGISTIC. El conjunto de datos creado en la salida de PARAMETERESTIMATES anterior guardado bajo el nombre tfg.parametros es el que pasa a ser la entrada en el PROC ANALYZE utilizando la opción PARMS. En esta sentencia se puede incluir cualquier archivo que contenga estimaciones de parámetros como los errores estándar asociados calculados a partir de conjuntos de datos imputados.

Este procedimiento lee nombres de efectos a partir de observaciones con la variable parameter, effect, variable o parm.

Cuando los efectos contienen variables categóricas, se utiliza la opción CLASSVAR=ctype para identificar las variables de clasificación asociadas al leer los niveles de clasificación a partir de observaciones. Los tipos disponibles son FULL, LEVEL, CLASSVAL. El valor por defecto es CLASSVAL=FULL. [18]

En este caso, se a puesto la opción CLASSVAL=CLASSVAR, donde se lee los niveles de clasificación para el efecto a partir de observaciones con las variables classval0, classval1, … donde la variable classval0 contiene el nivel de clasificación para la primera

variable de clasificación del efecto. Para cada efecto, estas variables categóricas se tienen que poner antes que el resto. El orden de las variables en la sentencia CLASS se usa para las variables dentro de cada lista. [18]

```
proc mianalyze parms(classvar=classval)=tfg.parametros ;
  class diasemana_r accidentado_r sexo_r clasificacion_r repeticion_r ;
 modeleffects intercept diasemana_r accidentado_r sexo_r clasificacion_r repeticion_r
a_o riesgo edad_r antiguedad_r puntuacion circunstancia_r    ;
ods output parameterestimates=tfg.main_parametros;
run ;
```

En la sentencia MODELEFFECTS se incluyen todas las variables de las que se han calculado sus parámetros estimados en el procedimiento PROC LOGISTIC. Con el ODS OUTPUT PARAMETERESTIMATES= se obtiene un fichero que se ha llamado tfg.main_parametros donde aparece la combinación de los resultados obtenidos para cada imputación, el cual se muestra en la tabla 19.

Parameter Estimates (5 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	967.34	51.089	867.2056	1067.479	29899	962.288	973.782971	0	18.93	<.0001
diasemana_r (1)	-0.0912	0.0453	-0.1802	-0.002	60448	-0.09735	-0.087799	0	-2.01	0.0442
diasemana_r (2)	-0.0540	0.0432	-0.1389	0.031	94898	-0.05822	-0.051005	0	-1.25	0.2117
diasemana_r (3)	-0.1603	0.0433	-0.2454	-0.075	182748	-0.16391	-0.157363	0	-3.70	0.0002
diasemana_r (4)	-0.2493	0.0432	-0.3341	-0.164	19183	-0.25550	-0.244511	0	-5.76	<.0001
diasemana_r (5)	-0.1636	0.0469	-0.2557	-0.072	52886	-0.16974	-0.159053	0	-3.48	0.0005
diasemana_r (6)	0.4473	0.0709	0.3084	0.586	10856	0.43225	0.454436	0	6.31	<.0001
accidentado_r (1)	0.7423	0.0345	0.6747	0.810	60985	0.73865	0.745288	0	21.51	<.0001
accidentado_r (2)	0.4327	0.0372	0.3597	0.506	85743	0.42827	0.435193	0	11.62	<.0001
sexo_r (0)	0.0258	0.0296	-0.0331	0.085	75.034	0.01151	0.038313	0	0.87	0.3848
clasificacion_r (1)	-0.417	0.1003	-0.6144	-0.221	12036	-0.43757	-0.404817	0	-4.16	<.0001
clasificacion_r (2)	-0.400	0.0710	-0.5394	-0.261	5085.1	-0.41767	-0.389209	0	-5.63	<.0001
clasificacion_r (3)	-0.097	0.0748	-0.2445	0.049	1185.7	-0.12020	-0.077500	0	-1.31	0.1917
repeticion_r (1)	-0.257	0.0501	-0.3570	-0.159	158.48	-0.28411	-0.232737	0	-5.14	<.0001
repeticion_r (2)	0.0927	0.0428	0.0086	0.177	633.19	0.07708	0.107053	0	2.16	0.0308
repeticion_r (3)	0.3788	0.0506	0.2790	0.479	194.31	0.35833	0.406070	0	7.48	<.0001
repeticion_r (4)	0.3831	0.0676	0.2502	0.516	518.25	0.36158	0.403905	0	5.66	<.0001
a_o	-0.4792	0.0253	-0.5290	-0.430	29597	-0.48247	-0.476800	0	-18.92	<.0001
riesgo	-0.0474	0.0062	-0.0598	-0.035	48643	-0.04797	-0.046717	0	-7.53	<.0001
edad_r	-0.0036	0.0029	-0.0098	0.003	16.457	-0.00631	-0.001327	0	-1.25	0.2287
antiguedad_r	-0.0350	0.0036	-0.0425	-0.028	27.506	-0.03725	-0.032908	0	-9.69	<.0001
puntuacion	0.0142	0.0061	0.0022	0.026	52393	0.01360	0.014822	0	2.33	0.0201
circunstancia_r	-0.0246	0.0061	-0.0367	-0.013	58600	-0.02535	-0.023974	0	-4.00	<.0001

