



UNIVERSIDAD AUTÓNOMA METROPOLITANA

DIVISIÓN DE CIENCIAS NATURALES E INGENIERÍA

EL DIAGNÓSTICO MÉDICO DESDE UNA
PERSPECTIVA DE MINERÍA DE DATOS

T E S I N A

TRABAJO ESCRITO DE INVESTIGACIÓN
CORRESPONDIENTE A LOS PROYECTOS
TERMINALES I, II Y III

PRESENTA:

DIEGO JOVAN CANSINO MALPICA

ASESOR:

DR. PEDRO PABLO GONZÁLEZ PÉREZ

Mayo, 2022



A mi madre y mi padre, quienes han dado todo para mi crecimiento e hicieron posible que finalizara este trabajo, ustedes siempre desearon lo mejor para mí y han hecho lo imposible para dármelo. Gracias por su amor y apoyo incondicional a lo largo de mi vida.

A mis amigos y profesores, quienes me ayudaron a transcender a lo largo de este camino, brindándome tiempo y conocimiento.

Índice general

Índice de figuras	v
1 Big Data	1
1.1 Definición	1
1.2 Características	1
1.2.1 Volumen	2
1.2.2 Velocidad	2
1.2.3 Variedad	2
1.2.4 Veracidad	3
1.2.5 Valor	3
1.3 Tipos de datos	4
1.3.1 Datos de grandes transacciones (Big Transaction Data)	4
1.3.2 Redes sociales y páginas web	5
1.3.3 Biométricas	5
1.3.4 Generados por los seres humanos	5
1.3.5 Máquinas (Machine to Machine M2M)	5
1.4 Tipos de datos según su formato	5
1.4.1 Datos estructurados	5
1.4.2 Datos semi estructurados	5
1.4.3 Datos no estructurados	6
1.5 Fuentes de datos	6
1.6 Desafíos	7
1.6.1 Crecimiento de los datos	7
1.6.2 Calidad de los datos	7
1.6.3 Infraestructura	8
1.6.4 Experiencia	8
1.6.5 Seguridad	8
1.7 Casos de uso	8
1.7.1 Consumo	8
1.7.2 Financiera	8
1.7.3 Sector público	8
1.7.4 Telecomunicaciones	8
1.7.5 Administración	9
1.7.6 Comercio electrónico	9
1.7.7 Salud	9

1.8 Funcionamiento	10
1.8.1 Integrar	10
1.8.2 Administrar	10
1.8.3 Analizar	10
2 Ciencia de datos	11
2.1 Introducción	11
2.2 Metodología CRISP-DM	11
2.2.1 Comprensión del problema o negocio	12
2.2.2 Comprensión de datos	12
2.2.3 Preparación de datos	12
2.2.4 Modelado	12
2.2.5 Evaluación del modelo	13
2.2.6 Implementación del modelo	13
2.3 Modelos dentro de la fase del modelado	13
2.3.1 Red Neuronal	13
2.3.2 Árbol de decisión	16
2.3.3 Algoritmo de regresión	18
3 Ciencias de la vida	20
3.1 Introducción	20
3.2 El área de salud	21
3.3 Diagnóstico médico	22
3.4 Características del diagnóstico médico	22
3.4.1 Validez	22
3.4.2 Reproductividad	22
3.4.3 Seguridad	23
3.5 Tipos de diagnóstico médico	23
3.5.1 Diagnóstico etiológico	23
3.5.2 Diagnóstico sintomatológico	23
3.5.3 Diagnóstico diferencial	23
3.5.4 Diagnóstico genérico	23
3.5.5 Diagnóstico nosológico	23
3.5.6 Diagnóstico patogénico	23
3.6 Predicción de enfermedades	24
3.7 Casos de estudio de los big data aplicados al área de la salud	26
3.8 Especialidades en el área de salud	27
3.9 Oncología	29
3.10 Cardiología	30
4 Aplicación de la metodología CRISP-DM	32
4.1 Comprensión del dominio del problema	32
4.1.1 Determinación de los objetivos del proyecto	32
4.1.2 Valoración de la situación actual del objetivo del proyecto	32
4.1.3 Determinación de los objetivos de minería de datos	33

4.1.4	Propuesta del enfoque metodológico (plan de proyecto de minería de datos)	34
4.2	Comprensión de los datos	35
4.2.1	Recopilación de los datos iniciales	35
4.2.2	Descripción de los datos	35
4.2.3	Exploración de los datos	37
4.2.4	Verificación de la calidad de los datos	43
4.3	Preparación de los datos	44
4.3.1	Selección de datos	44
4.3.2	Limpieza de datos	44
4.3.3	Construcción de nuevos datos	46
4.3.4	Integración de datos	48
4.3.5	Formato de datos	48
4.3.6	Nueva exploración de los datos	50
4.4	Modelado	64
4.4.1	Selección de técnicas de modelado	64
4.4.2	Métodos de comprobación	65
4.4.3	Generación de los modelos	65
4.4.4	Modelo Árbol de decisión C&R	66
4.4.5	Modelo Red Neuronal Perceptron Backpropagation	70
4.4.6	Modelo Regresión logística	77
4.4.7	Modelo Árbol aleatorio	84
4.4.8	Modelo Árbol de decisión en Analytic Server	89
4.4.9	Evaluación	93
4.5	Caso de estudio - "Breast Cancer"	103
5	Herramienta Intelligent Data Analysis Tool	109
5.1	Descripción de la herramienta	109
5.2	Comprobación de su funcionamiento	119
5.3	Futuros trabajos	124
Conclusiones		126
Bibliografía		128

Índice de figuras

1.1	Características del big data - 3 V's.	2
1.2	Características del big data - 5 V's.	3
1.3	Características del big data - 10 V's.	4
1.4	Tipos de datos en big data.	6
1.5	Fuentes de datos en big data.	7
1.6	Casos de uso de big data.	9
1.7	Acciones clave para trabajar con big data.	10
2.1	Diagrama de secuencia sobre la metodología CRISP-DM.	13
2.2	Estructura de una red neuronal.	14
2.3	Estructura de una árbol de decisión.	17
2.4	Principales algoritmos de regresión.	18
3.1	Principales ciencias que componen las ciencias de la vida.	21
3.2	Patrón de diseño Bridge enfocado a los diagnósticos.	24
3.3	Proceso de recolección de datos en el área de la salud.	25
3.4	Especialidades en el área de la Salud.	29
3.5	Enfermedades más comunes dentro de Oncología.	30
3.6	Enfermedades más comunes dentro de Cardiología.	31
4.1	Visualización del archivo "heart_failure_clinical_records_dataset.csv" . .	35
4.2	Histograma - Edades de los pacientes.	37
4.3	Histograma - serum_creatinine de los pacientes.	38
4.4	Histograma - serum_sodium de los pacientes.	39
4.5	Histograma - creatinine_phosphokinase de los pacientes.	40
4.6	Histograma - Platelets de los pacientes.	41
4.7	Histograma - Ejection_fraction de los pacientes.	42
4.8	Histograma - Time de los pacientes.	43
4.9	Filtro de los datos en la herramienta IBM SPSS Modeler.	44
4.10	Resultado de la limpieza de los datos con la herramienta IBM SPSS Modeler	45
4.11	Fórmula para derivar el campo DiferenciaDeMediaSC.	46
4.12	Fórmula para derivar el campo DiferenciaDeMediaCP.	47
4.13	Fórmula para derivar el campo DiferenciaDeMediaPlatelets.	47
4.14	Fórmula para derivar el campo DiferenciaDeMediaSS.	48
4.15	Resultado de la limpieza de los datos con la herramienta IBM SPSS Modeler	50
4.16	Histograma - Age	51

4.17 Histograma - serum_creatinine	52
4.18 Histograma - ejection_fraction	53
4.19 Histograma - creatinine_phosphokinase	54
4.20 Histograma - platelets_transformed	55
4.21 Recuento de la edad de los pacientes.	56
4.22 Diagrama de caja - creatinine_phosphokinase.	57
4.23 Diagrama de caja - platelets_transformed	58
4.24 Diagrama de caja - serum_sodium	59
4.25 Colección de anemia respecto a la edad del paciente.	60
4.26 Colección de diabetes respecto a la edad del paciente.	61
4.27 Colección de presión arterial alta respecto a la edad del paciente.	62
4.28 Colección de sexo respecto a la edad del paciente.	63
4.29 Colección de fumadores respecto a la edad del paciente.	64
4.30 Ruta para generar modelos dentro de la herramienta IBM SPSS Modeler.	65
4.31 Configuración del modelo Árbol de decisión C&R - 1.	66
4.32 Configuración del modelo Árbol de decisión C&R - 2.	67
4.33 Configuración del modelo Árbol de decisión C&R - 3.	68
4.34 Ejecución del modelo Árbol de decisión C&R - 1.	69
4.35 Ejecución del modelo Árbol de decisión C&R - 2.	69
4.36 Configuración del modelo Red Neuronal Perceptron Backpropagation - 1.	70
4.37 Configuración del modelo Red Neuronal Perceptron Backpropagation - 2.	71
4.38 Configuración del modelo Red Neuronal Perceptron Backpropagation - 3.	72
4.39 Ejecución del modelo Red Neuronal Perceptron Backpropagation - 1.	73
4.40 Ejecución del modelo Red Neuronal Perceptron Backpropagation - 2.	74
4.41 Ejecución del modelo Red Neuronal Perceptron Backpropagation - 3.	75
4.42 Ejecución del modelo Red Neuronal Perceptron Backpropagation - 4.	76
4.43 Configuración del modelo Regresión logística - 1.	77
4.44 Configuración del modelo Regresión logística - 2.	78
4.45 Ejecución del modelo Regresión logística - 1.	79
4.46 Ejecución del modelo Regresión logística - 2.	80
4.47 Ejecución del modelo Regresión logística - 3.	81
4.48 Ejecución del modelo Regresión logística - 4.	82
4.49 Ejecución del modelo Regresión logística - 5.	83
4.50 Configuración del modelo Árbol aleatorio - 1.	84
4.51 Configuración del modelo Árbol aleatorio - 2.	85
4.52 Ejecución del modelo Árbol aleatorio - 1.	86
4.53 Ejecución del modelo Árbol aleatorio - 2.	87
4.54 Ejecución del modelo Árbol aleatorio - 3.	88
4.55 Ejecución del modelo Árbol aleatorio - 4.	88
4.56 Configuración del modelo Árbol de decisión en Analytic Server - 1.	89
4.57 Configuración del modelo Árbol de decisión en Analytic Server - 2.	90
4.58 Ejecución del modelo Árbol de decisión en Analytic Server - 1.	91
4.59 Ejecución del modelo Árbol de decisión en Analytic Server - 2.	92
4.60 Ejecución del modelo Árbol de decisión en Analytic Server - 3.	93
4.61 Evaluación del modelo Árbol de decisión C&R.	94

4.62 Evaluación del modelo Red Neuronal Perceptron Backpropagation.	95
4.63 Evaluación del modelo Regresión logística.	96
4.64 Evaluación del modelo Árbol aleatorio.	97
4.65 Evaluación del modelo Árbol de decisión en Analytic Server.	98
4.66 Ruta para evaluar los modelos en la herramienta IBM SPSS Modeler.	99
4.67 Registros seleccionados para comprobar el funcionamiento de los modelos. .	99
4.68 Cambio de la medida de los campos.	100
4.69 Valores destino de los registros seleccionados.	100
4.70 Predicción del modelo Red Neuronal Perceptron Backpropagation.	101
4.71 Predicción del modelo Árbol de decisión C&R.	101
4.72 Configuración ingresada en la funcionalidad 'Datos Usuario'.	102
4.73 Predicción del modelo Árbol aleatorio.	102
4.74 Predicción del modelo Regresión logística.	103
4.75 Predicción del modelo Árbol de decisión en Analytic Server.	103
4.76 Ruta para evaluar los modelos con el data set "Breast Cancer".	104
4.77 Visualización del data set "Breast Cancer".	105
4.78 Medidas y valores de los campos del data set "Breast Cancer".	105
4.79 Evaluación del modelo Red Neuronal Perceptron Backpropagation.	106
4.80 Evaluación del modelo Árbol aleatorio.	107
4.81 Evaluación del modelo Regresión logística.	108
5.1 Interfaz introductoria de la aplicación.	110
5.2 Interfaz principal de la aplicación.	111
5.3 "Show file" dentro de la aplicación.	112
5.4 Interfaz para realizar gráficos.	112
5.5 Gráfico creado con la aplicación.	113
5.6 Pestaña "Preparation" de la aplicación.	114
5.7 Pestaña "Modeling" de la aplicación.	115
5.8 Interfaz para seleccionar los INPUT y TARGET.	116
5.9 Interfaz para cargar los datos de evaluación.	116
5.10 Datos de evaluación dentro de la aplicación.	117
5.11 Gráfica de la precisión del modelo.	118
5.12 Predicción del modelo.	119
5.13 "Show file" dentro de la aplicación - 2.	120
5.14 Interfaz para seleccionar los INPUT y TARGET - 2.	121
5.15 Datos de evaluación dentro de la aplicación - 3.	122
5.16 Gráfica de la precisión del modelo - 2.	123
5.17 Predicción del modelo - 2.	124

Capítulo 1

Big Data

1.1. Definición

Hoy en día, nos encontramos con múltiples definiciones de big data de importantes autores e instituciones alrededor del mundo, por ejemplo:

- IBM: "Es el uso de técnicas analíticas avanzadas contra conjuntos de datos muy grandes y diversos que incluyen datos estructurados, semiestructurados y no estructurados, de diferentes orígenes, y en tamaños diferentes de terabytes a zettabytes." [1]
- Oracle: "Los big data son conjuntos de datos de mayor tamaño y más complejos, procedentes particularmente de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede administrarlos." [2]

Finalmente, Pierson, L., & Porway, J., mencionan:

"Son datos que superan la capacidad de procesamiento de los sistemas de bases de datos convencionales porque son demasiado grandes, se mueven demasiado rápido o no se ajustan a los requisitos estructurales de las arquitecturas de bases de datos tradicionales. Tanto si los volúmenes de datos se clasifican en escalas de terabytes como petabytes, las soluciones de ingeniería de datos deben diseñarse para satisfacer los requisitos del destino y uso de los datos." [3]

La información antes presentada presenta ligeras diferencias, pero con ellas puedo determinar que big data es el conjunto de datos que son generados por humanos, redes sociales, sensores, transacciones, dispositivos inteligentes y otras diversas fuentes de datos, donde su fiabilidad, complejidad, variedad, velocidad de flujo de datos y volumen no permiten que sean almacenados, gestionados y/o procesados por medio de sistemas convencionales.

1.2. Características

En un principio, dadas las definiciones de big data, se plantearon tres características que la definen, estas son las tres V [4]:

1.2.1. Volumen

Es la característica que más asociamos a big data y se refiere a la cantidad de datos generados y almacenados, donde dicha cantidad de datos suele ser muy alta. El volumen se mide en gigabytes, terabytes, petabytes, entre otros.

1.2.2. Velocidad

En este caso es el volumen de datos por unidad de tiempo; es decir, nos referimos a la velocidad a la que fluyen los datos. Hoy en día, dentro de big data se manejan velocidades que oscilan entre 30 kilobytes por segundo hasta 30 gigabytes por segundo.

1.2.3. Variedad

La variedad se refiere a los diversos tipos de datos disponibles, entre estos están los datos estructurados, no estructurados y semiestructurados.

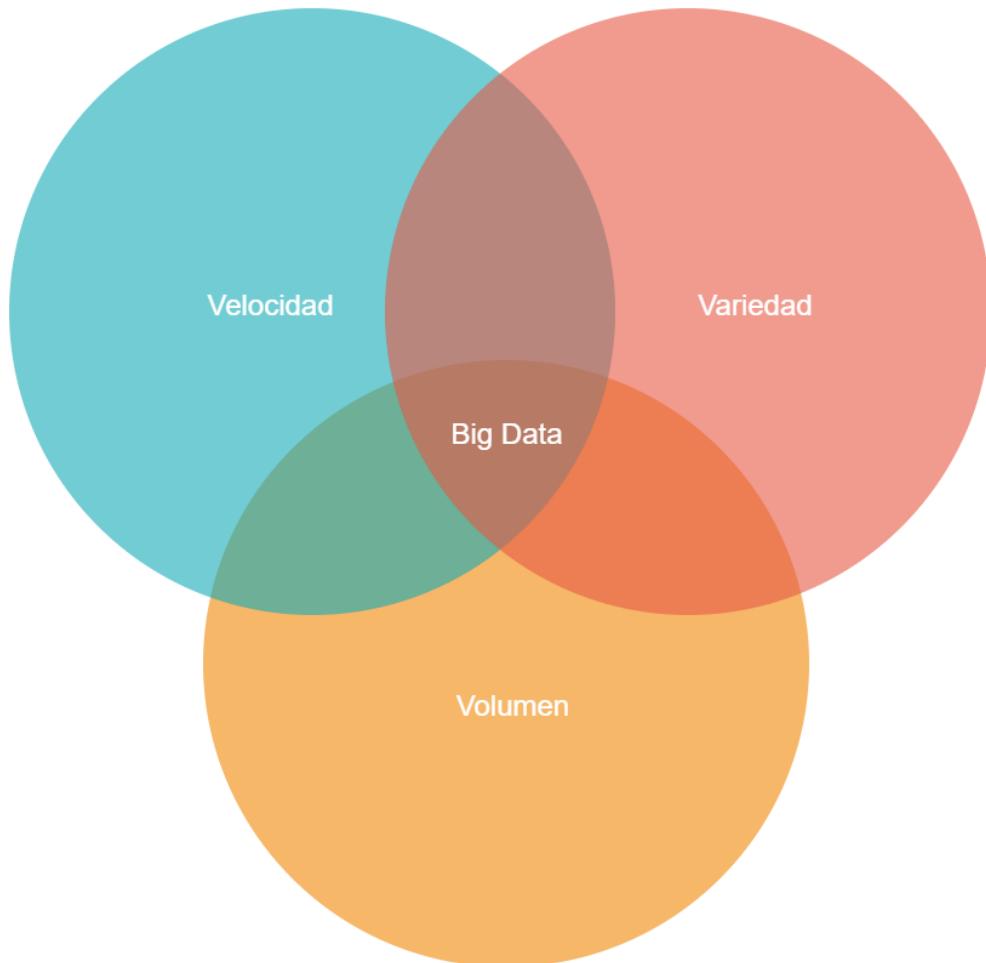


Figura 1.1: Características del big data - 3 V's.

Por otra parte, desde hace años se ha mencionado la integración de dos V's al modelo, los cuales se mencionarán a continuación.[4]

1.2.4. Veracidad

Se refiere al grado de fiabilidad de la información recolectada.

1.2.5. Valor

El valor hace referencia a la utilidad de los datos.

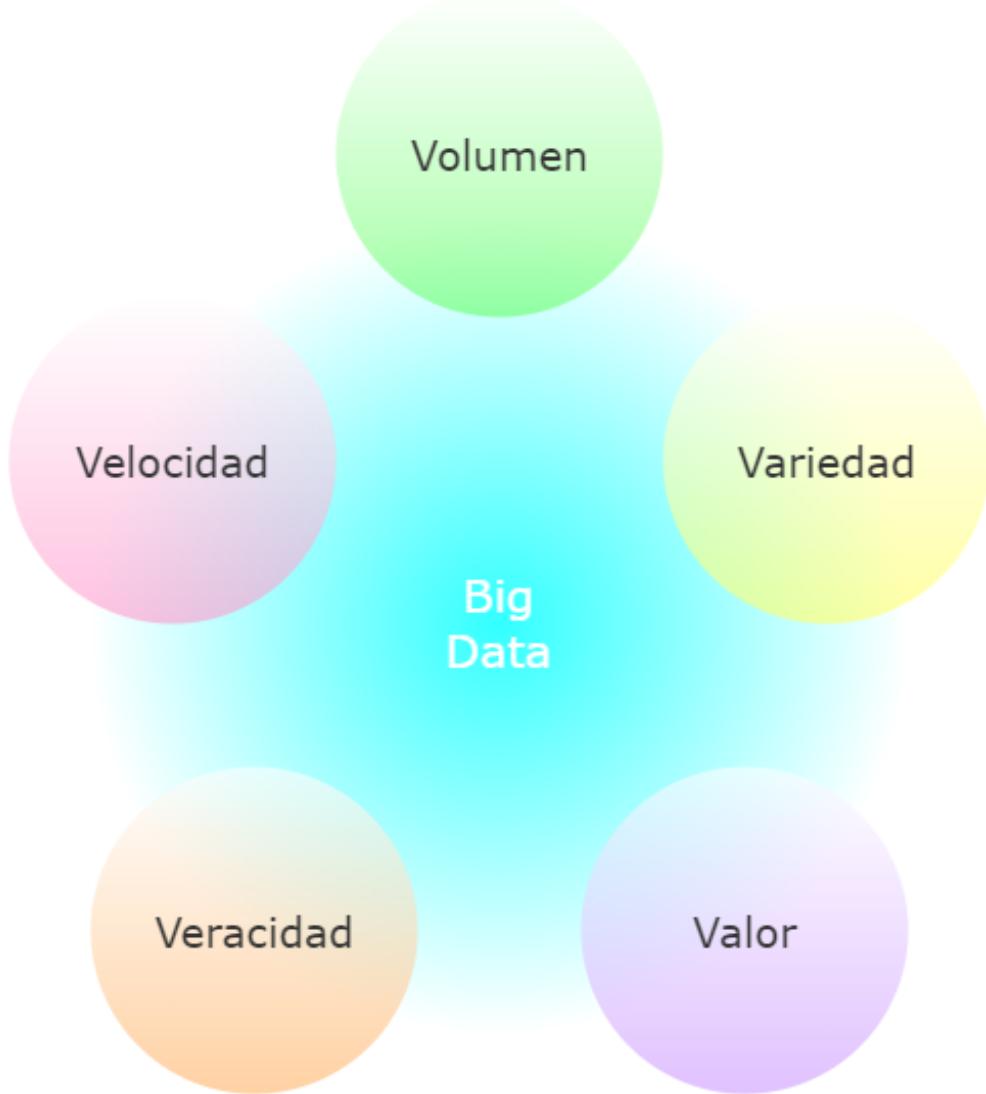


Figura 1.2: Características del big data - 5 V's.

Finalmente, este modelo sigue y seguirá evolucionando con el paso del tiempo, hoy en día se ha llegado a encontrar diferentes subconjuntos para que las organizaciones los tomen

en cuenta al momento de desarrollar estrategias. Uno de los subconjuntos del modelo más completos lo encontramos en la figura 1.3.



Figura 1.3: Características del big data - 10 V's.

1.3. Tipos de datos

Teniendo en cuenta la clasificación que nos presenta IBM[5], existen cinco grandes tipos.

1.3.1. Datos de grandes transacciones (Big Transaction Data)

Son datos que podemos encontrar en formatos semi estructurados o no estructurados. En este apartado se encuentran los registros de facturación, de las llamadas, telecomunicaciones, datos empresariales que se refieren a la información del cliente, la cual proviene de sistemas como el CRM; inventarios de ventas; datos transaccionales del ERP, etc.

1.3.2. Redes sociales y páginas web

Toda aquella información que se obtiene a través de las transacciones web, y el contenido que se adquiere de las redes sociales.

1.3.3. Biométricas

Se refiere a la información que incluye escaneo de la retina, huellas digitales, reconocimiento genético o facial, entre otros.

1.3.4. Generados por los seres humanos

Es aquella información generada por los humanos de manera indirecta; por ejemplo, cuando llamamos a un call center, escribimos correos electrónicos, documentos electrónicos, notas de voz, telecomunicaciones, uso de tarjetas de crédito o débito.

1.3.5. Máquinas (Machine to Machine M2M)

Datos generados por aquellas tecnologías que se conectan a otros dispositivos, y los utilizan como sensores o medidores; por ejemplo, cuando las compañías de servicios públicos miden el consumo de agua, gas o electricidad a través de medidores inteligentes.

1.4. Tipos de datos según su formato

Como se mencionó en la sección 1.3, existen diferentes tipos de datos donde cada uno de ellos tiene sus correspondientes propiedades, las cuales son:

1.4.1. Datos estructurados

Son aquellos datos con formatos fijos que poseen campos fijos; es decir, son los datos de las bases de datos relacionales, hojas de cálculo, entre otros. Estos datos se componen de información que se conoce de antemano, tiene un formato específico y se produce en un orden especificado.

1.4.2. Datos semi estructurados

Datos donde se tiene un flujo lógico y un formato que puede ser definido pero su comprensión es complicada para el usuario. Estos datos no tienen formato fijo, pero contienen etiquetas y otros marcadores que permiten separar los elementos; por ejemplo, texto de etiquetas de lenguajes XML y HTML.

1.4.3. Datos no estructurados

Tipo de datos sin tipos predefinidos, sin campos fijos, donde se tiene poco o ningún control sobre ellos; por ejemplo, audio, vídeo, fotografías, correos electrónicos, mensajes de texto, entre otros.

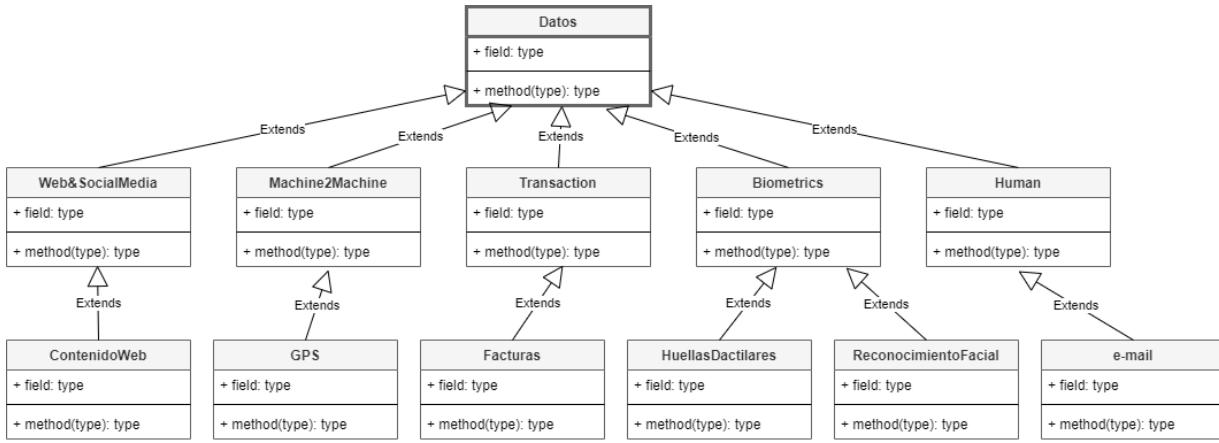


Figura 1.4: Tipos de datos en big data.

1.5. Fuentes de datos

En la actualidad, el volumen de datos son generados por humanos, máquinas y sensores que se encuentran alrededor del mundo. Normalmente, las fuentes de datos incluyen datos de redes sociales, transacciones bancarias, registros, secuencias de clic en páginas web, entre otros.

Las fuentes más populares en big data se muestran en la siguiente figura.

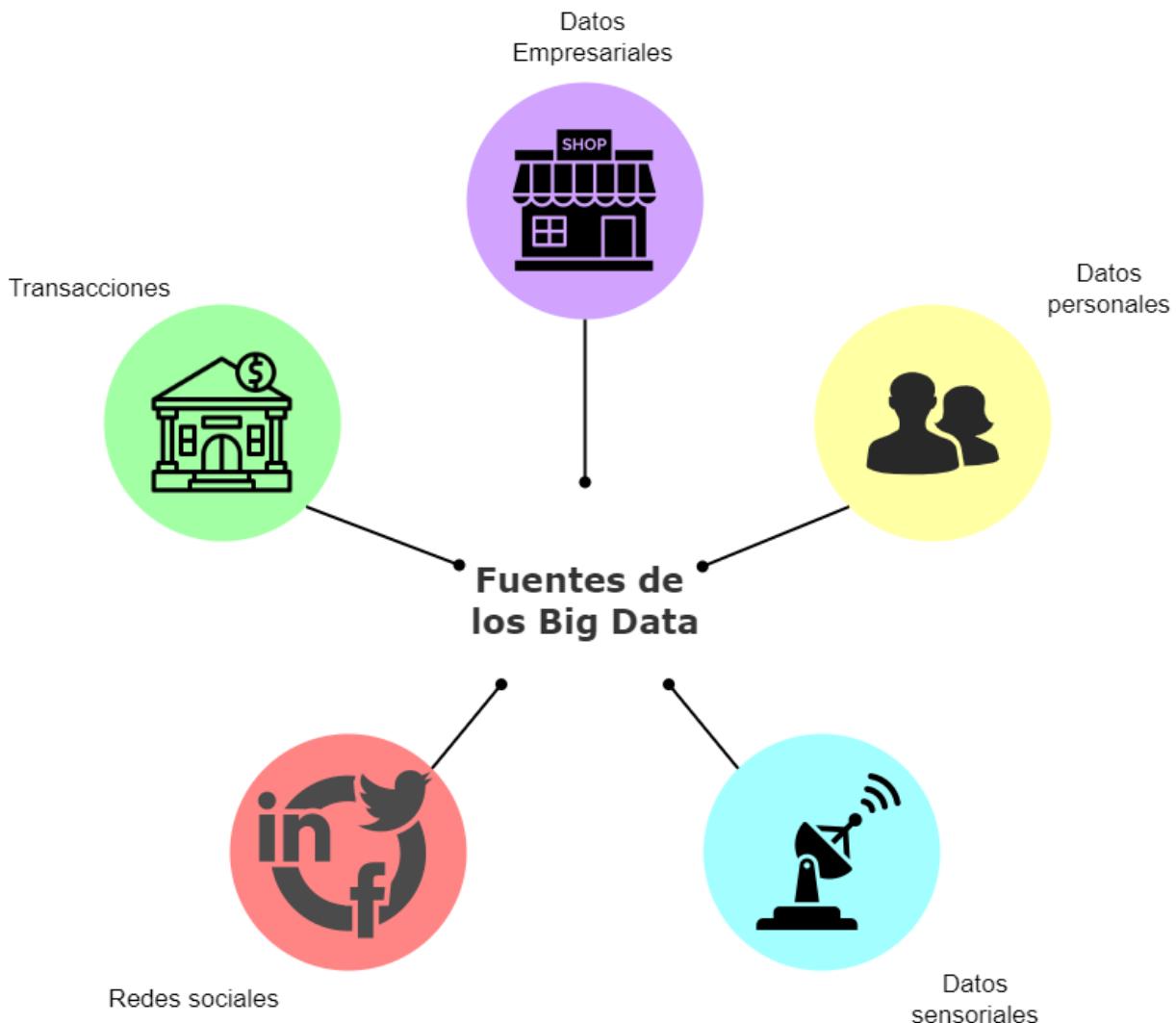


Figura 1.5: Fuentes de datos en big data.

1.6. Desafíos

Incluso con las grandes oportunidades que brindan los big data, hay muchos desafíos[6], entre ellos se destacan:

1.6.1. Crecimiento de los datos

Panesar[6] nos dice que el crecimiento de los datos en el universo digital se duplica cada dos años, lo cual implica un desafío en el almacenamiento.

1.6.2. Calidad de los datos

Es un reto la gestión de datos para mantenerlos y limpiarlos correctamente.

1.6.3. Infraestructura

En este ámbito, nos referimos a los recursos técnicos como el almacenamiento, ancho de banda, bases de datos, entre otros. Aquí el desafío reside en encontrar proveedores de servicios y soporte.

1.6.4. Experiencia

Es un desafío encontrar personal con experiencia de buena calidad en las disciplinas de análisis y ciencia de datos.

1.6.5. Seguridad

Es complicado lidiar con los riesgos de seguridad y privacidad en cierto tipo de datos; por ejemplo, en el área de la salud existe confidencialidad, por lo que se deben tratar dichos datos con mayor sensibilidad para procurar la privacidad.

1.7. Casos de uso

Las áreas con mayor volumen de datos son las que más uso hacen de los big data, entre las áreas más destacadas se encuentran [4]:

1.7.1. Consumo

Gran parte de las industrias de consumo recolectan datos de los clientes para mejorar sus ventas y brindar atención personalizada.

1.7.2. Financiera

En esta área se encuentran las entidades bancarias, aseguradoras u otras empresas de servicio financiero que buscan atraer y retener a sus clientes con mayor efectividad. Por otra parte, pueden facilitar ofertas, mejorar la detección de fraudes, gestión de riesgos, entre otros.

1.7.3. Sector público

El sector público incorpora el big data para creación de estrategias que respalden el medio ambiente, la seguridad pública, educación, política, entre otros.

1.7.4. Telecomunicaciones

Las empresas enfocadas en telecomunicaciones buscan nuevas perspectivas para ajustar la entrega de productos y servicios a las exigencias de los clientes mediante el análisis de redes sociales, sentimientos y en general, el análisis de los datos.

1.7.5. Administración

El uso de los big data en la administración ayuda en la toma de decisiones, esto mediante los datos que cuentan sobre la logística, el inventario, producción, defectos de fabricación, entre otros.

1.7.6. Comercio electrónico

Al recolectar las cantidades de texto, imágenes, el número de pulsaciones de clic dentro de una página y los datos de los clientes, es posible mejorar la eficacia y precisión del comercio; además, se le brinda al cliente una experiencia mejorada.

1.7.7. Salud

Los registros médicos, consultas, casos clínicos, investigaciones médicas de los hospitales u otro tipo de datos que son generados dentro de esta área permiten realizar predicciones, desarrollo de investigaciones, diagnósticos, desarrollo de fármacos, tratamientos personalizados, entre muchas otras aplicaciones.

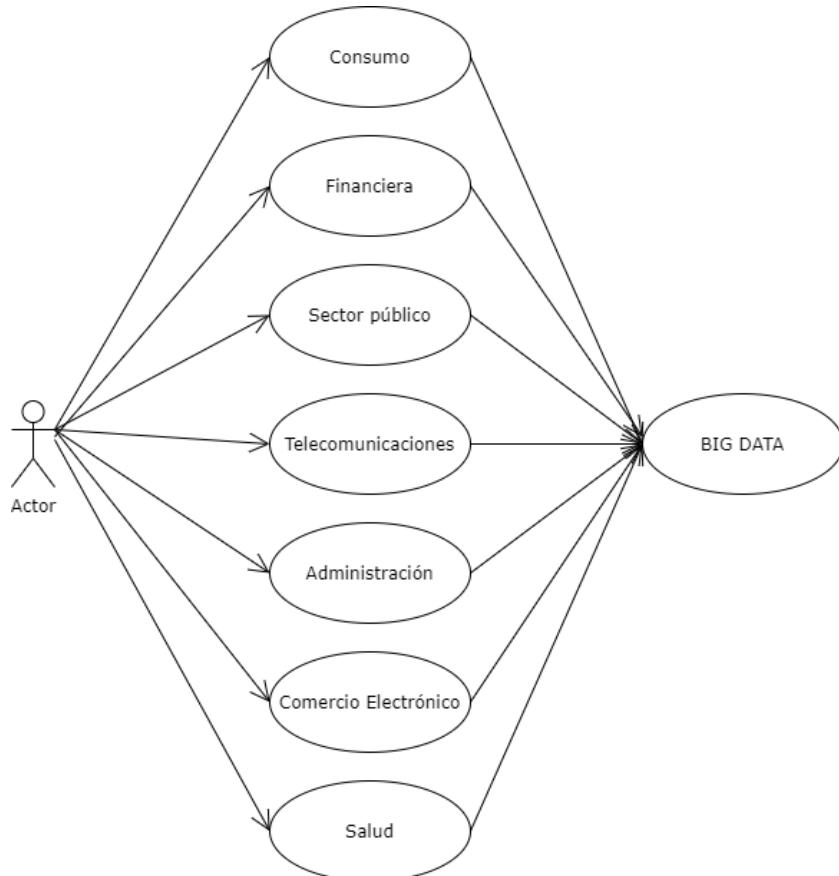


Figura 1.6: Casos de uso de big data.

1.8. Funcionamiento

Oracle[2] nos recomienda realizar tres acciones clave al comenzar a trabajar con los big data, las cuales son:

1.8.1. Integrar

Llevar a cabo la recolección de datos de diversas fuentes. En esta fase es necesario la incorporación, procesamiento y aseguramiento del formato de los datos.

1.8.2. Administrar

Realizar el almacenamiento de los datos, ya sea en la nube o en las instalaciones, para incorporar los requisitos de procesamiento y los motores de procesamiento necesarios a dichos conjuntos de datos.

1.8.3. Analizar

Comenzar con el análisis de datos, de ser posible, de forma visual. En esta fase es importante la construcción de modelos de datos con aprendizaje automático e inteligencia artificial.

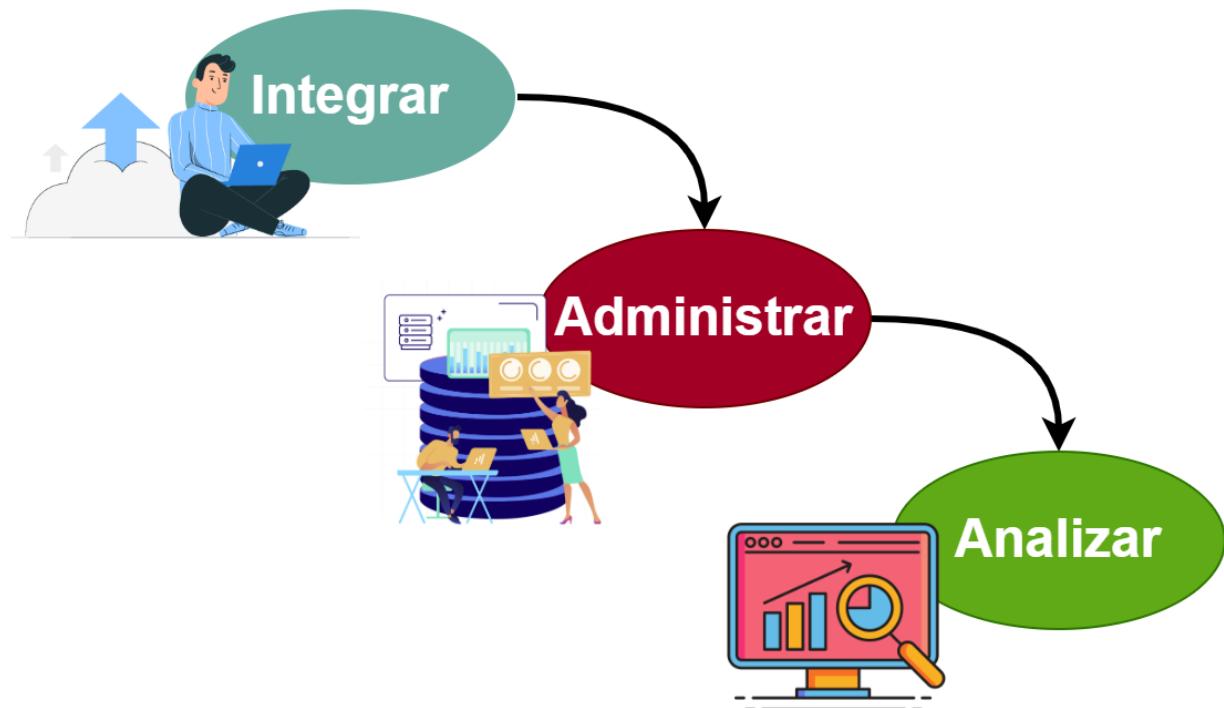


Figura 1.7: Acciones clave para trabajar con big data.

Capítulo 2

Ciencia de datos

2.1. Introducción

La ciencia de datos es la ciencia computacional de extraer información significativa de los datos en bruto y luego comunicar de manera efectiva las mismas para generar valor. Con lo antes mencionado, es posible comprender la importancia de la incorporación de la ciencia de los datos en los big data.

De manera simple, la ciencia de datos es la práctica de utilizar un conjunto de técnicas y metodologías analíticas para generar y comunicar información valiosa y procesable a partir de datos que no han sido procesados.

El objetivo de la ciencia de datos es optimizar los procesos y respaldar una toma de decisiones mejor informada de los datos, con lo que se genera un aumento del valor, ya sea que el valor esté representado por el número de vidas ahorradas, la cantidad de dinero ahorrada o el porcentaje de ingresos aumentado.

Gracias a la ciencia de los datos, tenemos la capacidad de predecir comportamientos futuros, descubrir patrones, proporcionar información procesable o extraer un significado importante de los datos sin explotar.[3]

2.2. Metodología CRISP-DM

Hoy en día, la ciencia de los datos es la clave para hacer que los grandes volúmenes de datos sean útiles. Para darle utilidad a los big data en este trabajo se utilizará la metodología CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, el cual es un método probado para orientar sus trabajos de minería de datos.[7]

La metodología CRISP-DM[8] consta de seis etapas, las cuales se presentan a continuación.

2.2.1. Comprensión del problema o negocio

Por muchos, considerada la etapa más importante, ya que si no se tiene una correcta comprensión del problema, o negocio, no es posible comenzar a trabajar con las siguientes etapas. Las actividades principales de esta etapa son:

- Identificación del problema.
- Determinación de objetivos.
- Evaluación de la situación actual.

2.2.2. Comprensión de datos

Entre sus principales actividades se destacan:

- Recolección de datos.
- Descripción de datos.
- Exploración de datos.

2.2.3. Preparación de datos

Generalmente esta es la etapa que consume más tiempo en el proyecto, y es donde se seleccionan los datos que se transforman de acuerdo con los resultados de la etapa anterior a fin de utilizarlos en la etapa de modelado. Las actividades principales de esta etapa son:

- Limpieza de datos.
- Creación de indicadores.
- Transformación de datos.

2.2.4. Modelado

En esta etapa se obtiene propiamente el modelo de minería de datos. Se centra en realizar actividades como:

- Selección de técnica de modelado.
- Selección de datos de prueba.
- Obtención del modelo.

2.2.5. Evaluación del modelo

Dentro de esta etapa se determina la calidad del modelo con base en el análisis de ciertas métricas estadísticas del mismo, comparando los resultados con resultados previos, o bien, analizando los resultados con apoyo de expertos en el dominio del problema. De acuerdo con los resultados de esta etapa se determina seguir con la última fase de la metodología, regresar a alguna de las etapas anteriores o incluso partir de cero con un nuevo proyecto.

2.2.6. Implementación del modelo

Esta etapa explota, mediante acciones concretas, el conocimiento adquirido mediante el modelo. Aquí también es importante documentar los resultados de manera clara para el usuario final y asegurarse de que todas las etapas de la metodología se documenten debidamente para hacer una revisión del proyecto a fin de obtener lecciones aprendidas durante el proceso. Asimismo, observa las acciones para detectar áreas de oportunidad o incluso nuevos problemas.

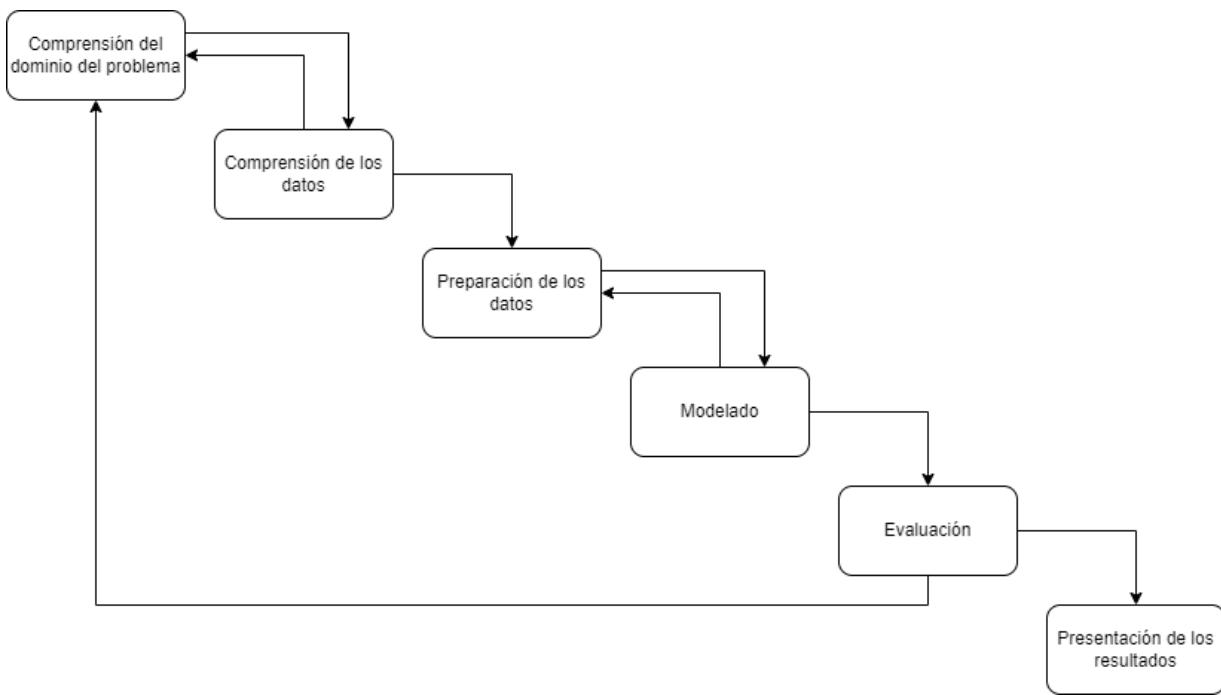


Figura 2.1: Diagrama de secuencia sobre la metodología CRISP-DM.

2.3. Modelos dentro de la fase del modelado

2.3.1. Red Neuronal

Según Berzal[9], "una red neuronal es un modelo con múltiples parámetros, cuyos valores se ajustan mediante un algoritmo de entrenamiento con la ayuda de un conjunto de

datos de entrenamiento.”

Por otra parte, IBM[10] nos dice que, ”una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas.”

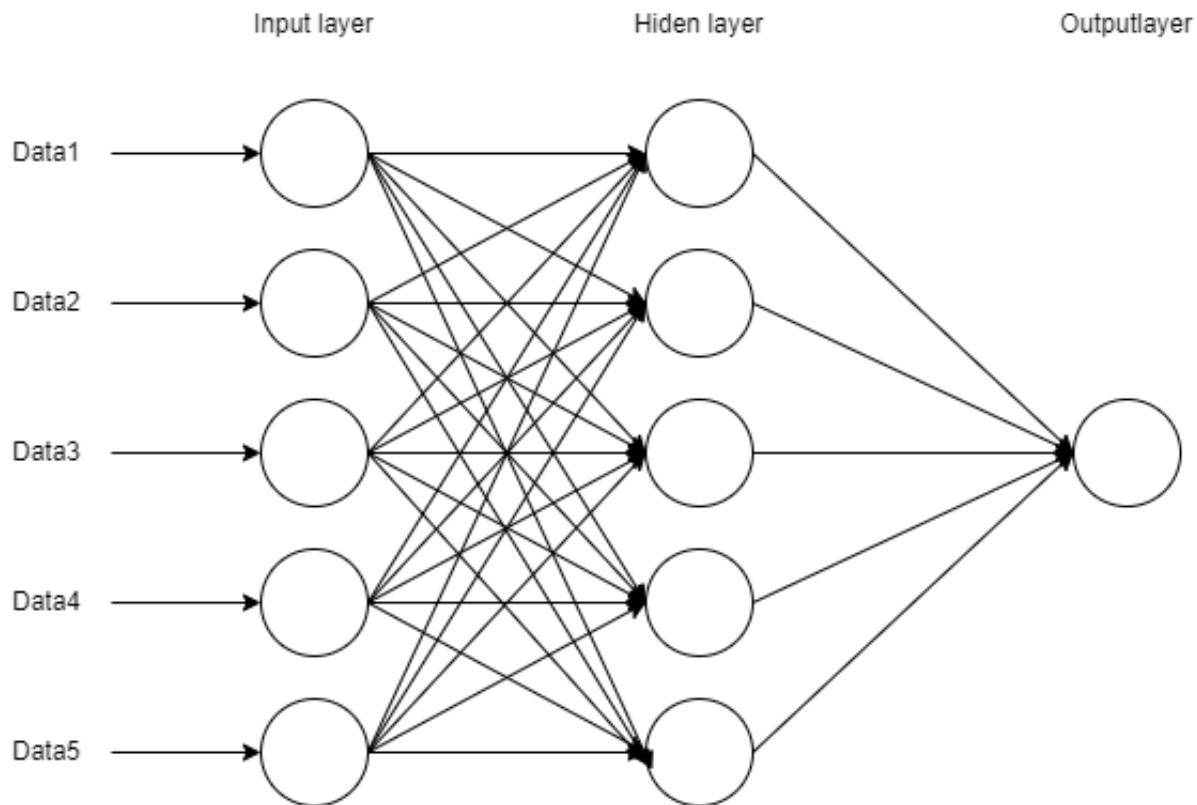


Figura 2.2: Estructura de una red neuronal.

Tal y como podemos ver en la figura 2.2, dentro de la estructura de una red neuronal tenemos:

- Una capa de entrada, con unidades que representan los campos de entrada.
- Una o varias capas ocultas; y una capa de salida.
- Una unidad o varias unidades que representan el campo o los campos de destino.

Las unidades se conectan con fuerzas de conexión variables. Los datos de entrada se presentan en la primera capa, y los valores se propagan desde cada neurona hasta cada neurona de la siguiente capa. Finalmente, se envía un resultado desde la capa de salida.

La red aprende examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada.

Las redes neuronales tienen tres características fundamentales [11]:

- **Aprender:** Adquirir el conocimiento de una cosa por medio del estudio, ejercicio o experiencia.
- **Generalizar:** Extender o ampliar una cosa.
- **Abstraer:** Aislar mentalmente o considerar por separado las cualidades de un objeto.

Tal y como se menciona en uno de los puntos anteriores, las redes neuronales artificiales tienen la capacidad del aprendizaje, para aprender deben tener un respectivo entrenamiento.

El objetivo de entrenar una red neuronal artificial es garantizar que una aplicación dada, para un conjunto dado de entradas, produzca el conjunto de salidas mínimamente consistente o deseado. El proceso de entrenamiento consiste en aplicar secuencialmente diferentes conjuntos o vectores de entrada para que los pesos de las asociaciones se ajusten según un procedimiento predeterminado. Durante el entrenamiento, los pesos convergen gradualmente a valores que hacen que cada entrada produzca el vector de salida deseado.[11]

Los algoritmos de entrenamiento se pueden clasificar en dos grupos:

- **Entrenamiento supervisado:** El entrenamiento consiste en exponer un vector de entrada a la red, calcular la salida de la red, compararla con la salida esperada, y el error o diferencia resultante se utiliza para realimentar la red y cambiar los pesos de acuerdo con un algoritmo que tiende a minimizar el error.
- **Entrenamiento no supervisado:** Son modelos de aprendizaje más lógicos en sistemas biológicos. El algoritmo de entrenamiento cambia los pesos de la red para producir vectores de salida consistentes. El proceso de entrenamiento extrae las propiedades estadísticas del conjunto de vectores de entrenamiento y agrupa vectores similares en clases.

La mayor parte de los algoritmos de entrenamiento surgieron de la evolución del modelo de aprendizaje no supervisado que propuso Hebb. [11] La ley de Hebb se representa de la siguiente forma:

$$w_{ij}(n+1) = w_{ij}(n) + \alpha OUT_i OUT_j \quad (2.1)$$

2.3.2. Árbol de decisión

Un árbol de decisión es un modelo predictivo cuya finalidad es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Además, sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema.[12]

Toda la información obtenida durante el proceso de aprendizaje inductivo se representa mediante un árbol, el cual se representa por un conjunto de nodos, hojas y ramas.

Su nodo raíz es el atributo a partir del cual se inicia el proceso de clasificación, cada nodo interno corresponde a una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo.[12]

Lo antes mencionado se representa en la siguiente figura.