Tabla 19: *Parámetros estimados del modelo de regresión.*

La tabla ha sido modificada de forma que se pueda interpretar lo mejor posible en el documento. La columna referida a las variables se ha insertado también el valor de la dummy (en la tabla real existen tantas columnas más como variables categóricas para indicar el valor de la dummy de cada variable). Las siguientes columnas son el parámetro estimado de cada una junto con su error estándar, intervalo de confianza, mínimo,

máximo y las tres últimas son para realizar el contraste de hipótesis de significatividad del parámetro. El contraste de hipótesis sería el siguiente:

$H_o: \beta_i = 0$ frente a $H_1: \beta_i \neq 0$

Donde i=1,…,23 que corresponde a cada uno de los parámetros de la tabla.

En cuanto al p-valor del estadístico se observa que, con un nivel de significación fijado de 0.05, no existe evidencia en contra de la hipótesis nula en las variables edad_r y sexo_r. Por lo que estas no aportarían información significativa al modelo. Por otro lado, algunas variables dummy tampoco son significativas pero sus otras variables dummy si lo son por ello no se eliminarían del modelo y si aportarían información, ya que si no habría que eliminar sus otras variables dummy también. El resto de variables si aportarían información significativa al modelo.

De todas formas, no se va a eliminar ninguna variable del modelo. La información anterior se tendrá en cuenta para no interpretarlas, ya que no aportan información.

4.4. Análisis del fichero.

Una vez obtenidos la combinación de todas las regresiones logísticas, se puede calcular los OR creando un nuevo fichero de datos a partir de este.

Para el cálculo de los OR, se hará la exponencial de los parámetros estimados, ya que esta es su fórmula con la regresión logística.

$$OR_i = \exp(\beta_i)$$

Para ello, en la siguiente sintaxis, se crea un fichero data en el cual se va a obtener todas las variables del fichero tfg.main_parametros añadiéndole el valor de los OR y su intervalo de confianza.

```
data tfg.regresion;
set tfg.main_parametros;
OR= exp(estimate);
lcl_or=OR*EXP(-1.96*STDERR);
ucl_or=OR*EXP(1.96*STDERR);
run;
```

Del fichero que se crea, tfg.regresion, solo se representará en la tabla 20, las nuevas variables calculadas junto con los estimadores, los nombres de la variable y los valores de su variable dummy.

Obs	Parm	diasemana_r	accidentado_r	sexo_r	clasificacion_r	repeticion_r	Estimate	StdErr	OR	lcl_or	ucl_or
1	Intercept						967.342549	51.089135	.	.	.
2	diasemana_r	1					-0.091267	0.045354	0.91277	0.83514	0.99763
3	diasemana_r	2					-0.054067	0.043294	0.94737	0.87029	1.03127
4	diasemana_r	3					-0.160345	0.043371	0.85185	0.78243	0.92743
5	diasemana_r	4					-0.249312	0.043270	0.77934	0.71597	0.84832
6	diasemana_r	5					-0.163646	0.046973	0.84904	0.77436	0.93092
7	diasemana_r	6					0.447398	0.070934	1.56424	1.36120	1.79756
8	accidentado_r		1				0.742365	0.034508	2.10090	1.96350	2.24791
9	accidentado_r		2				0.432707	0.037229	1.54143	1.43295	1.65811
10	sexo_r			0			0.025891	0.029618	1.02623	0.96835	1.08757
11	clasificacion_r				1		-0.417605	0.100398	0.65862	0.54097	0.80186
12	clasificacion_r				2		-0.400021	0.071083	0.67031	0.58313	0.77051
13	clasificacion_r				3		-0.097744	0.074823	0.90688	0.78318	1.05013
14	repeticion_r					1	-0.257932	0.050170	0.77265	0.70029	0.85248
15	repeticion_r					2	0.092776	0.042855	1.09722	1.00882	1.19336
16	repeticion_r					3	0.378885	0.050636	1.46066	1.32265	1.61306
17	repeticion_r					4	0.383157	0.067676	1.46691	1.28468	1.67498
18	a_o						-0.479293	0.025338	0.61922	0.58922	0.65075
19	Riesgo						-0.047431	0.006299	0.95368	0.94197	0.96552
20	edad_r						-0.003634	0.002907	0.99637	0.99071	1.00207
21	antiguedad_r						-0.035090	0.003620	0.96552	0.95869	0.97239
22	puntuacion						0.014236	0.006123	1.01434	1.00224	1.02658
23	circunstancia_r						-0.024637	0.006157	0.97566	0.96396	0.98751

Tabla 20: *Tabla de los parámetros junto con los OR.*

Las variables subrayadas en amarillo son aquellas en las que el intervalo de confianza contiene al uno y, por lo tanto, su OR no es significativo a la hora de interpretarlo, ya que no habría asociación entre ellas como bien se ha mencionado antes.

Se tienen variables que no son significativos sus odds ratio, ya que no aporta información suficiente. Las que se subrayaron en amarillo en la tabla de los parámetros XXX, que son las variables sexo_r y edad_r, sus coeficientes no eran significativos, por lo tanto sus odds ratio tampoco.

Ahora vamos a interpretar cada uno de los OR significativos del modelo.

OR(diasemana 1) = 0.91277 → $\frac{1}{0.91277} = 1.09$

Los domingos tienen un 1.09 probabilidades más de sufrir un accidente que los lunes.

OR(diasemana 3) = 0.85185 → $\frac{1}{0.85185} = 1.174$

Los domingos tienen 1.174 probabilidades más de sufrir un accidente que los miércoles.

OR(diasemana 4) = 0.77934 → $\frac{1}{0.77934} = 1.28$

Los domingos tienen un 1.28 probabilidades más de sufrir un accidente que los jueves.

OR(diasemana 5) = 0.84904 → $\frac{1}{0.84904} = 1.18$

Los domingos tienen un 1.18 probabilidades más de sufrir un accidente que los viernes.

OR(diasemana 6) = 1.56424

Los sábados tienen un 1.56 probabilidades más de sufrir un accidentes que los domingos.

OR(accidentado 1) = 2.1

Los empleados de tipo Propio tienen un 2.1 de probabilidades más de sufrir un accidente que los empleados de tipo Otro o mixto.

OR (accidentado 2) = 1.54

Los empleados de tipo Contratista tienen un 1.54 más de probabilidades de sufrir un accidente que los que son de tipo Otro o mixto.

OR (clasificacion 1) = 0.65862 → $\frac{1}{0.65862} = 1.52$

Los eventos con una clasificación muy grave tienen 1.52 probabilidades más de sufrir un accidente que los eventos de tipo leve.