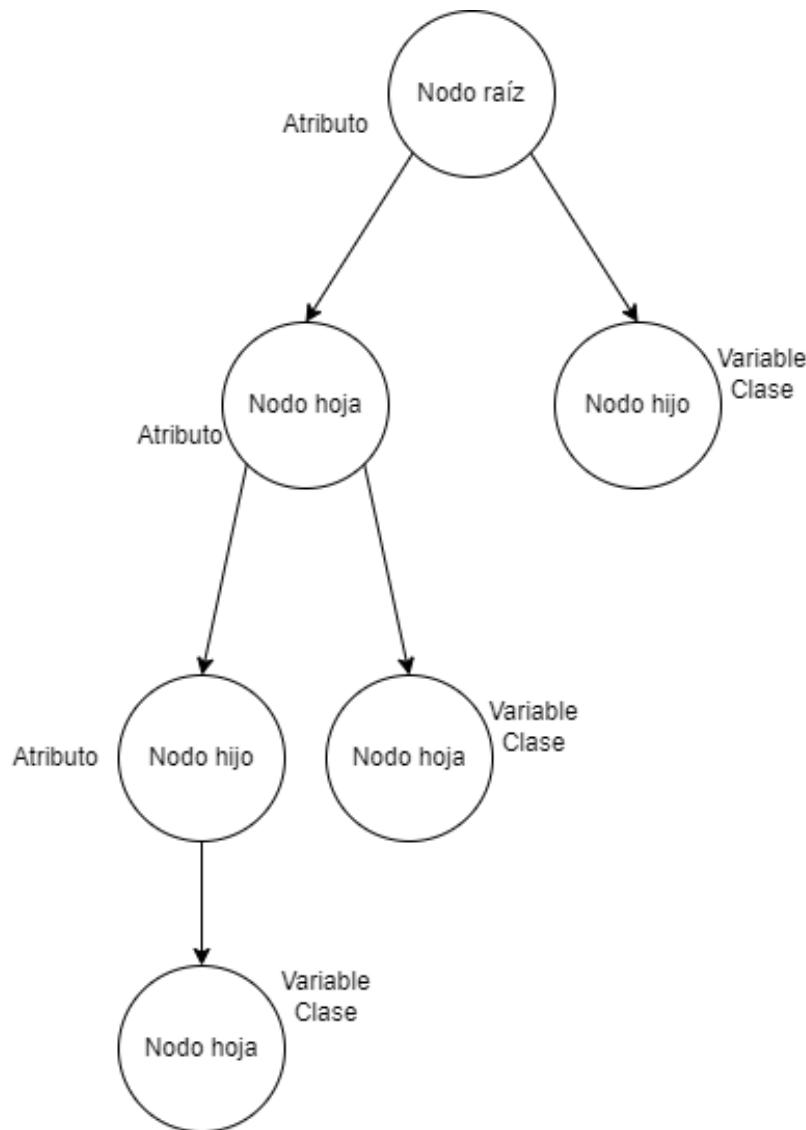


Figura 2.3: Estructura de una árbol de decisión.

Tal y como se presenta en la figura 2.3, las ramas que salen de cada uno de los nodos se encuentran etiquetadas con posibles valores del atributo, los nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase a resolver.

Los árboles de decisión se construyen a partir de la descripción narrativa de un problema, ya que provee una visión gráfica de la toma de decisión, especificando las variables que son evaluadas, las acciones que deben ser tomadas y el orden en el que la toma de decisión será efectuada.

Al ejecutar un árbol de decisión, solo un camino será seguido dependiendo del valor actual de la variable evaluada.

El algoritmo de generación de árboles de decisión consta de dos etapas:

- **La inducción del árbol:** Se construye el árbol de decisión a partir del conjunto de entrenamiento, donde cada nodo interno del árbol se compone de un atributo de prueba y la porción del conjunto de entrenamiento presente en el nodo es dividida de acuerdo con los valores que pueda tomar ese atributo. Su construcción inicia con la generación de su nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de entrenamientos en dos o más subconjuntos, para cada porción se genera un nuevo nodo y así sucesivamente.

Si en un nodo se tienen objetivos de más de una clase, se genera un nodo interno, cuando contiene objetivos de una sola clase, se genera una hoja a la que se le asigna la etiqueta de clase.

- **La clasificación:** Cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja, a partir de la que se determina la membresía del objeto a alguna clase. El camino a seguir en el árbol lo determinan las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente en él.

Dichos algoritmos suelen trabajar mediante la metodología *top-down*, eligiendo en cada fase la variable que mejor divide el conjunto de elementos [13]. No obstante, cada algoritmo utiliza diversos tipos de métricas para medir de manera óptima, estos se encargan de medir la homogeneidad de la variable de destino dentro de los subconjuntos.

2.3.3. Algoritmo de regresión

Este tipo de algoritmos son utilizados para modelar las relaciones entre las entidades de un conjunto de datos.[3]

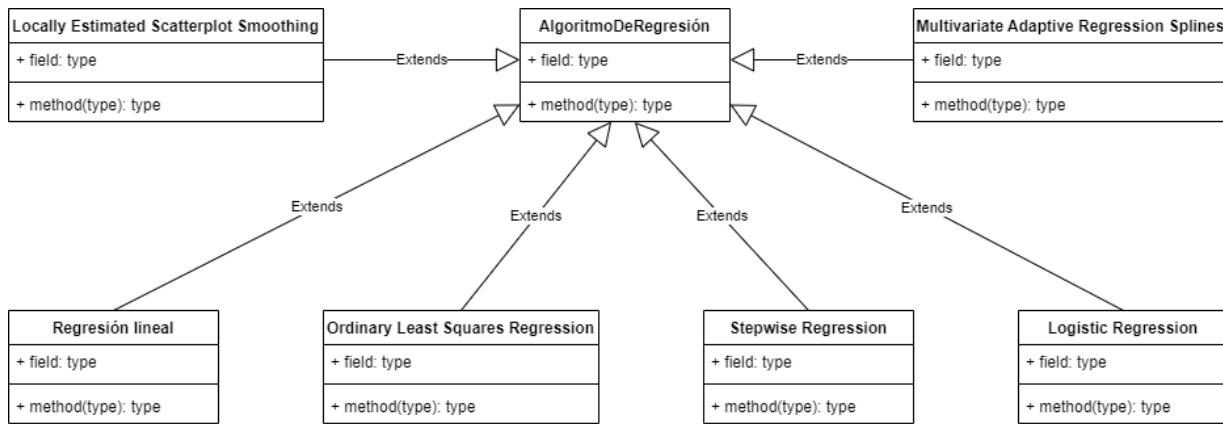


Figura 2.4: Principales algoritmos de regresión.

Mediante los algoritmos de regresión, es posible predecir valores futuros a partir de valores históricos; sin embargo, tienen restricciones. Los métodos de regresión asumen una relación de causa y efecto entre variables, pero las circunstancias actuales siempre están sujetas a flujo, esto implica que la predicción de valores futuros a partir de valores históricos generará resultados incorrectos cuando cambien las circunstancias presentes.

Como se muestra en la figura 2.4, existen múltiples tipos de algoritmos de regresión, a continuación, se abordarán los más utilizados.

Regresión lineal: Es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos. De la misma manera, es un método de aprendizaje automático que se puede utilizar para describir y cuantificar la relación entre la variable de destino y las características del conjunto de datos que ha elegido utilizar como variables predictoras. Cuando usas solo una variable como predictor, la regresión lineal puede ser tan simple como la fórmula:

$$y = mx + b \quad (2.2)$$

Tipos de regresión lineal:

- **Regresión lineal simple:** Modelos que utilizan un único predictor. La ecuación general es:

$$Y = \beta_0 + \beta_i X + \epsilon_i \quad (2.3)$$

- **Regresión lineal múltiple:** Modelos que utilizan múltiples predictores. Esta regresión tiene múltiples variables para predecir la respuesta. Este es un ejemplo de la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (2.4)$$

- **Regresión lineal multivariante:**[14] Modelos para varias variables de respuesta. Esta regresión tiene múltiples salidas que derivan de los mismos datos de salida . Se expresan con fórmulas diferentes. Este es un ejemplo del sistema con 2 ecuaciones:

$$Y_1 = \beta_{01} + \beta_{11} X_1 + \epsilon_1 \quad Y_2 = \beta_{02} + \beta_{12} X_1 + \epsilon_2 \quad (2.5)$$

Logistic regression: Es un método de aprendizaje automático que se puede utilizar para estimar los valores de una variable de destino categórica basada en las características seleccionadas. La variable de destino debe ser numérica y contener valores que describen la clase de destino o la categoría. Una cosa interesante sobre la regresión logística es que, además de predecir la clase de observaciones en su variable objetivo, indica la probabilidad de cada una de sus estimaciones. Este modelo tiene la forma[15]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (2.6)$$

Ordinary least squares (OLS) regression: Es un método estadístico que ajusta una línea de regresión lineal a un conjunto de datos. Con este método es posible hacer esto cuadrando los valores de distancia vertical que describen las distancias entre los puntos de datos y la línea de mejor ajuste, sumando esas distancias cuadradas y, a continuación, ajustando la ubicación de la línea de mejor ajuste para minimizar el valor de distancia cuadrada sumada. Es utilizado si se desea construir una función que sea una aproximación cercana a los datos. La ecuación del modelo se muestra a continuación[16]:

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon \quad (2.7)$$

Capítulo 3

Ciencias de la vida

3.1. Introducción

Las ciencias de la vida es el área que se compone de todo aquello que, de alguna forma u otra, estudia la vida en sus diferentes manifestaciones, ya sea por las ciencias biológicas, químicas y de la salud. Algunas de las ciencias que lo componen son:

- Biología
- Biomecánica
- Bioquímica
- Botánica
- Ciencias Agrogenómicas
- Ciencias Ambientales
- Ciencias Genómicas
- Ecología
- Genética
- Medicina
- Neurociencias
- Psicología
- Química
- Zoología

Como se presentó en la sección 1.7, los big data tienen gran impacto dentro del área de la salud, específicamente hablando, dentro de la medicina. La medicina se define como "La ciencia y el arte que se ocupa del mantenimiento de la salud y la prevención, el alivio o la cura de enfermedades." [17]

Con lo antes mencionado y la información presentada en el capítulo 1, es posible comprender la estrecha relación que existe entre los big data y el área de la salud.

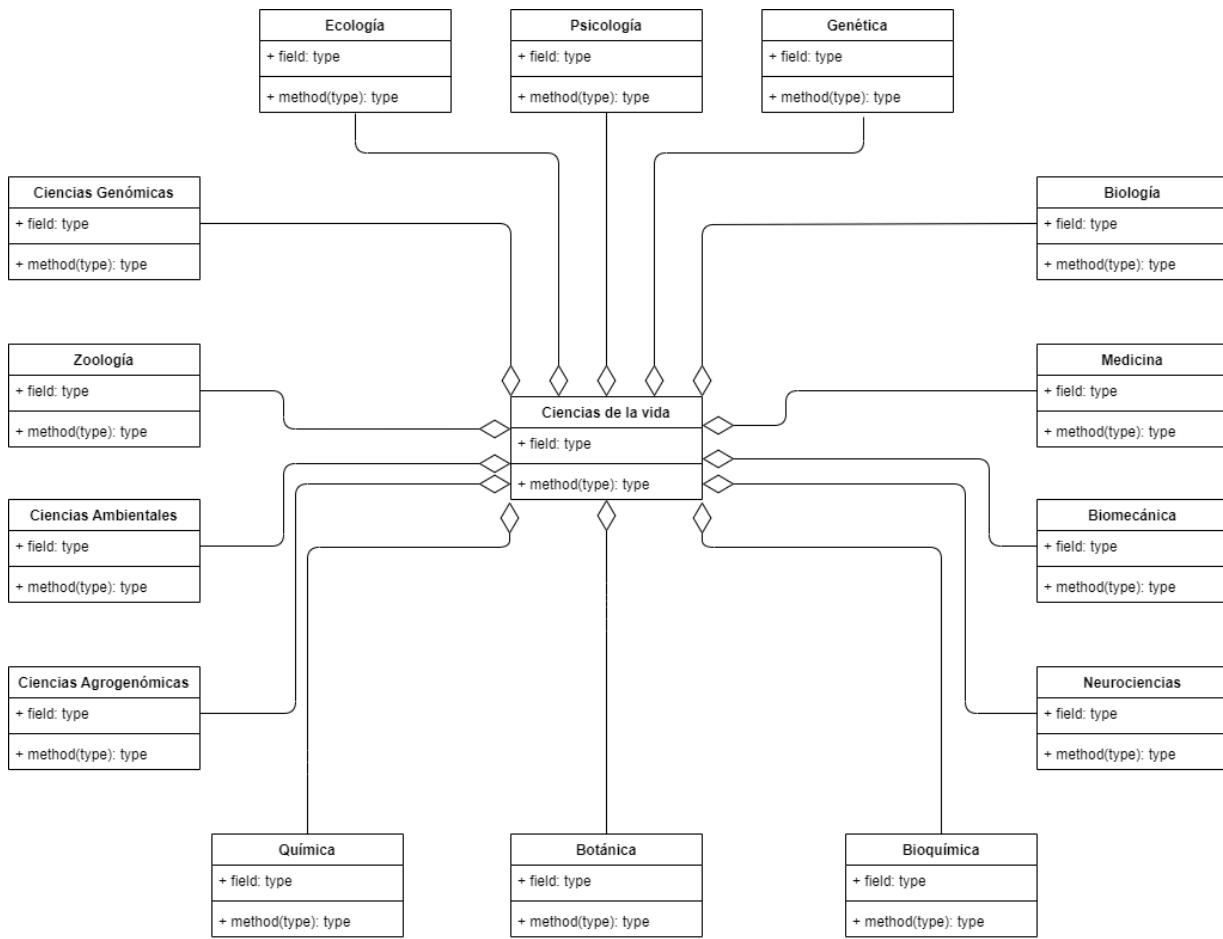


Figura 3.1: Principales ciencias que componen las ciencias de la vida.

3.2. El área de salud

En la actualidad, el área de la salud genera grandes volúmenes de datos que, haciendo uso de aprendizaje automático e inteligencia artificial, es posible transformar el sector y mejorar los resultados.[6]

Si bien es cierto que los agentes artificialmente inteligentes no lograrán sustituir a los médicos y enfermeras, sí lograrán mejorar el diagnóstico de los pacientes, ayudarán a predecir resultados y comenzarán a rayar la superficie de atención personalizada.

Se prevé que en la próxima década, será posible el desarrollo de planes de tratamiento de salud y estilo de vida desde el comienzo de la vida; es decir, desde el embrión. Por otra parte, con el manejo de los datos del perfil genético, será más rápido detectar posibles problemas y alterar o eliminar posibles defectos genéticos o características desfavorecidas.[6]

En el caso de las enfermedades, se tendrá la posibilidad de predecirlas y desarrollar planes de cuidado de vida saludable, esto gracias a un monitoreo constante de los datos del paciente mediante diversos medios.

3.3. Diagnóstico médico

El Instituto Nacional del Cáncer[18] en Estados Unidos proporciona la definición de diagnóstico como "Proceso en el que se identifica una enfermedad, afección o lesión por sus signos y síntomas. Para ayudar a hacer un diagnóstico, se pueden utilizar los antecedentes de salud o realizar un examen físico y pruebas, como análisis de sangre, pruebas con imágenes y biopsias."

Dentro de esta área existen diferentes técnicas diagnósticas, nuevamente, el Instituto Nacional del Cáncer[19] define la técnica diagnóstica como "Tipo de método o prueba que se usa como ayuda para diagnosticar una enfermedad o afección. Las pruebas de imaginología y las pruebas para medir la presión arterial, pulso y temperatura son ejemplos de técnicas diagnósticas."

3.4. Características del diagnóstico médico

Se considera que una prueba diagnóstica es buena cuando ofrece resultados positivos en pacientes enfermos y negativos en pacientes sanos con el menor rango de error posible. Por lo tanto, las condiciones que deben ser exigidas en un test diagnóstico son principalmente tres[20]:

3.4.1. Validez

Es la frecuencia con la que los resultados obtenidos con este test pueden ser confirmados por otros más complejos y rigurosos. Los parámetros que miden la validez de una prueba diagnóstica son la sensibilidad y la especificidad.

3.4.2. Reproductividad

Es la capacidad de un test de ofrecer los mismos resultados cuando se repite su aplicación en circunstancias similares.

3.4.3. Seguridad

Es la certeza de un test ya que predecirá la presencia o ausencia de enfermedad en un paciente.

3.5. Tipos de diagnóstico médico

Con lo antes mencionado, es indispensable recordar algunas premisas básicas sobre las que se apoya el diagnóstico médico, las cuales son[21]:

3.5.1. Diagnóstico etiológico

Su objetivo se basa en identificar los síntomas y signos para proceder a diagnosticar una enfermedad.

3.5.2. Diagnóstico sintomatológico

Es el reconocimiento de las causas de la enfermedad en cuestión, en nuestros tiempos casi siempre obtenida por estudios de laboratorio; por ende, su objetivo se centra en descubrir las causas o factores potencialmente reversibles que pueden beneficiarse de un tratamiento.

3.5.3. Diagnóstico diferencial

Se enfoca en determinar la enfermedad mediante la exclusión de otras posibles causas.

3.5.4. Diagnóstico genérico

Cuyo objetivo es simplemente determinar si el paciente sufre de alguna enfermedad o no.

3.5.5. Diagnóstico nosológico

Tiene por objetivo enunciar a la enfermedad, con la ventaja de dar por hecho su origen y sus características.

3.5.6. Diagnóstico patogénico

Su objetivo es la determinación específica de la enfermedad y el enunciado de los mecanismos que producen dicha enfermedad.

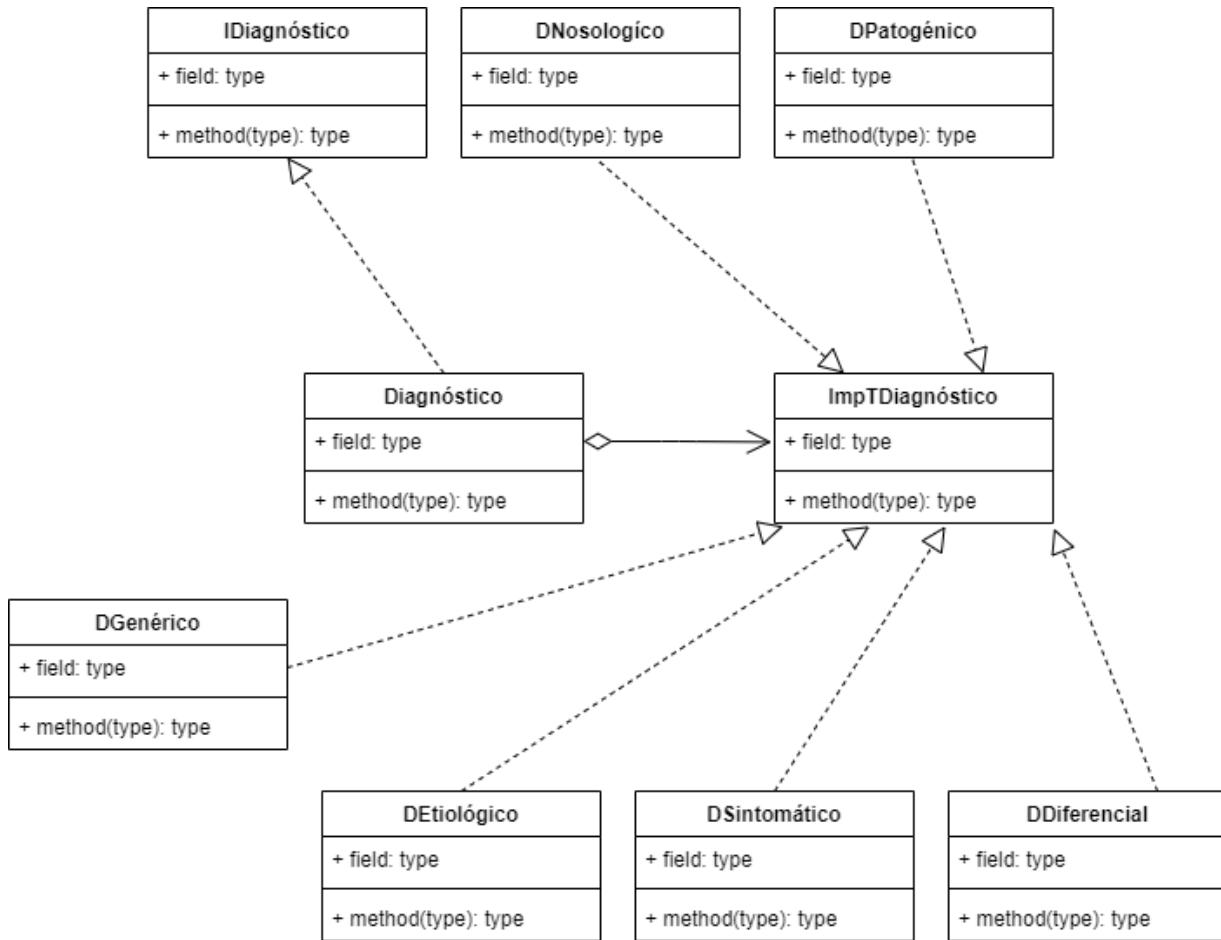


Figura 3.2: Patrón de diseño Bridge enfocado a los diagnósticos.

3.6. Predicción de enfermedades

Actualmente, los datos referidos al área clínica son utilizados por el médico para obtener información del paciente, estos les permiten efectuar diagnósticos, tratamientos, medición u otros; además, se generan datos por los propios sistemas informáticos para generar múltiples alertas como pueden ser interacciones medicamentosas, control sobre medicamentos contraindicados para una patología específica, resultados anormales en estudios, etc.[6]

La cantidad de información que disponen las instituciones de salud sobre sus pacientes, ya sea tanto a nivel administrativo como asistencial es inmensa. Como lo hemos leído en párrafos anteriores, gracias a la recopilación de dicha información es posible la creación de modelos predictivos, de clasificación, de agrupación y entre muchas otras aplicaciones.

Básicamente, el análisis del conjunto de datos antes mencionado nos permite realizar:

- Predicción de expectativas de vida.

- Identificación y clasificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.
- Atención médica personalizada.
- Identificación de terapias médicas y tratamientos erróneos para determinadas enfermedades.

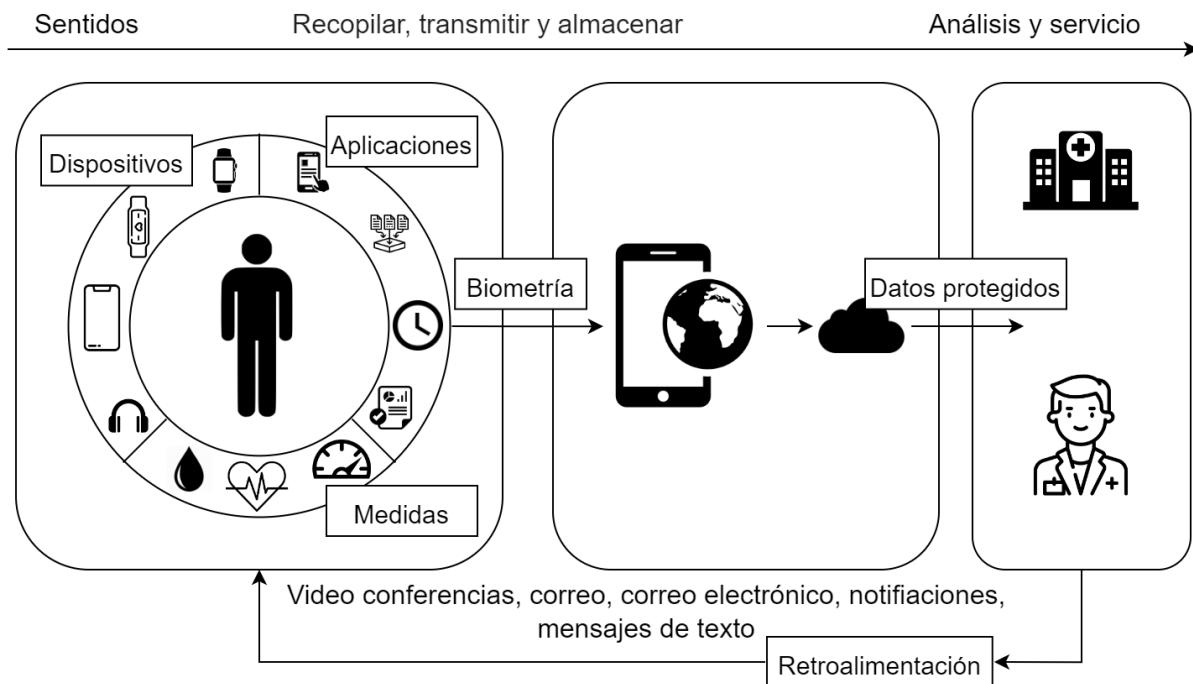


Figura 3.3: Proceso de recolección de datos en el área de la salud.

3.7. Casos de estudio de los big data aplicados al área de la salud

Tabla 3.1: Casos de estudio - aplicaciones de los big data en el área de la salud.

Propuesta	Descripción
Identificación de síntomas en personas asmáticas.	Propeller Health es una compañía que, mediante un sensor conectado a inhaladores y espirómetros para personas con asma o EPOC, rastrea las condiciones ambientales en las ubicaciones de los sensores y envía informes a los teléfonos de los pacientes, para que puedan comprender mejor las causas de sus síntomas y tomar medidas para prevenir ataques.
Detección de fibrilación auricular.	Apple, una empresa a nivel global, se ha asociado con investigadores de Stanford para determinar si el sensor cardíaco del Apple Watch se puede usar para detectar la fibrilación auricular, una condición que causa la muerte cientos de miles de estadounidenses al año. Esta investigación determinaría si una persona es diagnosticada con la enfermedad.
Reducir errores de prescripción.	MedAware, una StartUp israelí, se está asociando con organizaciones de atención médica para implementar su herramienta de apoyo a la toma de decisiones que utiliza big data para detectar errores de prescripción antes de que ocurran.
Diagnóstico temprano de Alzheimer.	Con la creación de un modelo de clasificación para un gran volumen de información longitudinal de modalidades de imágenes no invasivas, es posible realizar un diagnóstico temprano de la enfermedad.
Servicio diagnóstico virtual.	Aetna Health App, una aplicación para dispositivos móviles, asesoran a los pacientes sobre su afección médica utilizando datos agregados y pueden recomendar a los pacientes que busquen atención médica en función de la información recibida en la aplicación.
Manejo de la pandemia COVID-19.	Es posible emplearlo para el diagnóstico e identificación de población que está en mayor riesgo de contagio. También se emplea para el desarrollo más rápido de medicamentos, incluyendo el estudio de reutilización de medicamentos que han sido probados para el tratamiento de otras enfermedades.

Tal y como se presenta en la tabla 3.1, se demuestra que en la actualidad es posible hacer uso de los big data como apoyo para la toma de decisiones médicas hasta realizar predicciones de enfermedades.

3.8. Especialidades en el área de salud

El área de la salud cuenta con múltiples especialidades que, con el paso de los años, van incrementando y evolucionando. Actualmente, entre las más conocidas se encuentran:

- Andrología
- Anestesiología
- Angiología
- Artrología
- Bacteriología
- Bioquímica
- Cardiología
- Ciencia veterinaria
- Cirugía General
- Dermatología
- Diabetología
- Endocrinología
- Enfermedades comunicativas
- Enfermedades infecciosas
- Epidemiología
- Farmacología clínica
- Gastroenterología
- Ginecología
- Hematología
- Hepatología
- Inmunología
- Medicina de Terapia Intensiva
- Medicina deportiva
- Medicina general

- Medicina Geriátrica
- Medicina Interna
- Miología / Sarcología
- Nefrología
- Neumología
- Neurología
- Obstetricia
- Odontología
- Oftalmología
- Oncología
- Osteopatía / Ortopedia
- Otología
- Otorrinolaringología
- Pediatría
- Psiquiatría
- Radiología
- Reumatología
- Rinología
- Sexología
- Teratología
- Tricología
- Urología
- Virología

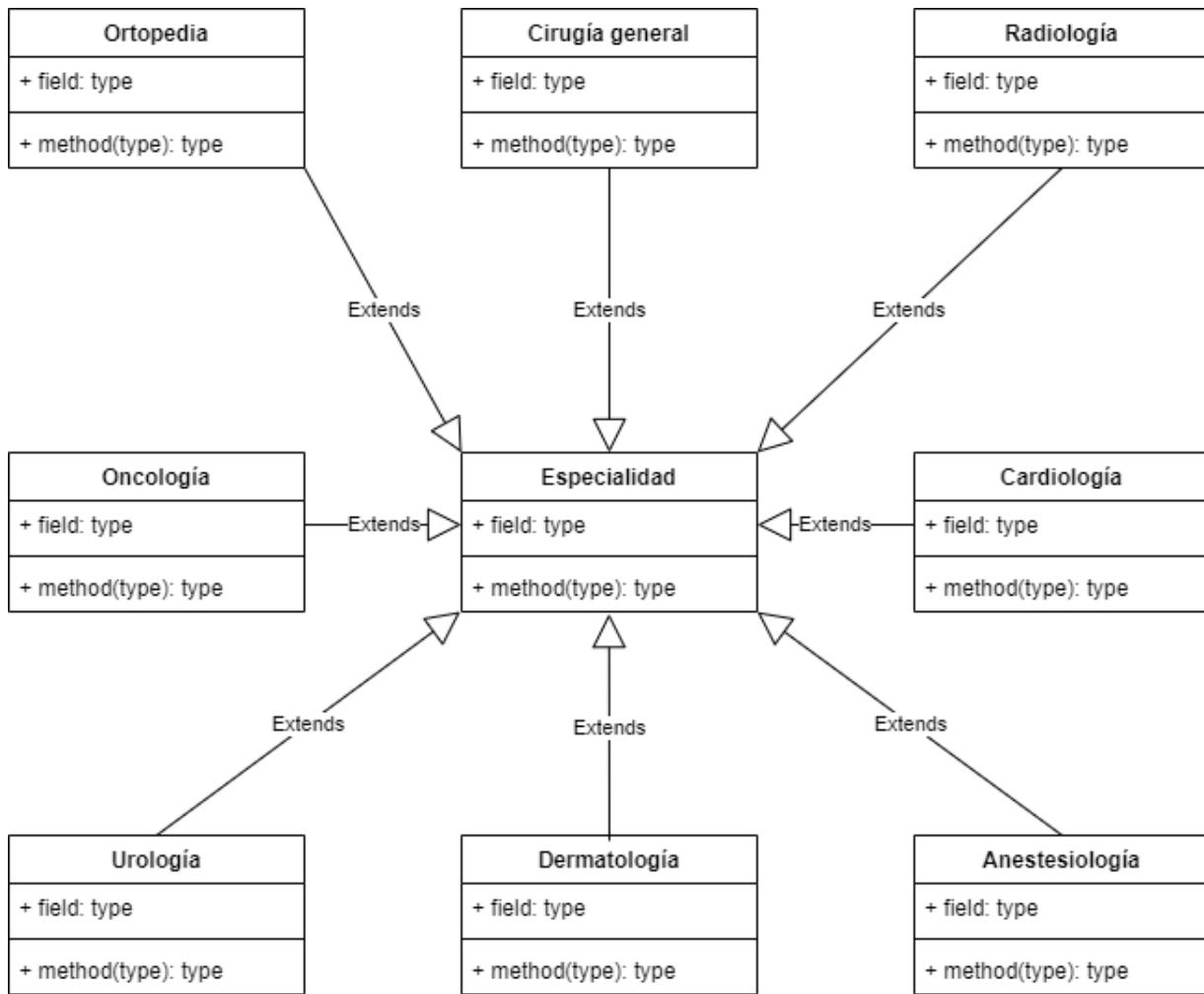


Figura 3.4: Especialidades en el área de la Salud.

3.9. Oncología

Según el Instituto Nacional del Cáncer[22] en Estados Unidos, la oncología se define como "rama de la medicina especializada en el diagnóstico y tratamiento del cáncer. Incluye la oncología médica (uso de quimioterapia, terapia con hormonas y otros medicamentos para tratar el cáncer), la radioncología (uso de radioterapia para tratar el cáncer) y la oncología quirúrgica (uso de cirugía y otros procedimientos para tratar el cáncer)."

Como se ha mencionado a lo largo del capítulo, los big data podrían ser útiles para desarrollar y remodelar estrategias de prevención de enfermedades.

La combinación de grandes conjuntos de datos genómicos y datos ambientales ayudará a predecir qué individuos o grupos están en riesgo de desarrollar ciertas enfermedades crónicas y cáncer. Esto podría provocar especificaciones destinadas a influir en los factores ambientales y el comportamiento que contribuyen a los riesgos para la salud en los grupos

destinatarios. Los big data también serán útiles para evaluar los programas de prevención actuales y podrían ayudar a identificar nuevos conocimientos para mejorarlo.

Por otra parte, en un entorno terapéutico, los big data son fundamentales para controlar, por ejemplo, los efectos de terapias específicas, especialmente en relación con las características del paciente y del tumor.

Esto ayudará a mejorar la medicina de precisión y a generar conocimientos importantes para calcular el costo-eficiencia de ciertos regímenes de tratamiento.

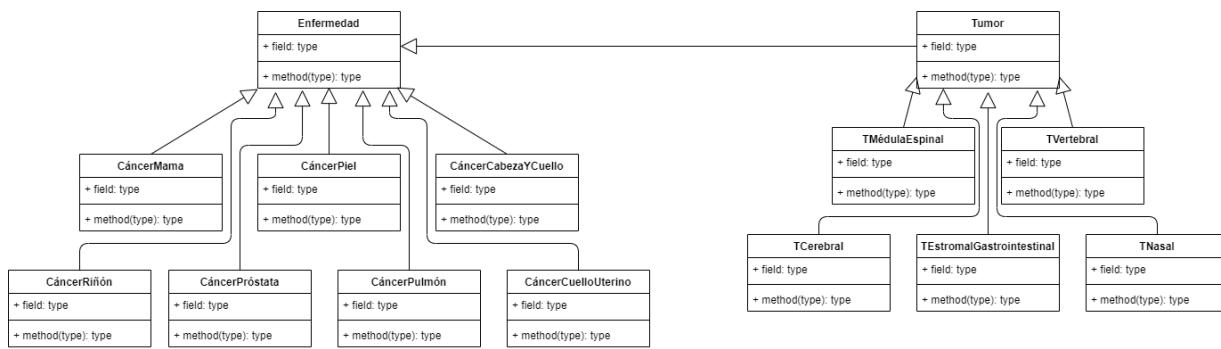


Figura 3.5: Enfermedades más comunes dentro de Oncología.

3.10. Cardiología

En este caso, el Instituto Nacional del Cáncer[23] en Estados Unidos, define la cardiología como "rama de la medicina que se especializa en el diagnóstico y tratamiento de enfermedades del corazón, los vasos sanguíneos y el sistema circulatorio. Estas enfermedades incluyen enfermedad de las arterias coronarias, problemas del ritmo del corazón e insuficiencia cardíaca."

Los análisis de big data y las aplicaciones de Machine Learning se utilizan en la cardiología por muchas razones, como por ejemplo para la toma de decisiones clínicas, la detección del factor de riesgo de enfermedades cardiovasculares, y para la medicina de precisión mediante información genómica.

El análisis de big data puede producir una predicción más poderosa de resultados que van desde la mortalidad hasta los resultados informados por el paciente y la utilización de recursos y, por lo tanto, podría ser más práctico desde el punto de vista clínico.

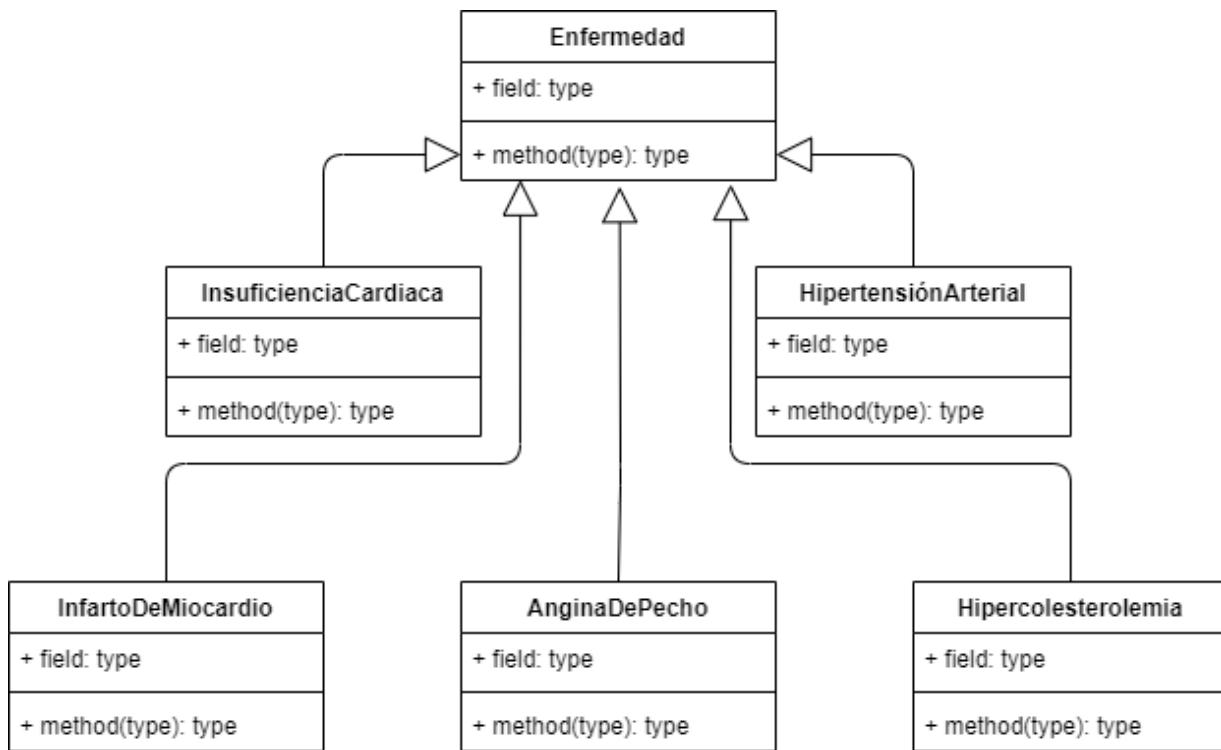


Figura 3.6: Enfermedades más comunes dentro de Cardiología.

Capítulo 4

Aplicación de la metodología CRISP-DM

A lo largo de este capítulo, se llevarán a cabo la aplicación de cuatro de las seis etapas de la metodología CRISP-DM para un determinado caso de estudio, las cuales son:

- Comprensión del dominio del problema.
- Comprensión de los datos.
- Preparación de los datos.
- Modelado.

Dentro de la etapa del *Modelado*, se comentará brevemente la evaluación de los modelos al igual que la presentación de los resultados.

4.1. Comprensión del dominio del problema

4.1.1. Determinación de los objetivos del proyecto

- Conseguir que, mediante la información médica de una persona, sea posible realizar el diagnóstico de enfermedades oncológicas y cardiológicas.
- Salvaguardar la vida de una persona por medio de la predicción de enfermedades.
- Por otra parte, identificar patrones que caracterizan a las personas que son propensas a ser diagnosticadas con dichas enfermedades.

4.1.2. Valoración de la situación actual del objetivo del proyecto

¿Se comprende de forma clara el problema que se intenta abordar?

Sí, el diagnóstico médico de enfermedades se refiere a realizar un proceso en el que se identifica una enfermedad, afección o lesión por sus signos y síntomas. Los diagnósticos

médicos permiten que los expertos en el área de la salud elaboren un tratamiento que resguarde la vida del paciente, es aquí donde surge la importancia de realizar la predicción de dichas enfermedades dentro de oncología y cardiología.

Dado el alto índice de mortalidad en las enfermedades de dichas ramas de la medicina, ofrecer un diagnóstico médico permitirá que los pacientes aumenten sus probabilidades de tener un tratamiento.

¿Existen datos disponibles para efectuar el análisis?

Sí, se contará con data set extraídos de sitios web con fines científicos, los cuales son formados por un conjunto de tablas que contienen datos personales de los usuarios, datos de sus signos vitales, datos sobre lesiones y/o malformaciones corporales.

¿Se dispone de recursos humanos y tecnológicos para desarrollar el proyecto?

Sí, por la parte de los Recursos humanos, tenemos:

- Especialista en el área de la salud.
- Especialista en minería de datos.
- Especialista en aprendizaje automatizado.

Mientras que, los recursos tecnológicos son:

- IBM SPSS Modeler
- R Studio
- Python

¿Se han identificado factores de riesgo que afecten el desarrollo del proyecto?

De momento no se han identificado riesgos.

4.1.3. Determinación de los objetivos de minería de datos

Problema de predicción: A partir de los datos disponibles de los data set, elaborar un modelo de predicción para diagnosticar la posibilidad de tener una enfermedad oncológica y/o cardiológica.

Problema de clasificación: Construir un modelo que permita clasificar la enfermedad cardiológica u oncológica según el diagnóstico de un paciente determinado.

4.1.4. Propuesta del enfoque metodológico (plan de proyecto de minería de datos)

Tabla 4.1: Plan de proyecto.

Fase	Tiempo a dedicar	Recursos humanos y tecnológicos	Riesgos atribuibles
Comprensión del dominio del problema.	Cuatro semanas	Experto en el área de la salud, experto en minería de datos.	No se han registrado riesgos.
Comprensión de los datos.	Cuatro semanas	Experto en el área de la salud, experto en minería de datos. Tablas, gráficos, resúmenes estadísticos que faciliten la comprensión de los datos. Uso de IBM SPSS Modeler, Python y R Studio.	La precisión del modelo de predicción podría ser baja debido a la escasa cantidad de registros de los data set.
Preparación de los datos.	Tres semanas	Experto en el área de la salud, experto en minería de datos. Herramientas para el análisis exploratorio de datos. Uso de IBM SPSS Modeler, Python y R Studio.	No se han registrado riesgos.
Modelado.	Tres semanas	Experto en minería de datos para aplicar en Big data, experto en técnicas de machine learning. Herramientas para la implementación de modelos de machine learning. Uso de IBM SPSS Modeler y Python.	No se han registrado riesgos.
Evaluación.	Una semana	Experto en el área de la salud, experto en minería de datos. Uso de IBM SPSS Modeler, Python y R Studio.	No se han registrado riesgos.
Presentación.	Una semana	Experto en el área de la salud, experto en minería de datos. Uso de IBM SPSS Modeler, Python y R Studio.	No se han registrado riesgos.

Tal y como se muestra en la tabla 4.1, vemos que en un periodo de tres meses se tiene pensado concluir el proyecto.

4.2. Comprensión de los datos

4.2.1. Recopilación de los datos iniciales

El archivo "heart_failure_clinical_records_dataset.csv" comprende un total de 299 instancias con 13 atributos cada una, donde cada instancia es la información clínica de un paciente y se representa de la siguiente manera:

A	B	C	D	E	F	G	H	I	J	K	L	M
age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
2	75	0	582	0	20	1	265000	1.9	130	1	0	4
3	55	0	7861	0	38	0	263358.03	1.1	136	1	0	6
4	65	0	146	0	20	0	162000	1.3	129	1	1	7
5	50	1	111	0	20	0	210000	1.9	137	1	0	7
6	65	1	160	1	20	0	327000	2.7	116	0	0	8
7	90	1	47	0	40	1	204000	2.1	132	1	1	8
8	75	1	246	0	15	0	127000	1.2	137	1	0	10
9	60	1	315	1	60	0	454000	1.1	131	1	1	10
10	65	0	157	0	65	0	263358.03	1.5	138	0	0	10
11	80	1	123	0	35	1	388000	9.4	133	1	1	10
12	75	1	81	0	38	1	368000	4	131	1	1	10
13	62	0	231	0	25	1	253000	0.9	140	1	1	10
14	45	1	981	0	30	0	136000	1.1	137	1	0	11
15	50	1	168	0	38	1	276000	1.1	137	1	0	11
16	49	1	80	0	30	1	427000	1	138	0	0	12
17	82	1	379	0	50	0	47000	1.3	136	1	0	13
18	87	1	149	0	38	0	262000	0.9	140	1	0	14
19	45	0	582	0	14	0	166000	0.8	127	1	0	14
20	70	1	125	0	25	1	237000	1	140	0	0	15
21	48	1	582	1	55	0	87000	1.9	121	0	0	15
22	65	1	52	0	25	1	276000	1.3	137	0	0	16

Figura 4.1: Visualización del archivo "heart_failure_clinical_records_dataset.csv"

4.2.2. Descripción de los datos

Este conjunto de datos contiene 13 atributos que pueden utilizarse para predecir la mortalidad por insuficiencia cardíaca.

Gran parte de las enfermedades cardiovasculares se pueden prevenir abordando los factores de riesgo conductuales como el consumo de tabaco, la obesidad, la escasa activación física y el consumo de alcohol utilizando estrategias para toda la población.

Con este conjunto de datos, es posible realizar un modelo de aprendizaje automático que sea de gran ayuda para las personas con enfermedades cardiovasculares o que se encuentran en alto riesgo cardiovascular cuya necesidad es una detección.

Tabla 4.3: Descripción de los campos.

Campo	Explicación	Tipo
Age - Edad	Edad del paciente	Años

Anaemia - Anemia	Disminución de los glóbulos rojos o hemoglobina	Booleano
creatinine_phosphokinase - creatinina fosfoquinasa	Nivel de la enzima CPK en la sangre	mcg/L
diabetes	Si el paciente tiene diabetes	Booleano
ejection_fraction - fracción de eyeccción	Porcentaje de sangre que sale del corazón en cada contracción	Porcentaje
high_blood_pressure - presión arterial alta	Si el paciente tiene hipertensión	Booleano
Platelets - plaquetas	Plaquetas en la sangre	kiloplaquetas/mL
serum_creatinine - creatinina sérica	Nivel de creatinina sérica en sangre	mg/dL
serum_sodium - sodio sérico	Nivel de sodio sérico en sangre	mEq/L
Sex - Sexo	Mujer u hombre	Binario
Smoking - Tabaquismo	Si el paciente fuma o no	Booleano
Time - Tiempo	Período de seguimiento	Días
DEATH_EVENT	Si el paciente falleció durante el período de seguimiento	Booleano

En la tabla 4.3, podemos visualizar que nuestro data set consta de 13 atributos.

Considero que son de suma importancia para el diagnóstico y/o prevención de enfermedades:

- Age
- Anemia
- High blood pressure
- Diabetes
- Smoking
- Death Event

Dadas sus principales características del conjunto de datos como Volumen, Variedad y Veracidad, se puede interpretar que los datos contenidos son suficientes para cumplir con el objetivo del proyecto.

4.2.3. Exploración de los datos

Los datos que se pueden graficar antes de comenzar a trabajar con los mismos son los siguientes:

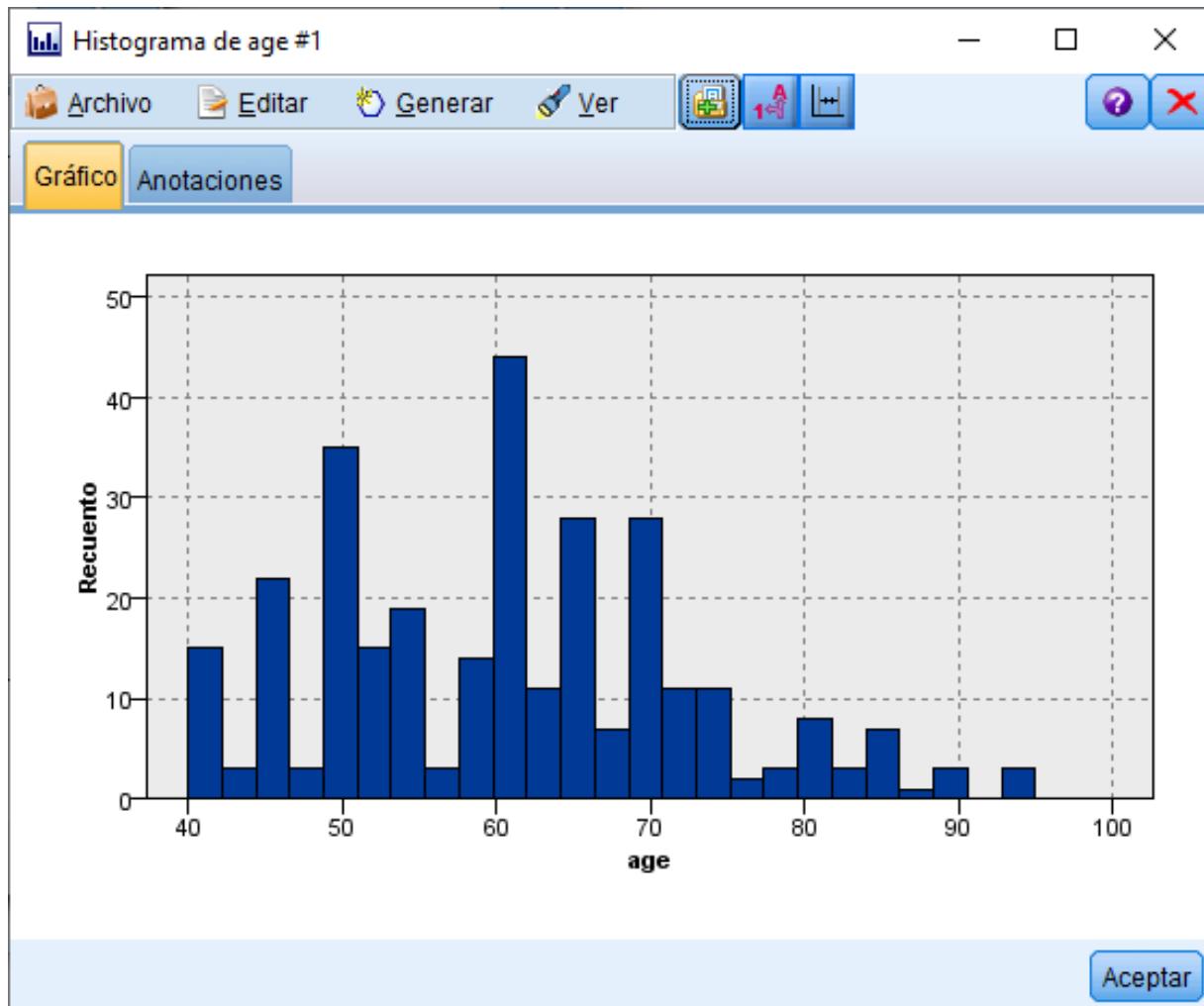


Figura 4.2: Histograma - Edades de los pacientes.

En la figura 4.2, vemos que la mayor cantidad de pacientes ronda entre los 50 a 70 años.