OR (clasificación 2) = 0.67031 → $\frac{1}{0.67031} = 1.49$

Los eventos con una clasificación muy grave tienen 1.49 probabilidades más de sufrir un accidente que los eventos de tipo menos grave.

OR (clasificación 3) = 0.90688 → $\frac{1}{0.90688} = 1.1$

Los eventos con una clasificación muy grave tienen 1.1 probabilidades más de sufrir un accidente que los eventos de tipo grave.

OR (repetición_r 1) = 0.77265 → $\frac{1}{0.77265} = 1.29$

Los eventos con una repetición muy alta tienen un 1.29 probabilidades más de que sea un accidente que si el evento tiene una probabilidad de repetición muy baja.

OR (repetición_r 2) = 1.09

Los eventos con una repetición baja tienen 1.09 probabilidades más de que sea un accidente que los eventos de repetición muy grave.

OR (repetición_r 3) = 1.46

Los eventos con repetición media tienen 1.46 probabilidades más de que sea un accidente que los eventos con repetición tipo muy grave.

OR (repetición_r 4) = 1.47

Los eventos con repetición grave tienen 1.47 probabilidades más de que sea un accidente que los eventos con repetición muy grave

OR (año) = 0.62

Cuando el evento surgido es más antiguo más probabilidades tiene de ser un accidente.

OR (riesgo) = 0.95368

Cuanto el riesgo es más pequeño más probabilidades hay de que sea un accidente.

OR (antigüedad) = 0.96552

Cuanto menos tiempo lleva un empleado trabajando en la empresa más probabilidades hay de que tenga un accidente.

OR (puntuación) = 1.01

Cuanta más puntuación tenga el evento tiene un 1.01 de probabilidades más de que sea un accidente que si la puntuación es más baja.

OR (circunstancia) = 0.97566

Cuanto más baja es la circunstancia asociada más probabilidades hay de que sea un accidente que si es más alta.

5. CONCLUSIONES.

En este trabajo, se ha explicado cómo proceder a la depuración de una base de datos. Se ha visto que el fichero que se obtuvo tenía muchos errores en su base de datos. La cuál se ha "limpiado" para poder analizarla.

A consecuencia de diversos estudios realizados de comparativas de análisis de datos completos contra análisis con conjuntos de datos imputados, donde se demostraba la superioridad del análisis con el conjunto de datos imputados y especialmente con la imputación múltiple, se decidió realizar el estudio con imputación múltiple.

Se pasó a la fase de imputación en varios conjuntos de datos, en la cual se realizaron los 3 pasos generales para ello.

En el primer pasó se encontró un grave problema con la estructura de los datos ausentes del fichero, ya que sin una estructura monótona no se podían imputar los datos faltantes.

Se obtuvieron distintos análisis de regresión logística para los 5 conjuntos de datos imputados. Con estas salidas, se combinaron los 5 análisis obteniendo así unos

parámetros estimados de regresión logística. No se demostró la significatividad del modelo, ya que lo que se buscaba con este análisis eran las relaciones entre cada una de las covariables con la variable objetivo de estudio. Para ello, se calcularon los odds-ratio, medidas de asociación. En las cuales se obtuvieron las siguientes conclusiones.

El objetivo de análisis fue la relación con la probabilidad de tener un accidente en la empresa que recogió los eventos ocurridos. Se obtuvo que los domingos tienen mayor probabilidad de que ocurra un accidente que de lunes a viernes y los sábados mayor probabilidad que los domingos. Luego, los fines de semana aumenta la probabilidad de accidente. Esto puede ser debido a que los trabajos que se realizan los fines de semana suelen ser de emergencia y el empleado tiene que dejar lo que este haciendo para ir a trabajar. Por lo tanto, suelen ser trabajos que se realizan con la mayor rapidez posible para poder volver a casa, de esta manera suelen haber más despistes y por lo tanto accidentes.