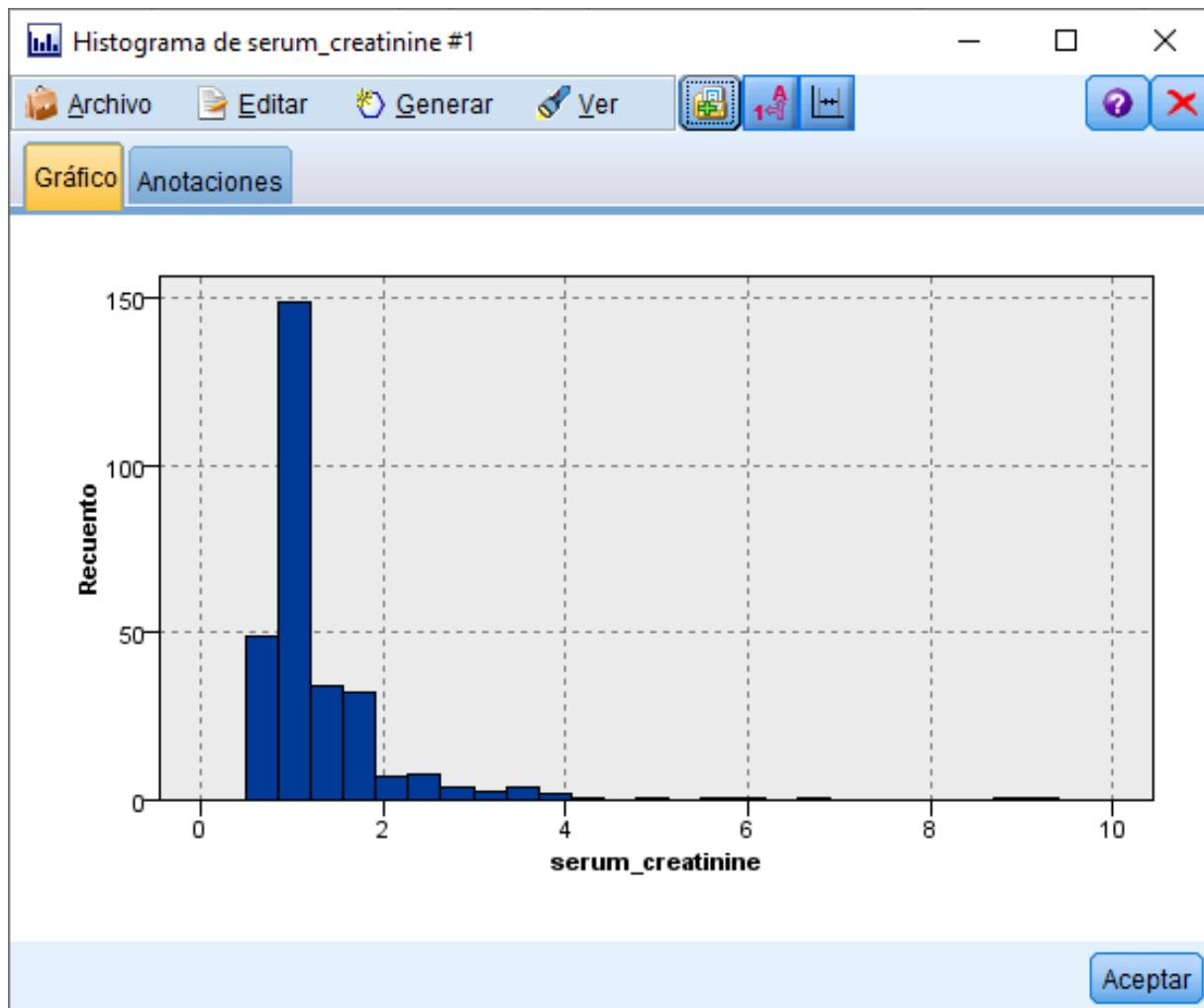


Figura 4.3: Histograma - serum_creatinine de los pacientes.

La figura 4.3 nos demuestra que el nivel creatinina sérica en la sangre de los pacientes se concentra en un rango menor a dos.

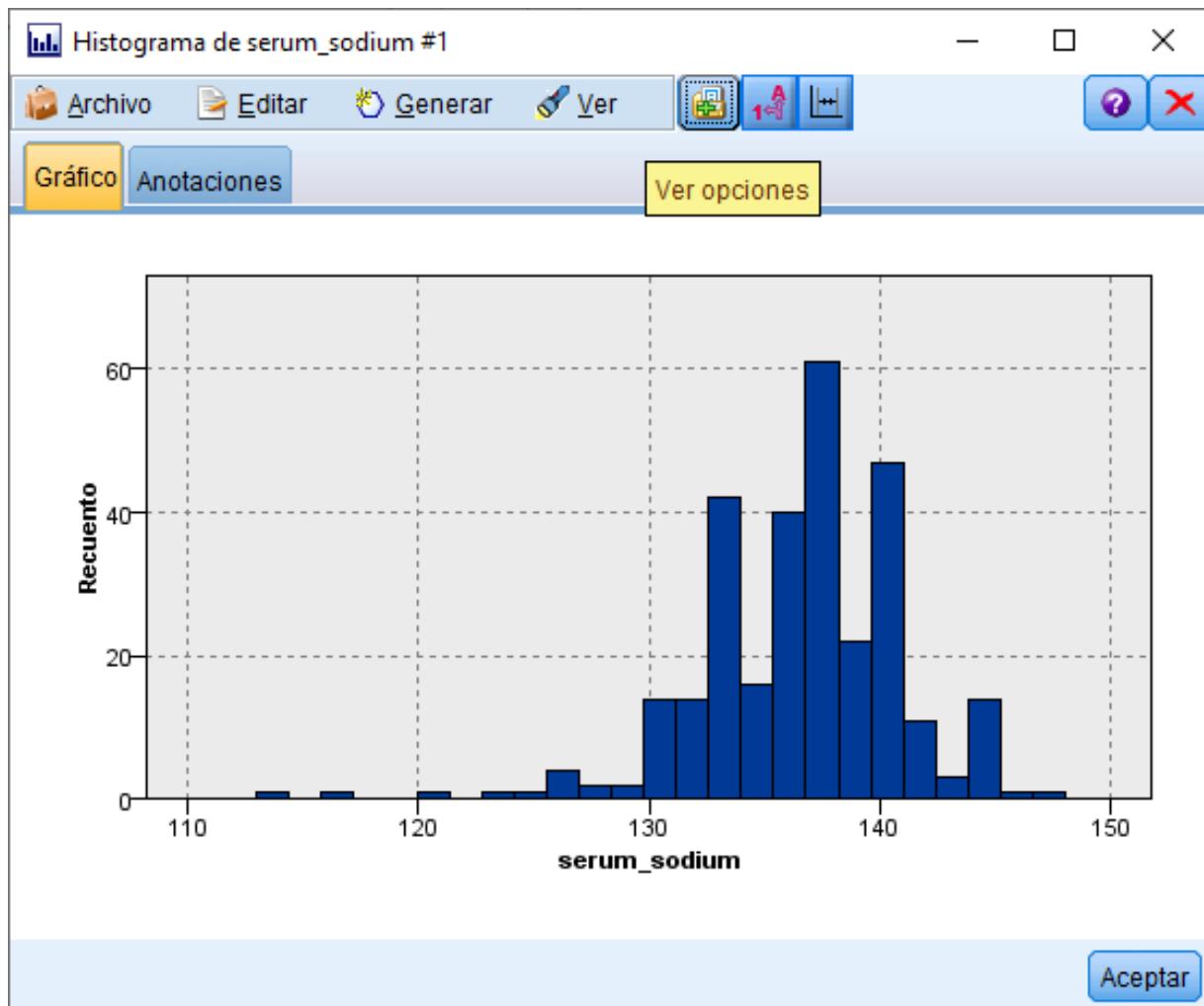


Figura 4.4: Histograma - serum_sodium de los pacientes.

La figura 4.4 muestra que el nivel sódico en la sangre de los pacientes se concentra en un rango de 130 a 150 mEq/L.

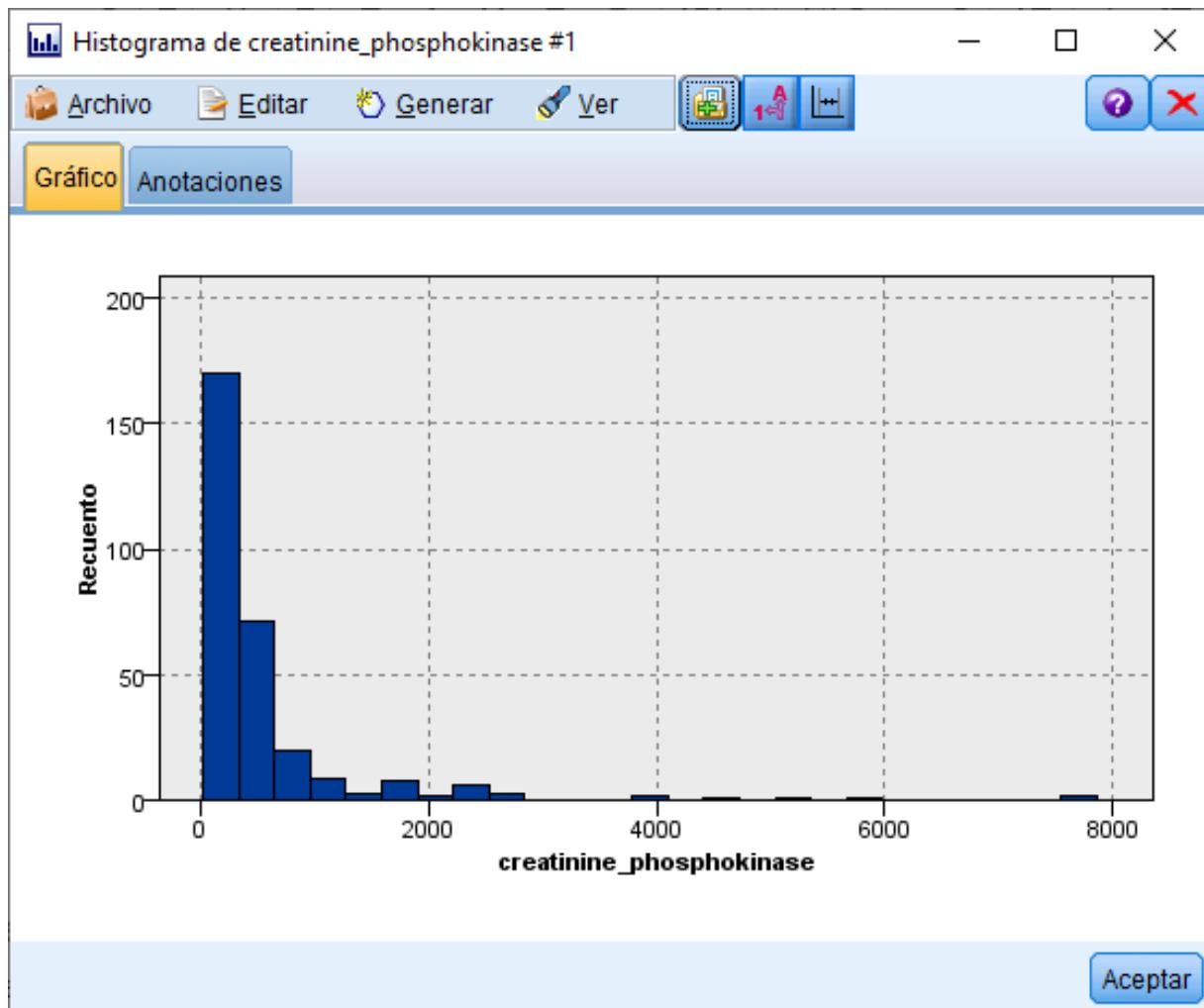


Figura 4.5: Histograma - creatinine_phosphokinase de los pacientes.

En la figura 4.5 vemos que el nivel de la enzima CPK en la sangre de los pacientes se concentra en un rango inferior a 100.

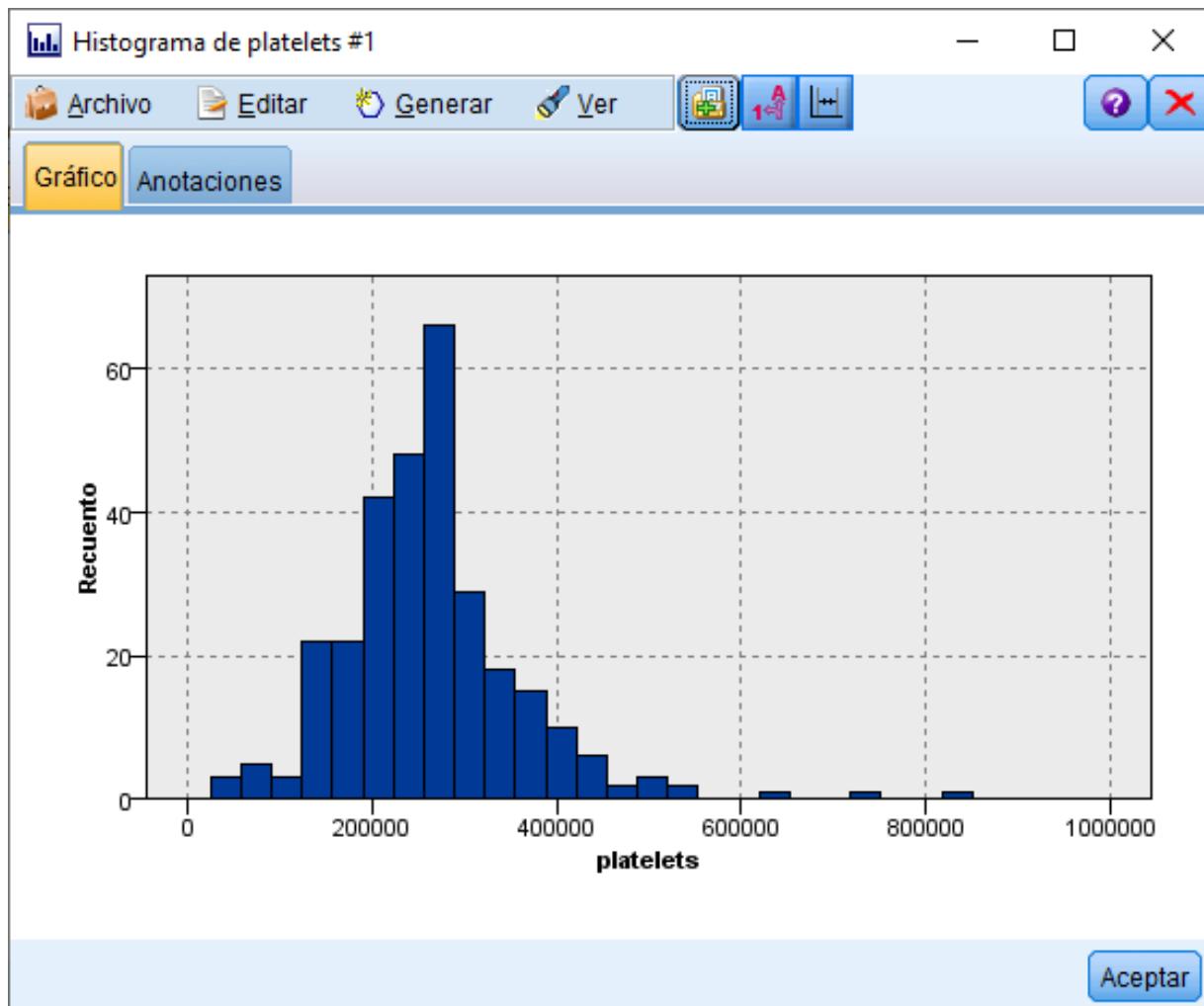


Figura 4.6: Histograma - Platelets de los pacientes.

Por otra parte, la figura 4.6 demuestra que las plaquetas en la sangre de los pacientes se concentra en un rango de 100,000 a 400,000.

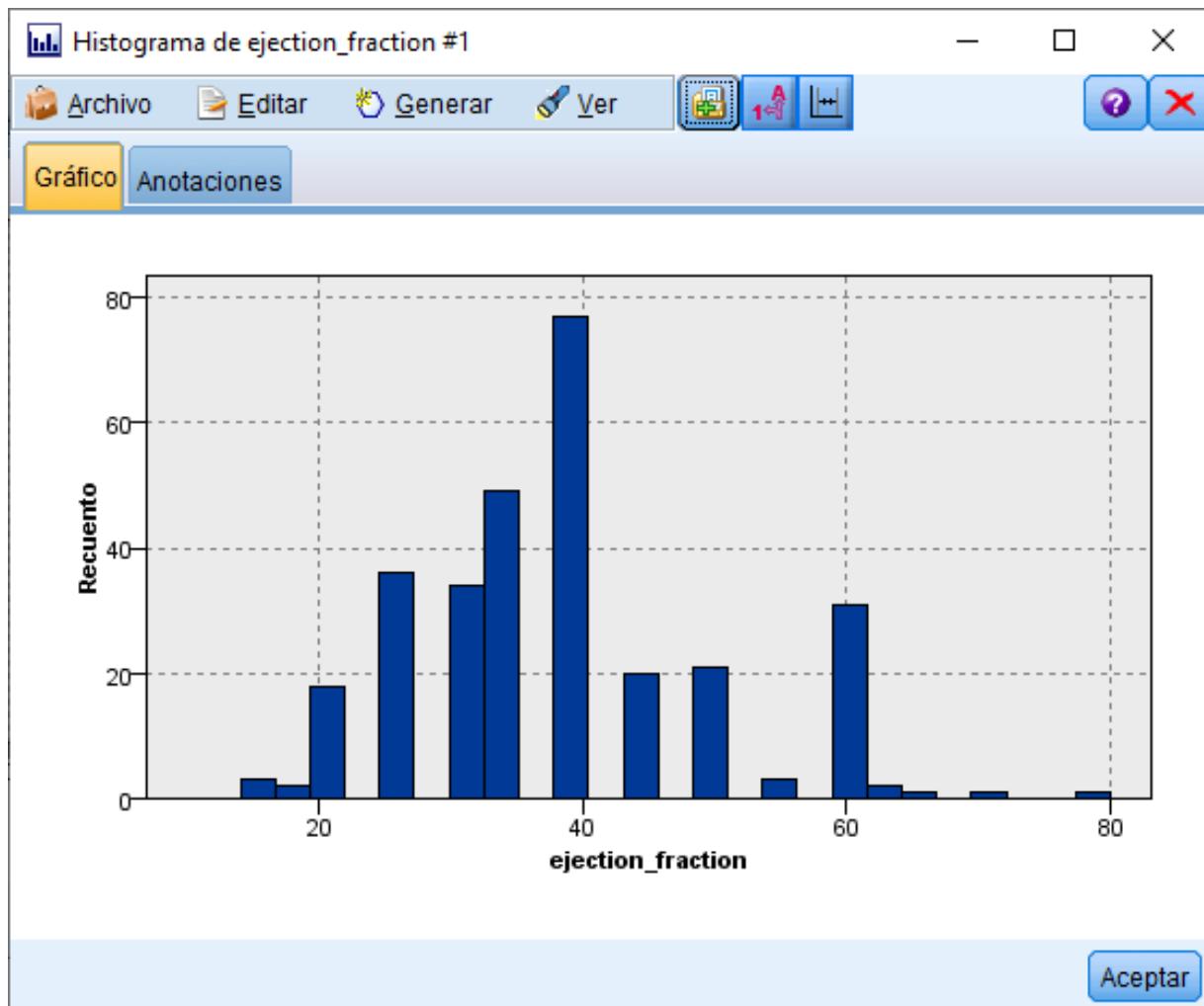


Figura 4.7: Histograma - Ejection_fraction de los pacientes.

No obstante, la figura 4.7 nos muestra que el porcentaje de sangre que sale del corazón en cada contracción de los pacientes es variable pero su punto máximo es 40 por ciento.

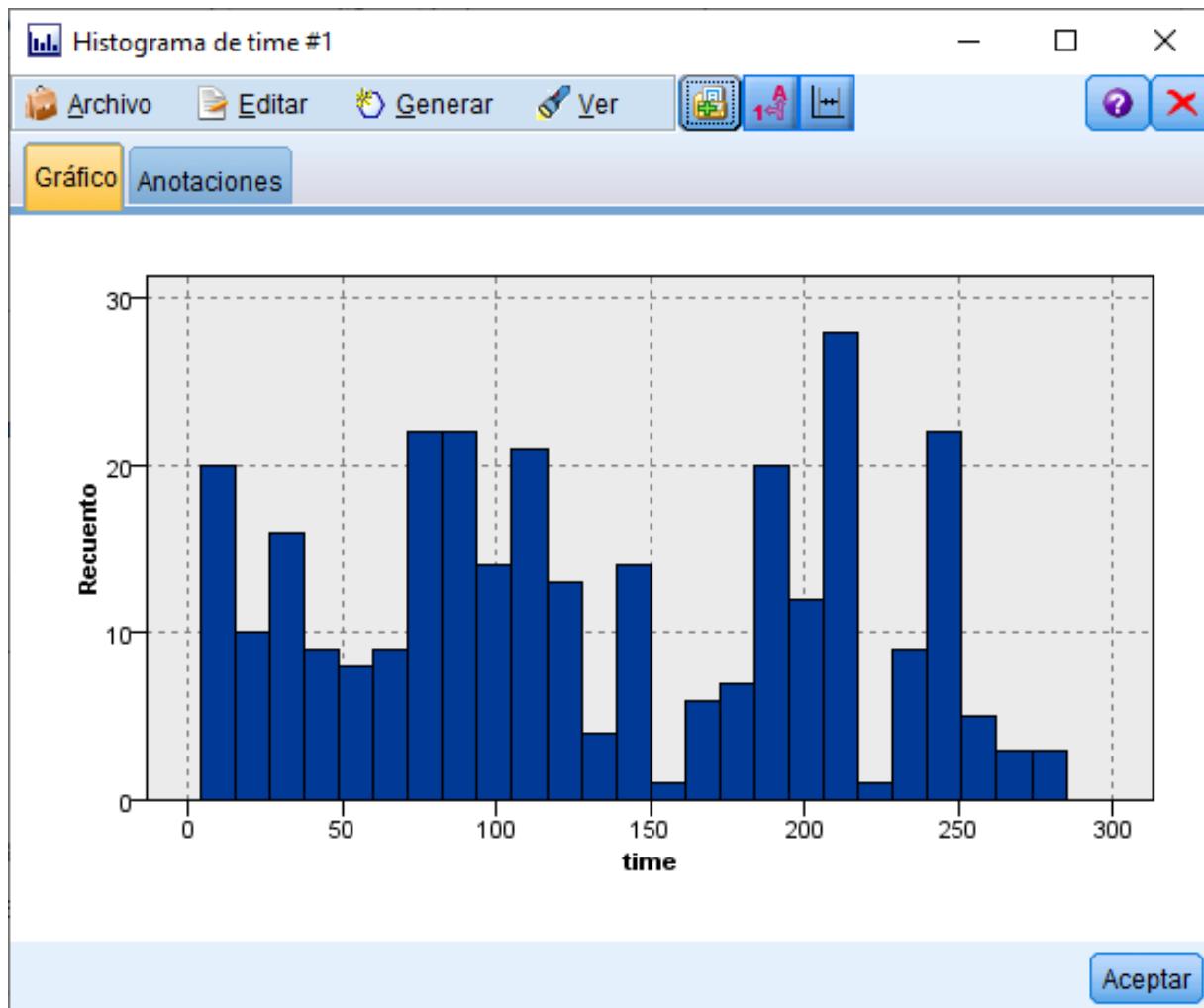


Figura 4.8: Histograma - Time de los pacientes.

Finalmente, la figura 4.8 nos muestra que el tiempo de seguimiento de los pacientes es el más variable de los datos antes presentados, pero su punto máximo ronda los 210 días.

En las figuras antes presentadas, podemos ver la distribución de los datos mediante un histograma, el cual nos permite comprender cuales son los rangos que tienen mayor y menor número datos y con base a esto es posible comenzar con la creación de conjecturas.

4.2.4. Verificación de la calidad de los datos

Dentro del archivo se encuentra:

- **Datos perdidos o vacíos:** No existen.
- **Errores en los datos:** Aparentemente no existen.
- **Errores de medición:** Aparentemente no existen.

- **Errores de codificación:** Aparentemente no existen.
- **Atributos redundantes o de escasa utilidad:** Aparentemente no existen.

4.3. Preparación de los datos

4.3.1. Selección de datos

Dentro de este data set se considera que el atributo "Time" no es un campo relevante para la aplicación de un modelo. Como consecuencia, se trabajará a partir de los 12 atributos restantes y, como veremos más adelante, se considerará la derivación de nuevos atributos, lo que podría ocasionar la exclusión de algunos atributos existentes al dejar de ser relevantes.

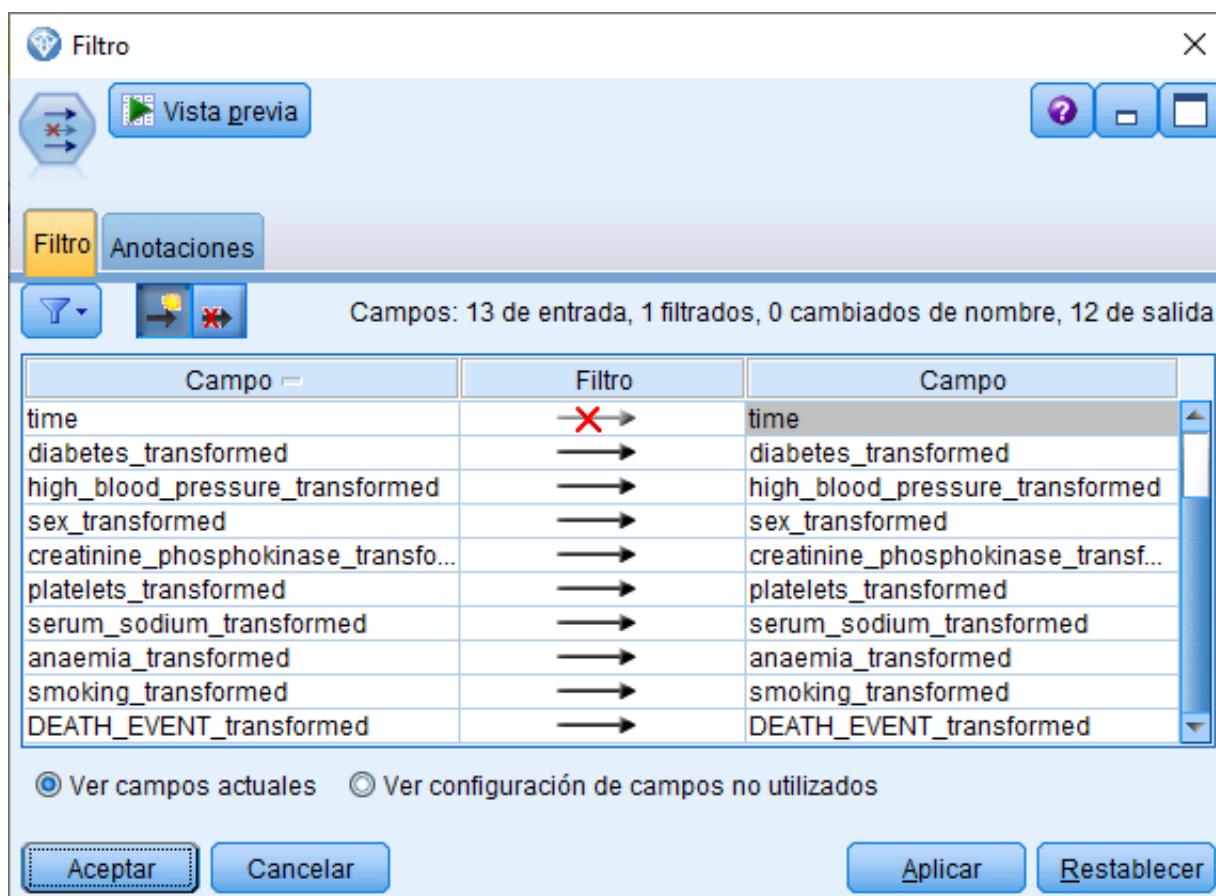


Figura 4.9: Filtro de los datos en la herramienta IBM SPSS Modeler.

4.3.2. Limpieza de datos

Como se mencionó en la fase "Comprensión de los datos", no era necesario corregir problemas relacionados con datos perdidos o vacíos; sin embargo, se optó por borrar un

par de datos del big data para implementar técnicas de preparación de los datos.

Por ende:

- **Datos perdidos o vacíos:** Sí existen.
- **Errores en los datos:** Aparentemente no existen.
- **Errores de medición:** Aparentemente no existen.
- **Errores de codificación:** Aparentemente no existen.
- **Atributos redundantes o de escasa utilidad:** Aparentemente no existen.

Dada la situación, por medio de la herramienta *IBM SPSS Modeler*, se limpiaron los datos en nueve campos del data set, los cuales son:

- Anaemia
- Diabetes
- High Blood Pressure
- Sex
- Smoking
- DEATH EVENT
- Creatinine phosphokinase
- Plateles
- Serum sodium

	anaemia_transformed	DEATH_EVENT_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transformed	smoking_transformed	age_transformed	creatinine_phosphokinase_transformed	ejection_frac
1	0	1	0	1	1	0	75.000	582.000	
2	0	1	0	0	1	0	55.000	1930.379	
3	0	1	0	0	1	1	65.000	146.000	
4	1	1	0	0	1	0	50.000	111.000	
5	1	1	1	0	0	0	65.000	160.000	
6	1	1	0	1	1	1	90.000	47.000	
7	1	1	0	0	1	0	75.000	246.000	
8	1	1	1	0	1	1	60.000	315.000	
9	0	1	0	0	0	0	65.000	157.000	
10	1	1	0	1	1	1	80.000	123.000	
11	1	1	0	1	1	1	75.000	81.000	
12	0	1	0	1	1	1	62.000	231.000	
13	1	1	0	0	1	0	45.000	981.000	
14	1	1	0	1	1	0	50.000	168.000	
15	1	0	0	1	0	0	49.000	80.000	
16	1	1	0	0	1	0	82.000	379.000	
17	1	1	0	0	1	0	87.000	149.000	
18	0	1	0	0	1	0	45.000	582.000	
19	1	1	0	1	0	0	70.000	125.000	
20	1	1	1	0	0	0	48.000	582.000	

Figura 4.10: Resultado de la limpieza de los datos con la herramienta IBM SPSS Modeler

4.3.3. Construcción de nuevos datos

Debido a la complejidad para comprender los datos, decidí crear nuevos campos que me permitieran asumir la información de cada registro, los campos que usé son:

- Serum creatinine
- Creatinine Phosphokinase
- Platelets
- Serum sodium

Los campos derivados son:

- DiferenciaDeMediaSC
- DiferenciaDeMediaCP
- DiferenciaDeMediaPlatelets
- DiferenciaDeMediaSS

Cada campo creado contiene el porcentaje de la diferencia de la media.

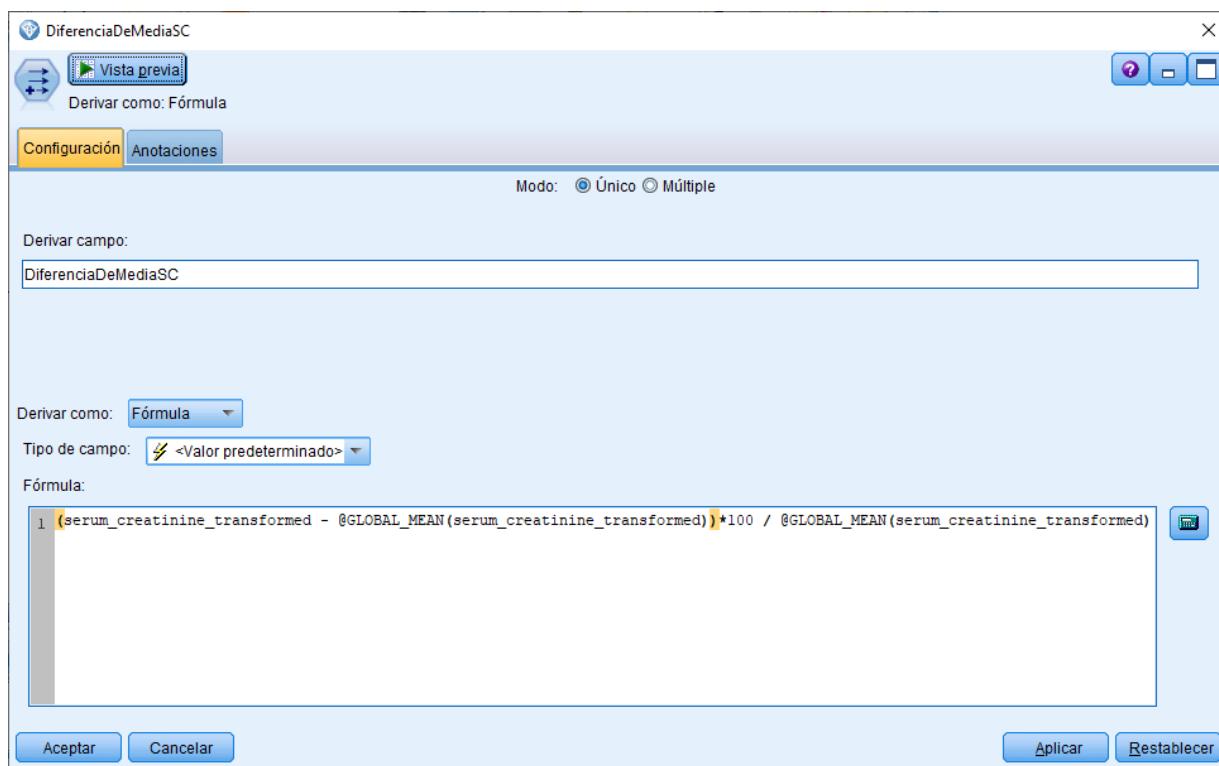


Figura 4.11: Fórmula para衍生 el campo DiferenciaDeMediaSC.

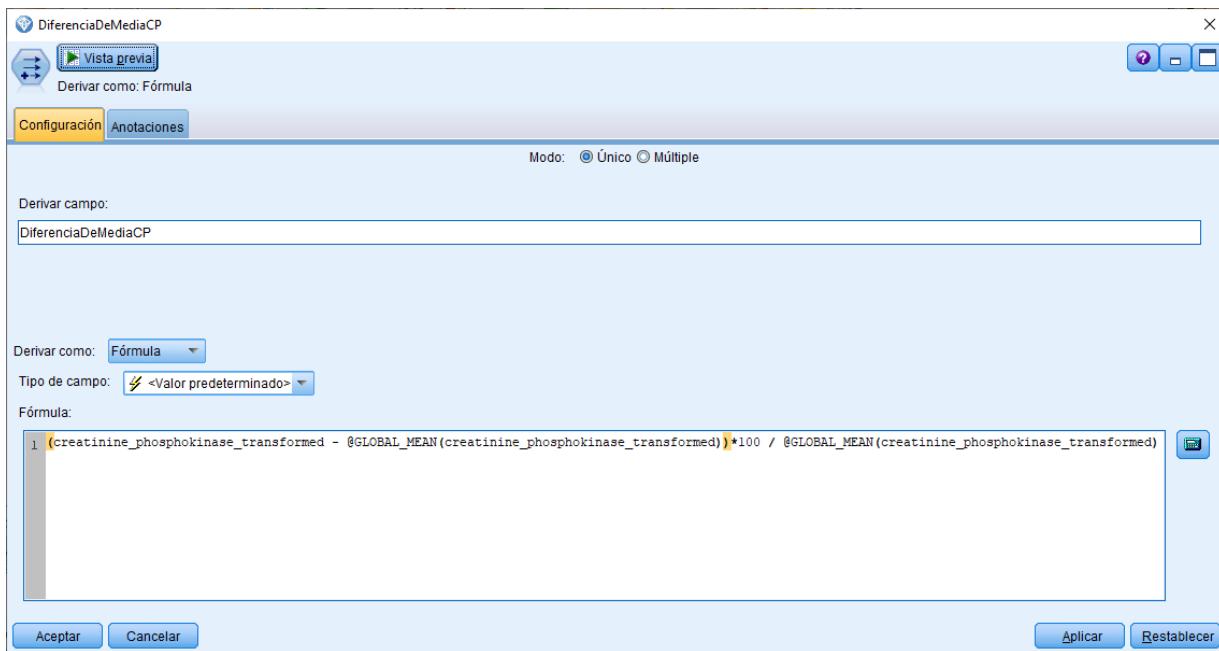


Figura 4.12: Fórmula para derivar el campo DiferenciaDeMediaCP.

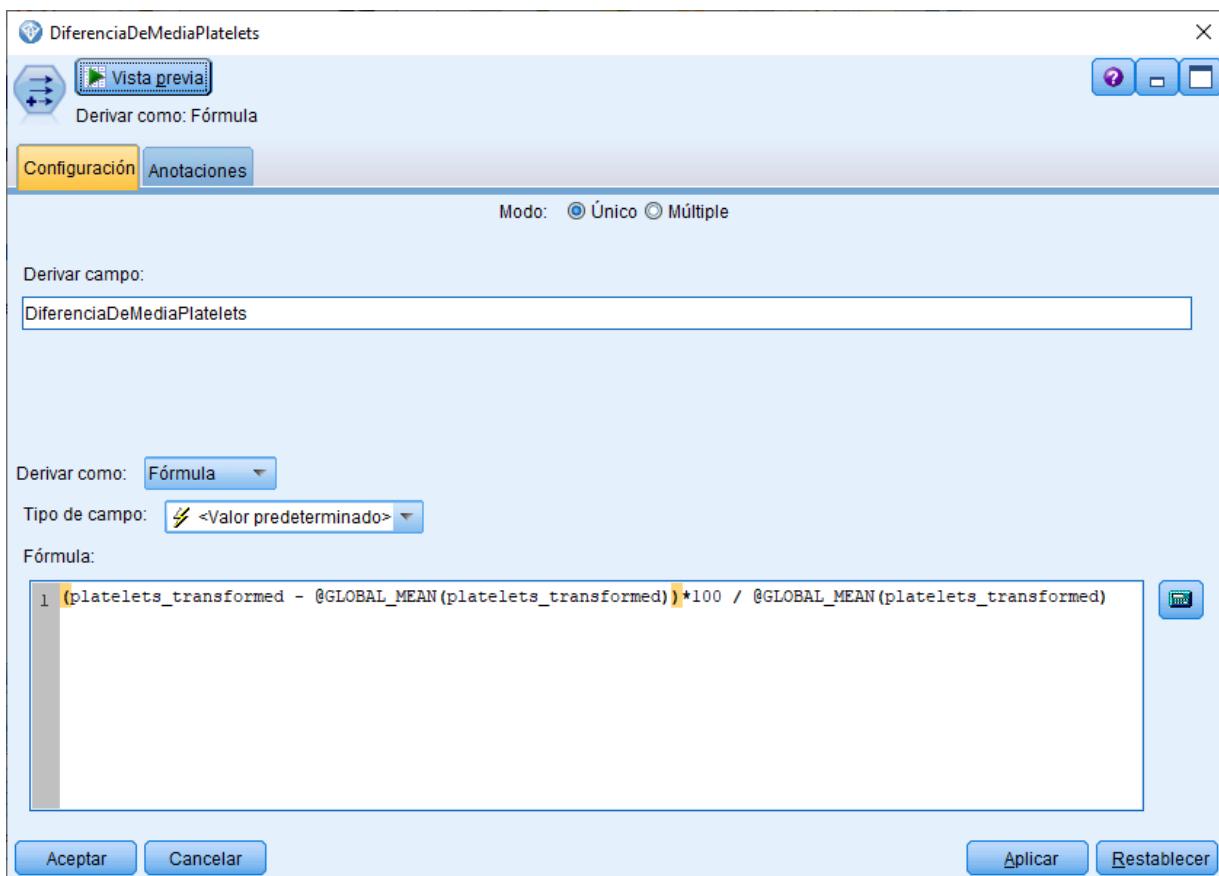


Figura 4.13: Fórmula para derivar el campo DiferenciaDeMediaPlatelets.

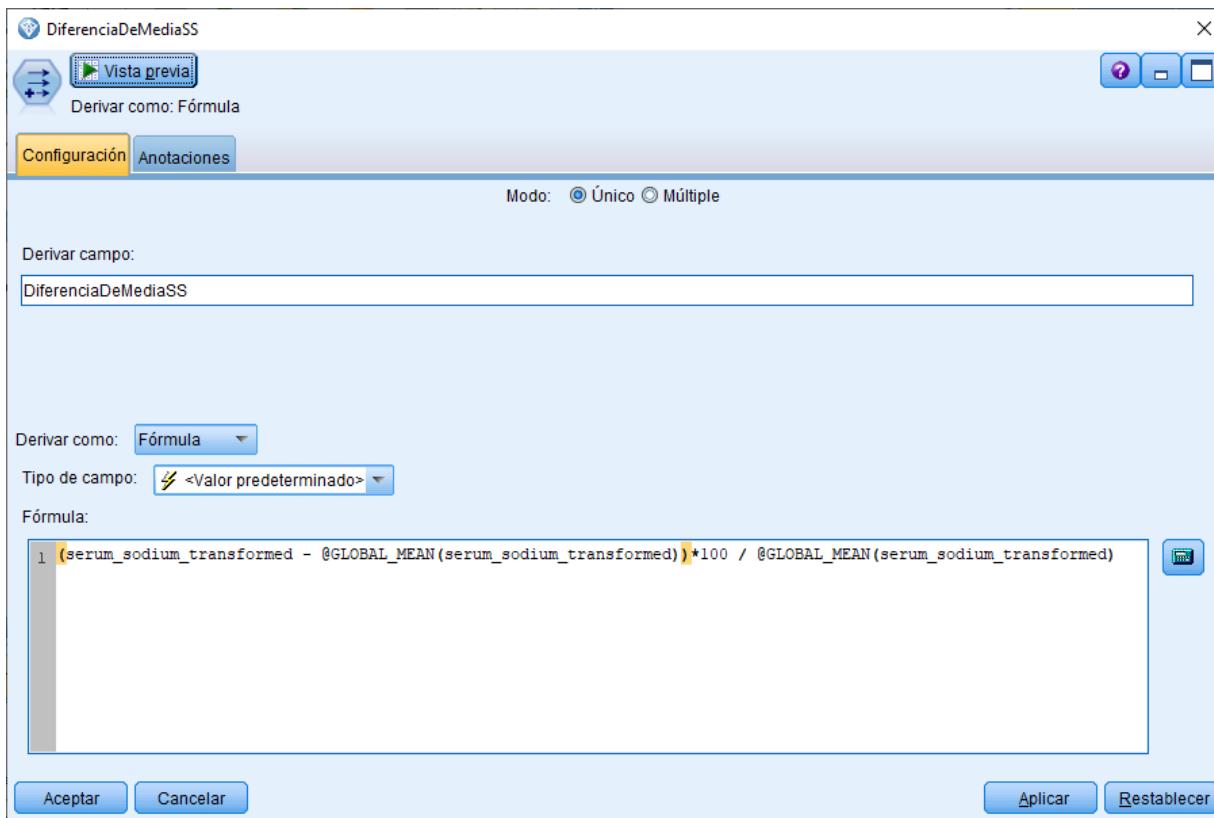


Figura 4.14: Fórmula para derivar el campo DiferenciaDeMediaSS.

4.3.4. Integración de datos

Para este data set, no fue posible realizar integración de los datos, no se cuenta con otras fuentes de datos a integrar al conjunto inicial de datos. Por lo tanto, el conjunto inicial de datos será el conjunto de datos a analizar con los modelos que se propongan en la fase "Modelado".

4.3.5. Formato de datos

Tabla 4.5: Formato de los campos.

Campo	Explicación	Tipo	Rango
Age - Edad	Edad del paciente	Años	40 - 90
Anaemia - Anemia	Disminución de los glóbulos rojos o hemoglobina	Booleano - Ordinal	0 - 1
creatinine phosphokinase - creatinina fosfoquinasa	Nivel de la enzima CPK en la sangre	mcg/L - Continuo	23 - 1930

diabetes	Si el paciente tiene diabetes	Booleano - Ordinal	0 - 1
ejection_fraction - fracción de eyección	Porcentaje de sangre que sale del corazón en cada contracción	Porcentaje - Continuo	14 - 70
high_blood_pressure - presión arterial alta	Si el paciente tiene hipertensión	Booleano - Ordinal	0 - 1
Platelets - plaquetas	Plaquetas en la sangre	kiloplaquetas/mL - Continuo	47,000 - 475,585
serum_creatinine - creatinina sérica	Nivel de creatinina sérica en sangre	mg/dL - Continuo	0.5 - 2.7
serum_sodium - sodio sérico	Nivel de sodio sérico en sangre	mEq/L - Continuo	126 - 147
Sex - Sexo	Mujer u hombre	Binario - Ordinal	0 - 1
Smoking - Tabaquismo	Si el paciente fuma o no	Booleano - Ordinal	0 - 1
DEATH_EVENT	Si el paciente falleció durante el período de seguimiento	Booleano - Ordinal	0 - 1
DiferenciaDeMediaSC	Porcentaje de la diferencia de la media.	Porcentaje - Continuo	-60 - 111
DiferenciaDeMediaCP	Porcentaje de la diferencia de la media.	Porcentaje - Continuo	-95 - 300
DiferenciaDeMediaPlatelets	Porcentaje de la diferencia de la media.	Porcentaje - Continuo	-81 - 82
DiferenciaDeMediaSS	Porcentaje de la diferencia de la media.	Porcentaje - Continuo	-7 - 7

En la tabla 4.5, vemos que los datos se encuentran en el formato requerido para la aplicación de los modelos predictivos.

Respecto al rol de cada campo, se mantiene que el único campo destino es DEATH_EVENT, esto se demuestra en la siguiente figura.

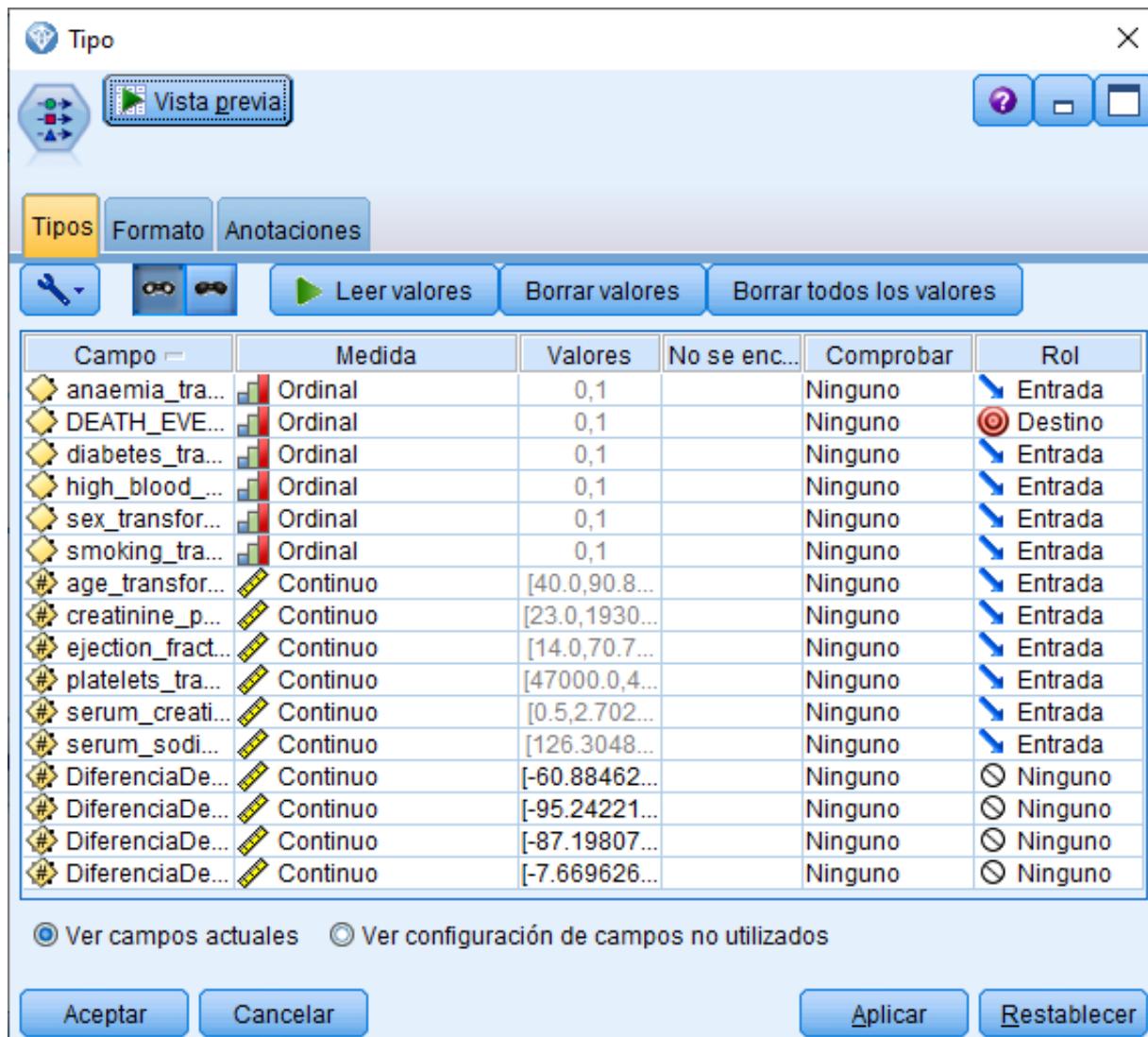


Figura 4.15: Resultado de la limpieza de los datos con la herramienta IBM SPSS Modeler

4.3.6. Nueva exploración de los datos

Una vez que hemos preparado los datos, es posible volver a explorarlos debido a que poseen una nueva presentación, esto nos permitirá comprenderlos mejor y será posible formular nuevas hipótesis de los mismos en caso de ser necesario.

En las siguientes figuras, podremos apreciar gráficos que nos permitirán comprender de manera visual la relación que existe entre los datos, al igual que su colección, agrupación, densidad, entre otros.

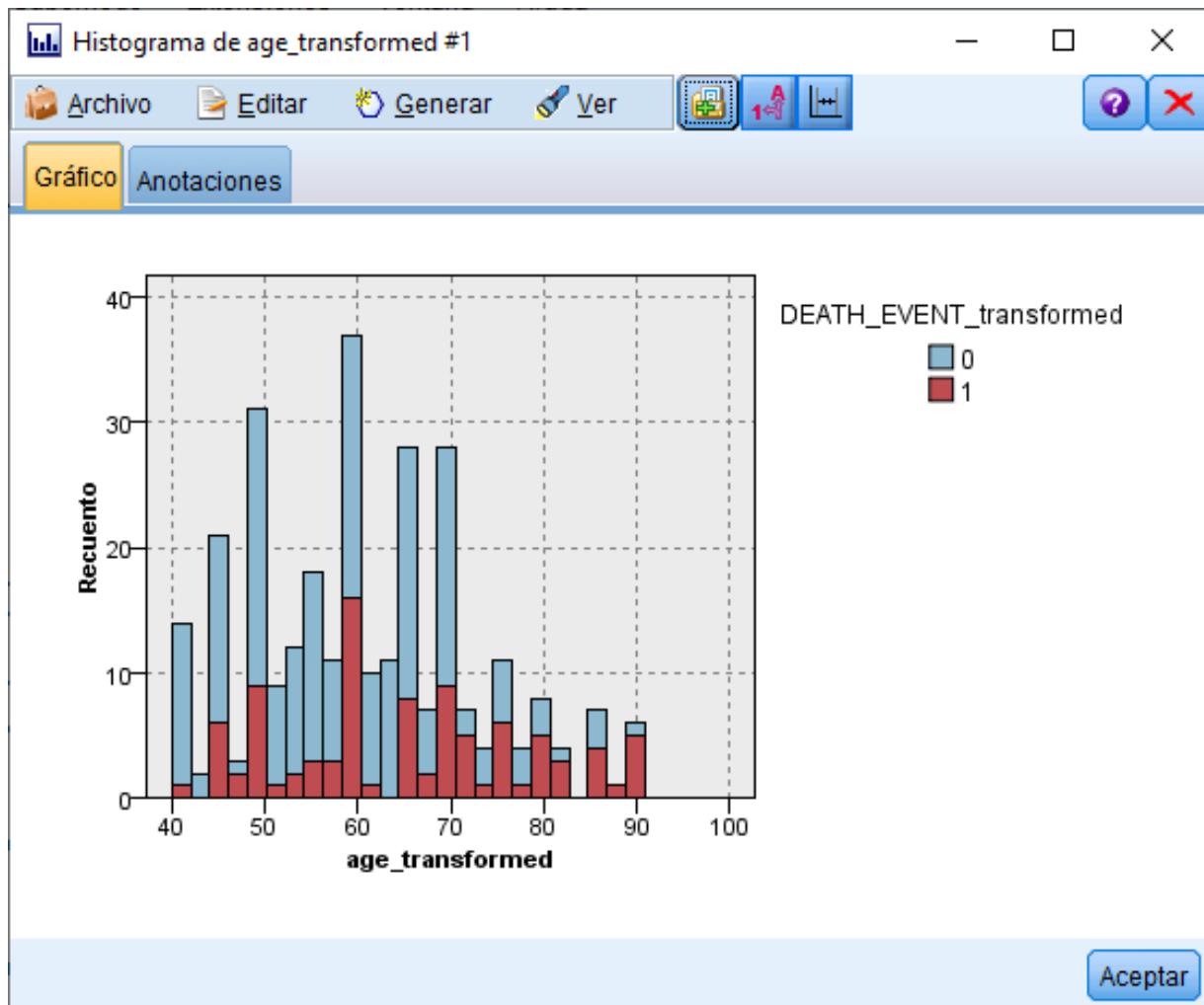


Figura 4.16: Histograma - Age

En la figura 4.16, vemos que mientras más edad se tiene, más probabilidad existe de que el campo *DEATH_EVENT* sea 1.