La relación más fuerte con la probabilidad de sufrir un accidente es que el empleado sea de tipo 'Propio', es decir, los empleados de la propia empresa. Y en menor medida los empleados de tipo 'Contratista', que son aquellos que se han contratado para un determinado trabajo.

También es lógico, que los eventos ocurridos clasificados como muy grave son aquellos que tienen más probabilidad de ser un accidente.

En cuanto a la probabilidad de repetición, un evento tiene más posibilidad de ser un accidente si su repetición es de tipo muy grave o media, ya que hay más accidentes con esta categoría de repetición.

Por otro lado, la base de datos tiene recogido el año en el que sucedió el evento, del que se ha obtenido que cuanto antes sucedió el evento, más probabilidades hay de que sea un accidente.

La antigüedad del empleado también tiene una relación significativa con la probabilidad de que sea un accidente, en especial, aquellos trabajadores que llevan menos tiempo trabajando en la empresa. Por último, destacar las circunstancias asociadas al accidente, cuantas menos circunstancias halla más probable es que sea un accidente.

6. BIBLIOGRAFÍA.

[1] F. Sánchez, «El país,» 24 Marzo 2016. [En línea]. Available: https://elpais.com/internacional/2016/03/23/actualidad/1458753834_097198.html.

[2] «quironpreventon,» 17 Marzo 2015. [En línea]. Available: https://www.quironprevencion.com/blogs/es/prevenidos/prevencion-riesgos-laborales-prl. [Último acceso: 4 Junio 2018].

[3] «Safety+Health,» 21 Febrero 2012. [En línea]. Available: http://www.safetyandhealthmagazine.com/articles/white-paper-predictive-analytics-in-workplace-safety-four-safety-truths-that-reduce-workplace-injuries-2. [Último acceso: 4 Junio 2018].

[4] «Prediction of occupational accidents using decision tree approach,» 2 Febrero 2017. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/7838969/. [Último acceso: 4 Junio 2018].

[5] «difiere,» [En línea]. Available: https://difiere.com/diferencia-accidente-e-incidente/. [Último acceso: 4 Junio 2018].

[6] «PowerData,» 19 Octubre 2016. [En línea]. Available: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/algunas-ventajas-que-proporciona-la-depuracion-de-datos.

[7] v. d. Heijden, «NCBI,» 11 Julio 2006. [En línea]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16980151.

[8] I. L. y. M. O. Bohdana Ratitch, «PHARMA Sug,» 2013. [En línea]. Available: https://www.pharmasug.org/proceedings/2013/SP/PharmaSUG-2013-SP03.pdf. [Último acceso: 4 Junio 2018].

[9] «PowerData,» [En línea]. Available: https://www.tecnologias-informacion.com/transformacion.html. [Último acceso: 4 Junio 2018].

[10] v. d. H. GJ, «NCBI,» 11 Julio 2006. [En línea]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16980151.

[11] J. M. A. Revenga, «Datos Missing: Detección y tratamiento».

[12] Á. Planchuelo, «Trabajo Fin de Máster,» Julio, Madrid, 2017.

[13] I. L. y. M. O. Bohdana Ratitch, «Pharma SUG,» 2013. [En línea]. Available: https://www.pharmasug.org/proceedings/2013/SP/PharmaSUG-2013-SP03.pdf.

[14] J. Cardenas, «Networkianos,» 1 Diciembre 2015. [En línea]. Available: http://networkianos.com/odd-ratio-que-es-como-se-interpreta/. [Último acceso: 4 Junio 2018].

[15] «SAS Suport,» [En línea]. Available: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mi_sect004.htm. [Último acceso: 4 Junio 2018].

[16] Patricia A. Berglund, «SAS suport,» 2010. [En línea]. Available: http://support.sas.com/resources/papers/proceedings10/265-2010.pdf. [Último acceso: 4 Junio 2018].

[17] «idre,» [En línea]. Available: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/. [Último acceso: 4 Junio 2018].

[18] «SAS Suport,» [En línea]. Available: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mianalyze_sect011.htm. [Último acceso: 4 Junio 2018].