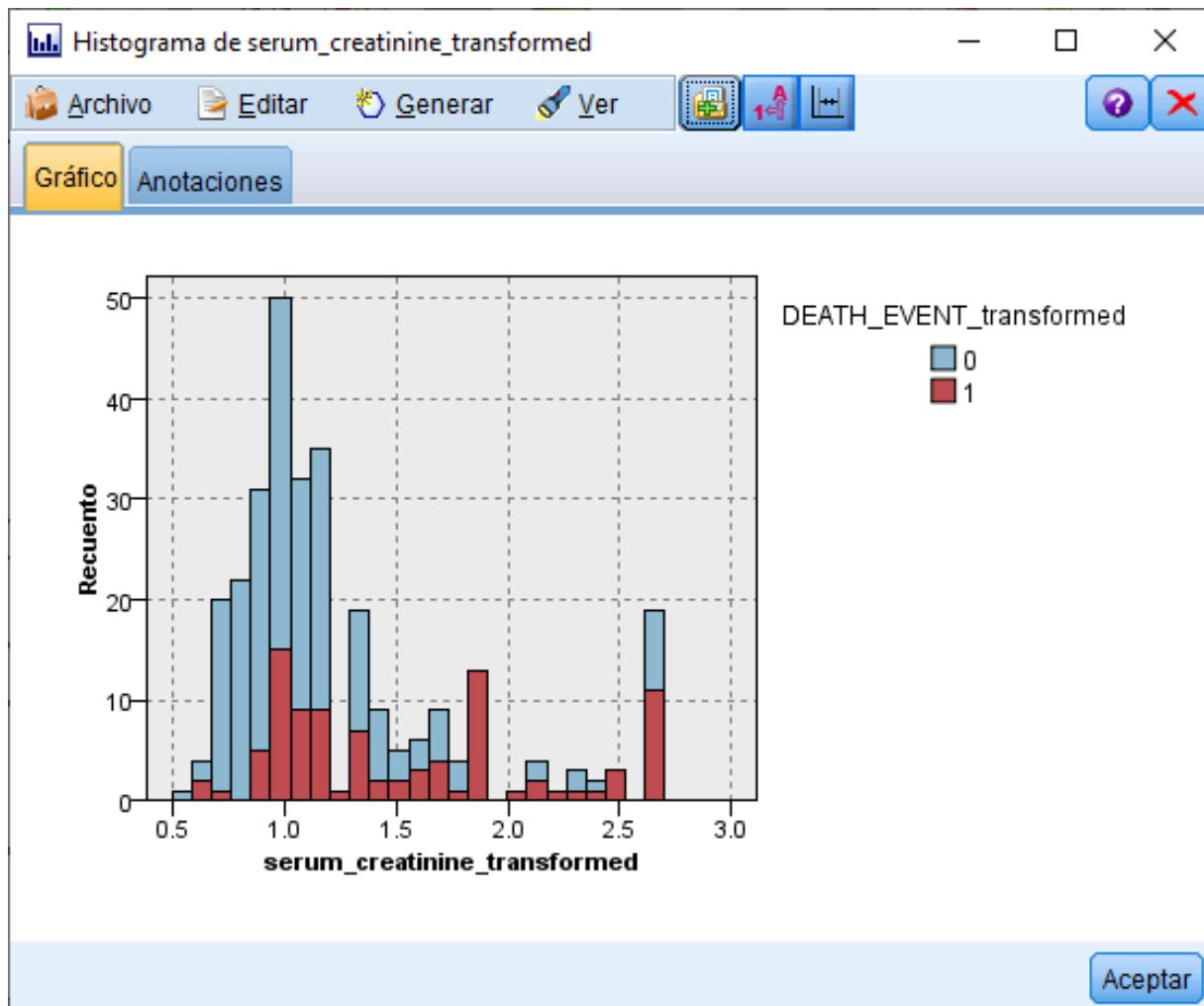


Figura 4.17: Histograma - serum_creatinine

En la figura 4.17, vemos que la mayoría de los pacientes lo mantienen en un rango de 0.5 a 2, donde gran parte de ellos obtienen 0 en el campo DEATH_EVENT.

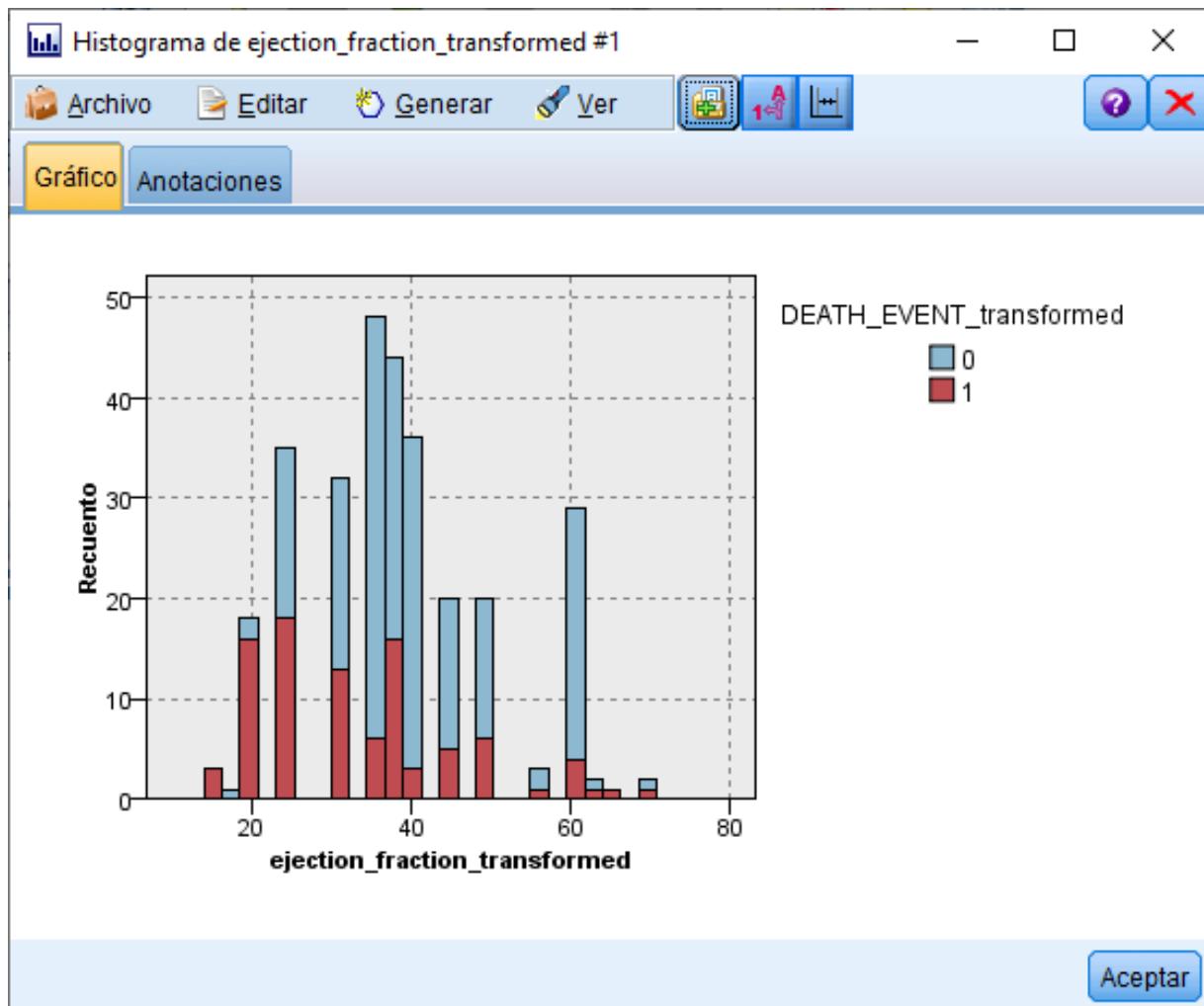


Figura 4.18: Histograma - ejection_fraction

La figura 4.18 nos muestra que los pacientes tienen más probabilidad de tener 1 en el campo DEATH_EVENT cuando se encuentran en el rango de 20 a 40.

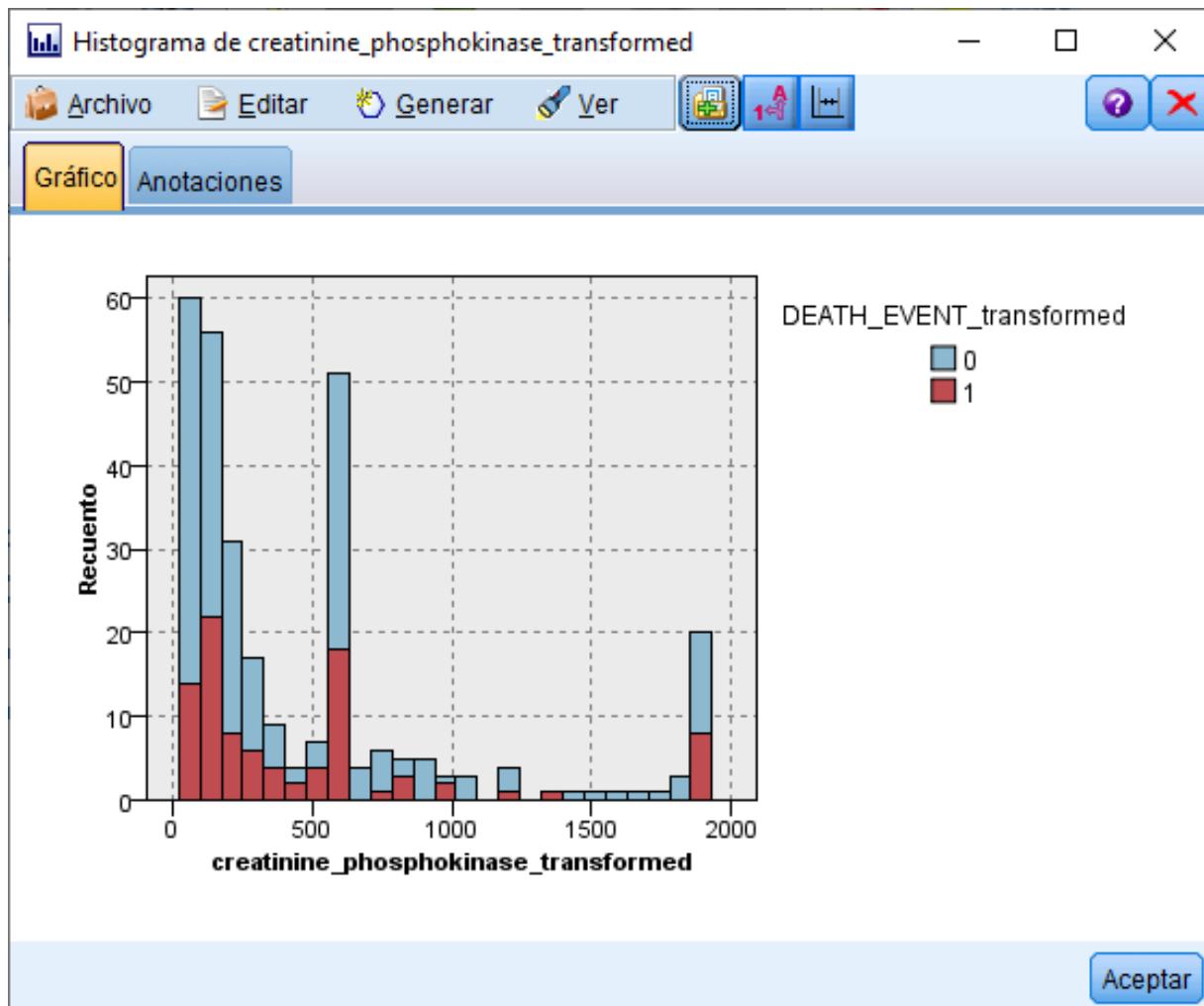


Figura 4.19: Histograma - creatinine_phosphokinase

La figura 4.19 nos muestra que los pacientes tienen más probabilidad de tener 1 en el campo DEATH_EVENT cuando se encuentran en el rango de 0 a 600.

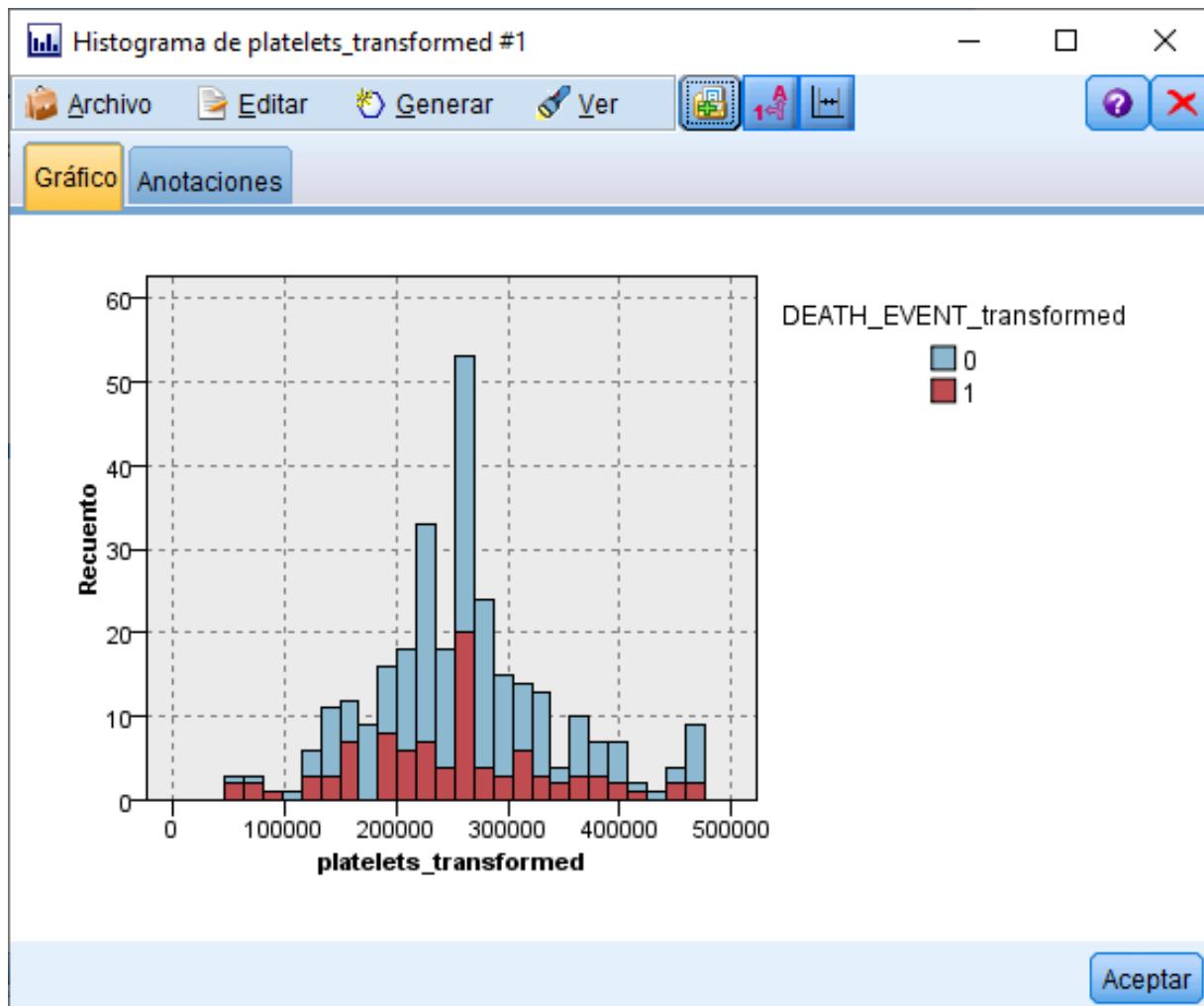


Figura 4.20: Histograma - platelets_transformed

En la figura 4.20, observamos que los pacientes tienen más probabilidad de tener 1 en el campo DEATH_EVENT cuando se encuentran en el rango de 260,000 a 280,000.

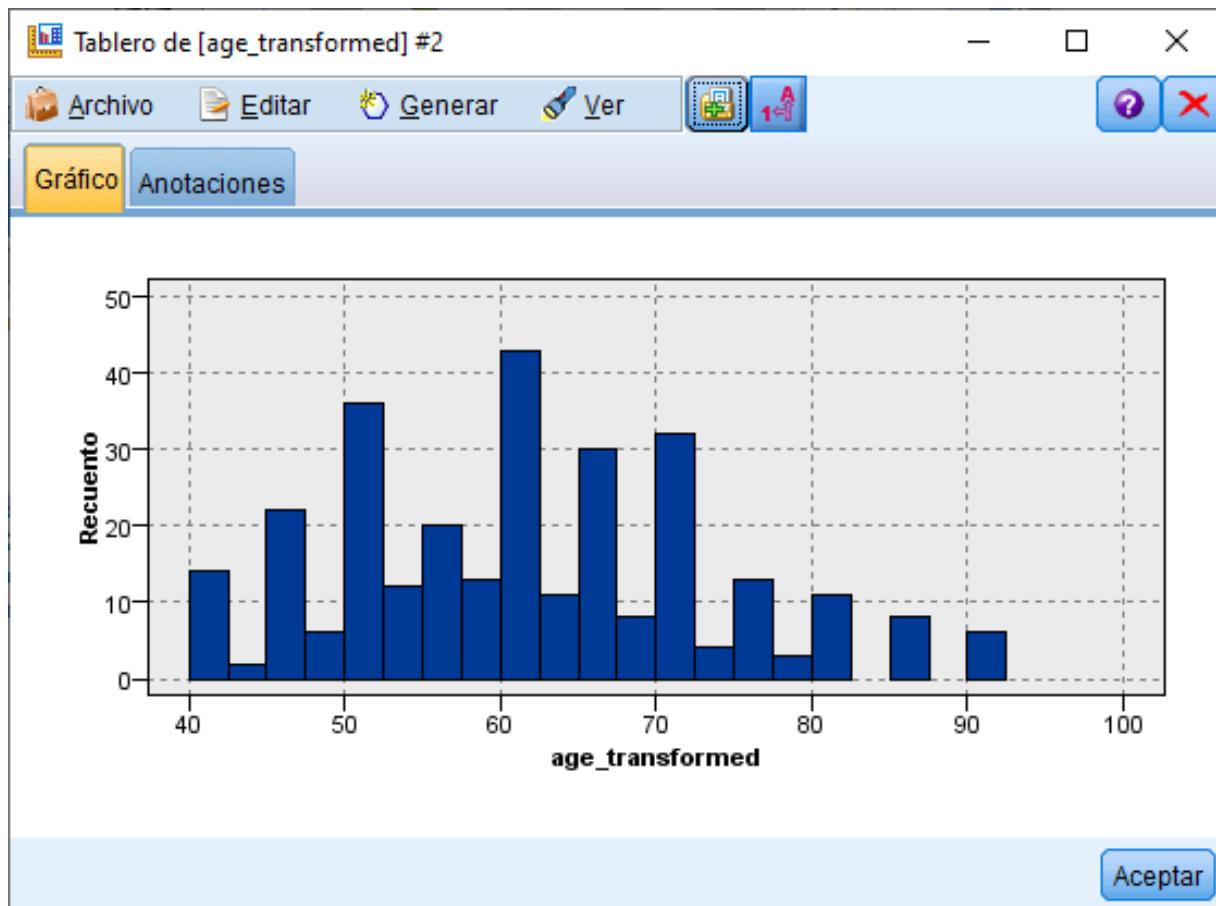


Figura 4.21: Recuento de la edad de los pacientes.

Dentro de la figura 4.21, nos percatamos que gran parte de los pacientes tienen entre 50 a 70 años.

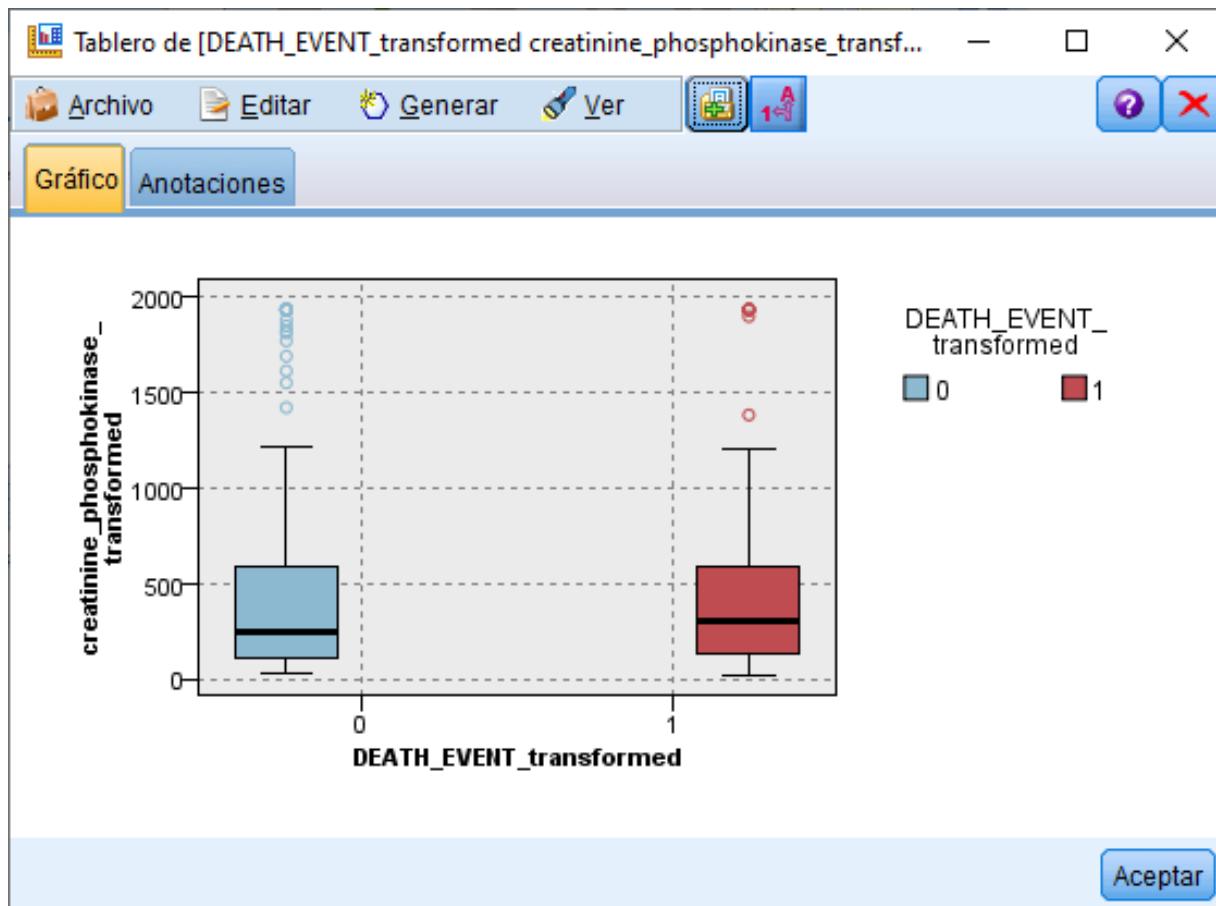


Figura 4.22: Diagrama de caja - creatinine_phosphokinase.

Dentro de la figura 4.22 podemos observar que gran parte de los pacientes mantienen valores en el rango de 0 a 600.

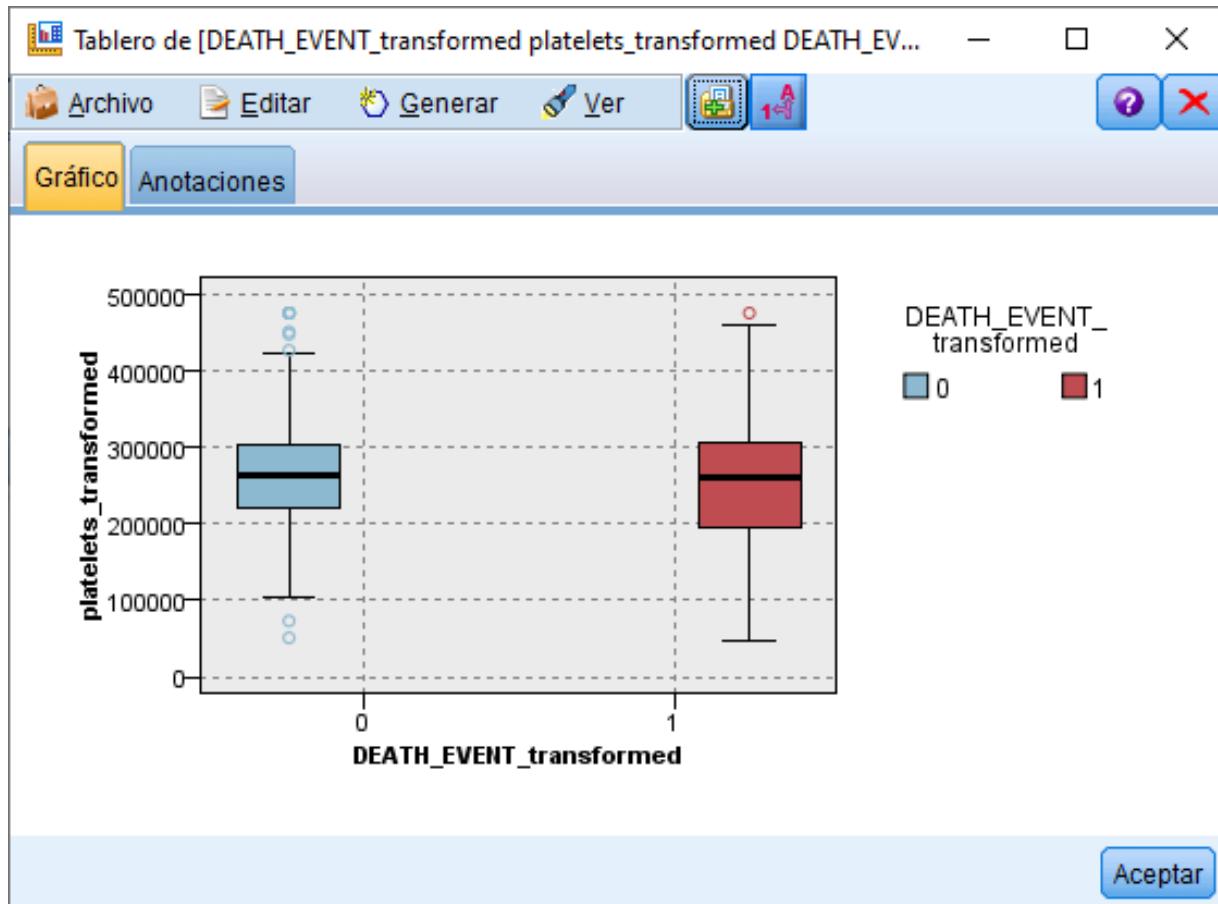


Figura 4.23: Diagrama de caja - platelets_transformed

Nuevamente, en la figura 4.23 tenemos el mismo análisis que se realizó para la figura 4.20 pero con una presentación diferente.

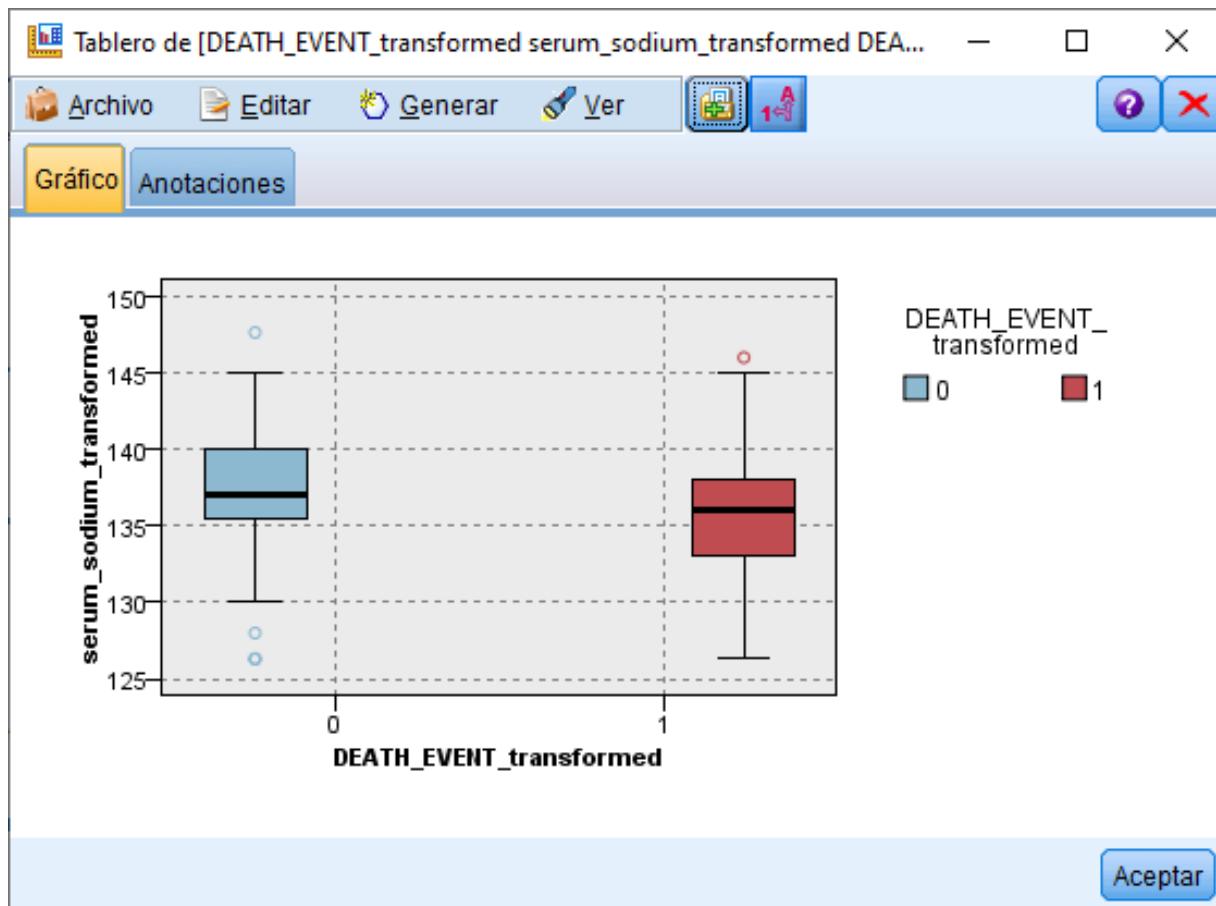


Figura 4.24: Diagrama de caja - serum_sodium

Por otra parte, en la figura 4.24 vemos que más del 50% de los pacientes se encuentran en un rango de 130 a 140 mEq/L.

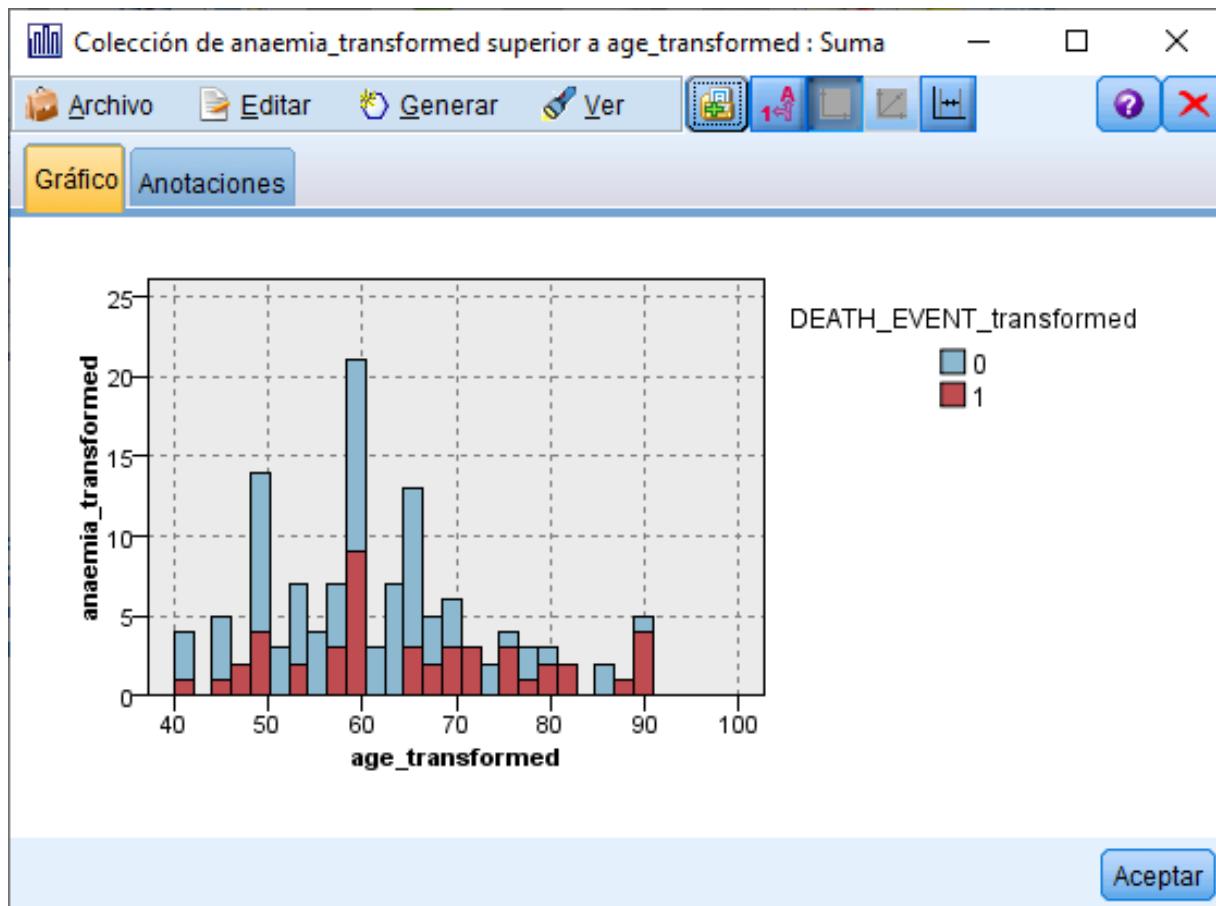


Figura 4.25: Colección de anemia respecto a la edad del paciente.

Por otra parte, en la figura 4.25 se demuestra que la mayor cantidad de pacientes con anemia rondan los 60 años, mientras que de los 70 años a los 85 años se tiene mayor probabilidad de tener anemia.

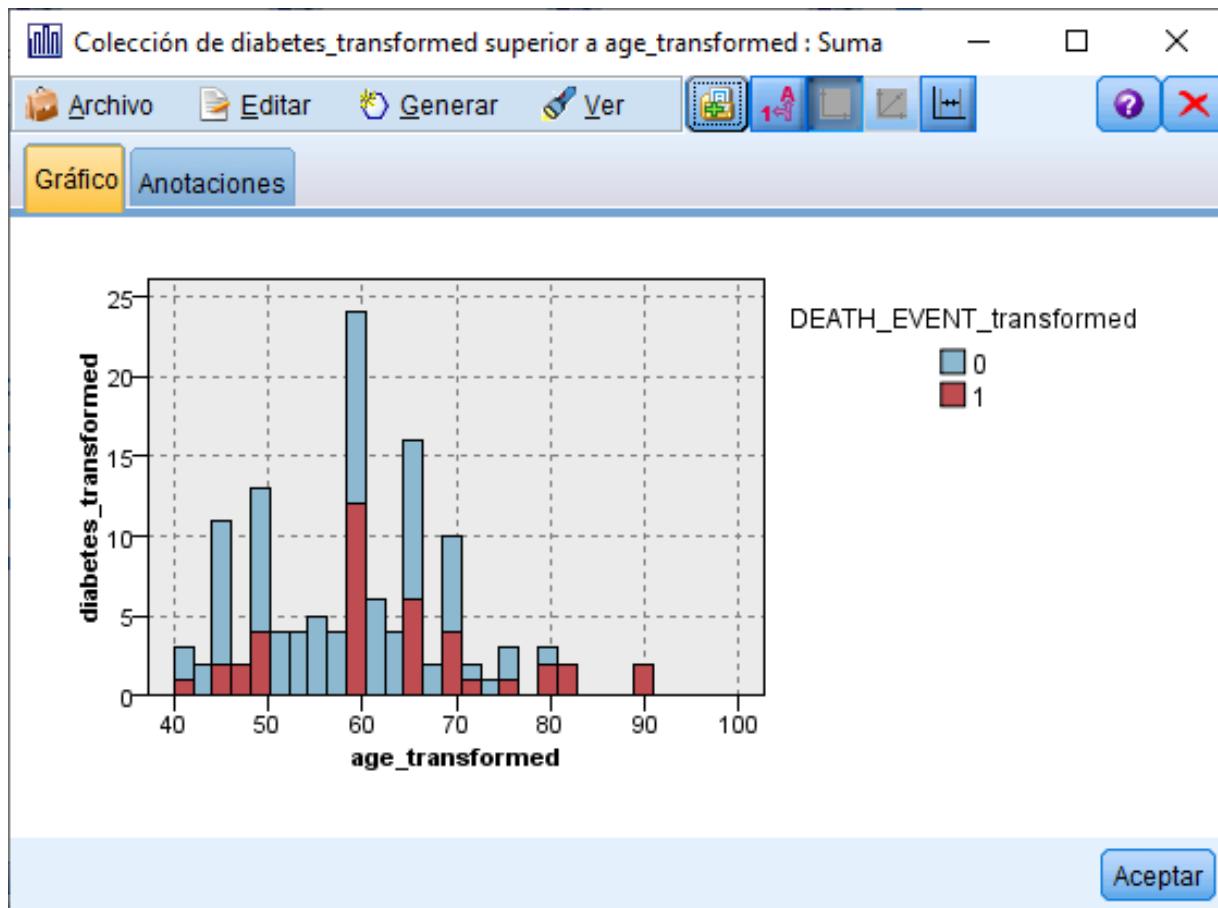


Figura 4.26: Colección de diabetes respecto a la edad del paciente.

No obstante, en la figura 4.26 vemos que la mayor cantidad de pacientes con diabetes ronda los 60 años.

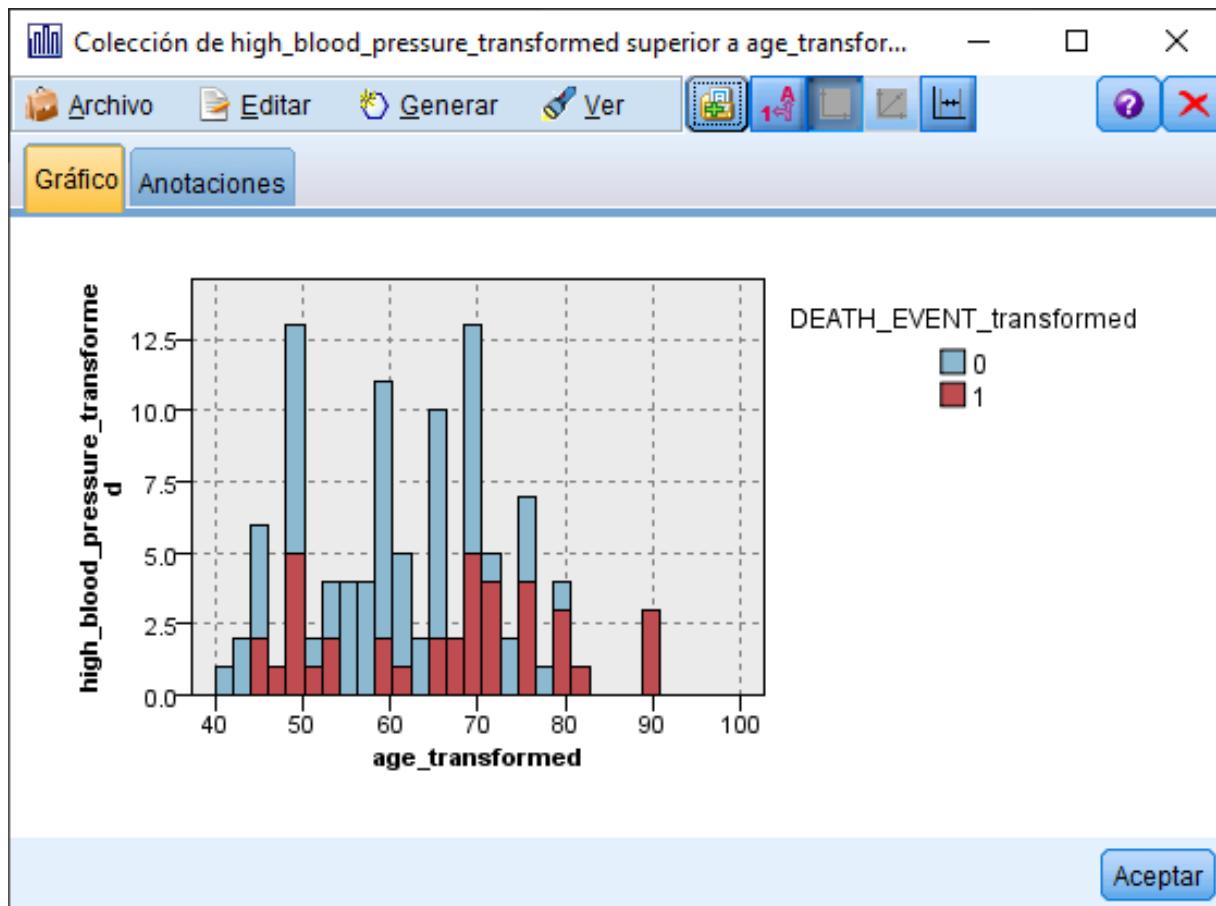


Figura 4.27: Colección de presión arterial alta respecto a la edad del paciente.

Así mismo, la figura 4.27 demuestra que, al cumplir los 70 años, un paciente tiene una alta probabilidad de tener presión arterial alta.

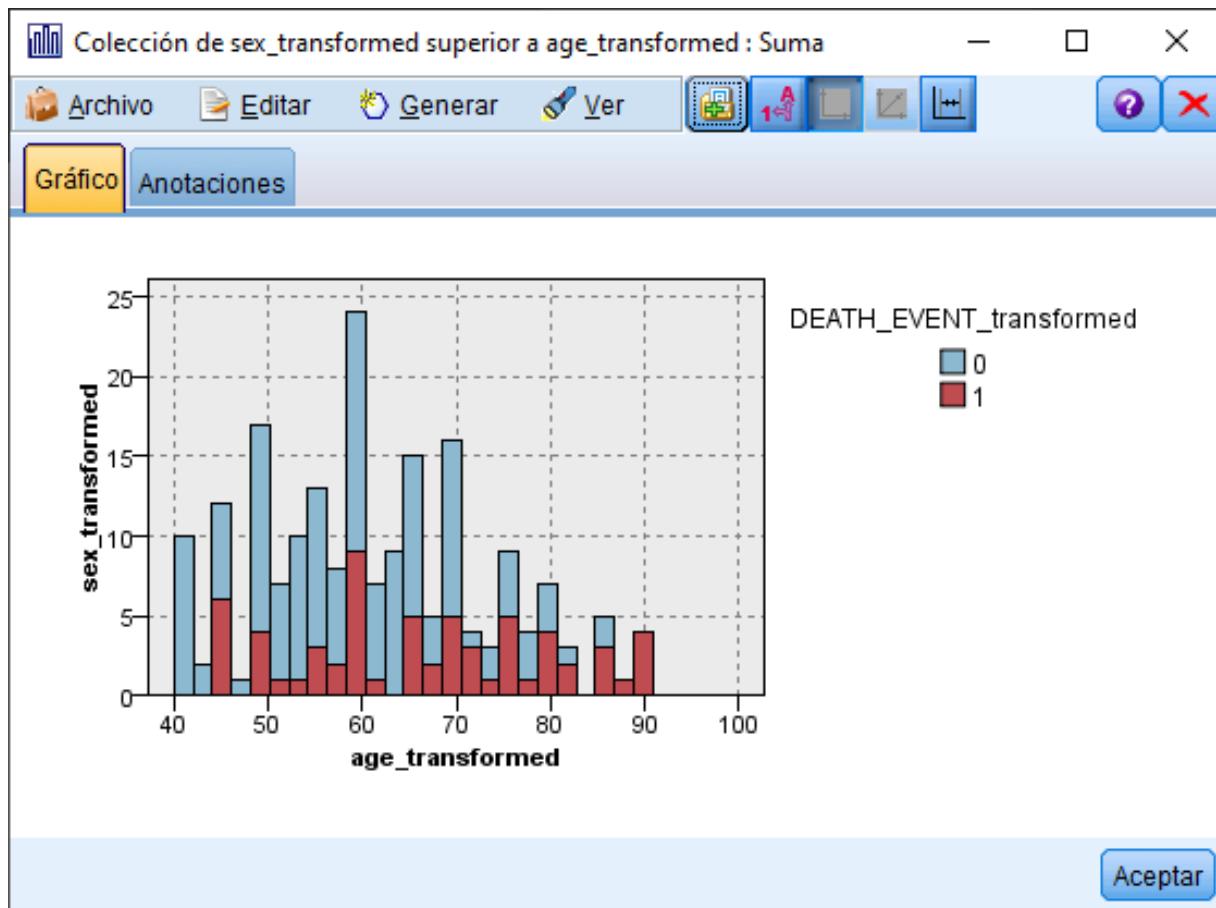


Figura 4.28: Colección de sexo respecto a la edad del paciente.

En esa misma línea, la figura 4.28 señala que gran parte de los pacientes dentro del data set son mujeres.

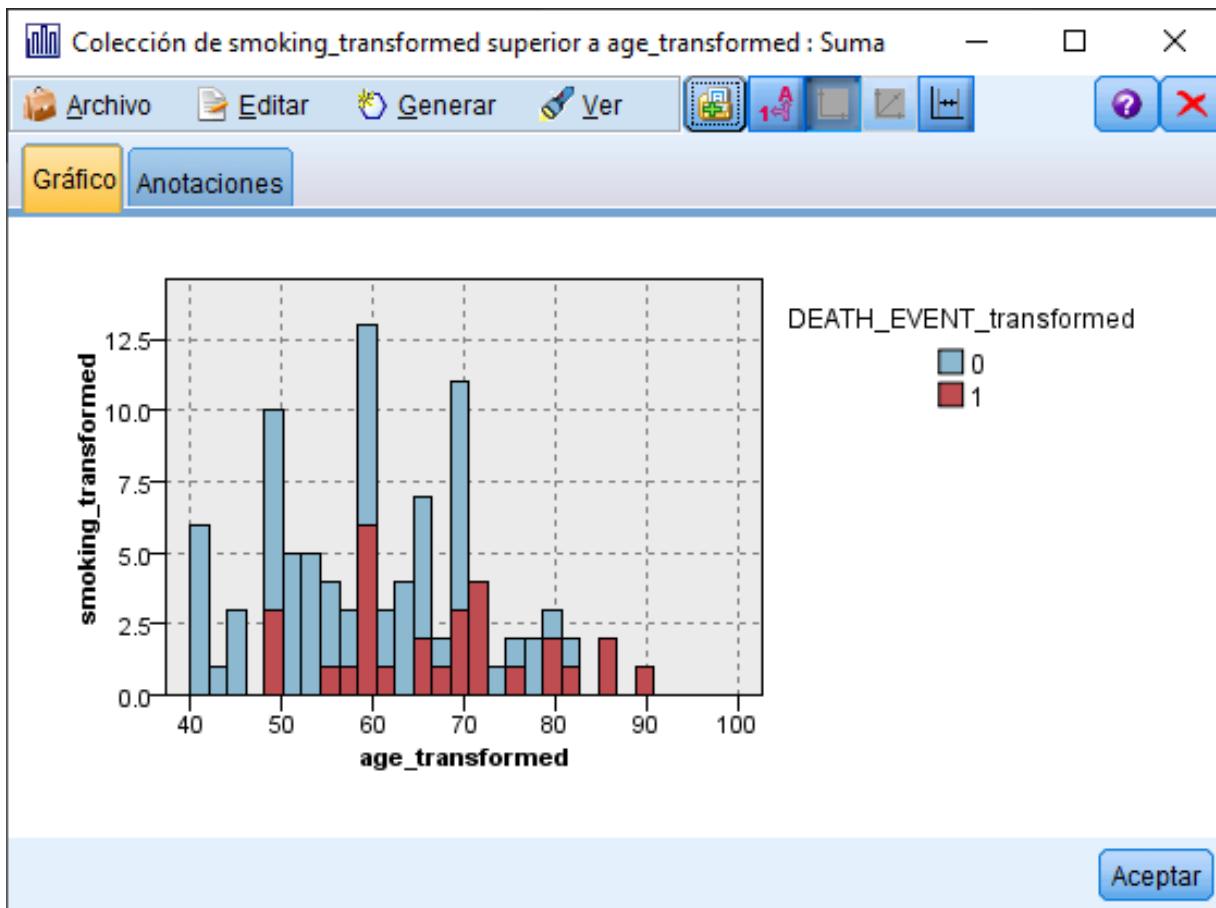


Figura 4.29: Colección de fumadores respecto a la edad del paciente.

Finalmente, la figura 4.29 prueba que, dentro de este data set, los pacientes cuya edad es inferior a los 70 años tienen baja probabilidad de ser fumadores.

4.4. Modelado

4.4.1. Selección de técnicas de modelado

Para este caso de estudio, se escogieron tres modelos que nos ayudarán a conseguir el objetivo de predicción de causas de muerte por fallos cardiacos según su historial médico. Los modelos seleccionados fueron los siguientes:

- Red Neuronal.
- Árbol de decisión.
- Algoritmo de regresión.

4.4.2. Métodos de comprobación

Dado que todos los modelos seleccionados son técnicas de aprendizaje supervisado, el método de comprobación seleccionado se encuentra en el criterio de bondad del modelo; es decir, la tasa de error del modelo.

Respecto a la relación de los datos para comprobar el criterio de bondad del modelo, cada uno de ellos usará particiones de los datos en dos conjuntos, uno para el entrenamiento del modelo y el otro de prueba o verificación de su funcionamiento.

4.4.3. Generación de los modelos

Mediante la herramienta *IBM SPSS Modeler*, se han generado los modelos antes mencionados con los parámetros por defecto que marca dicha herramienta.

Como ya se indicó, los tres modelos utilizados pertenecen a la categoría de aprendizaje supervisado, por lo que como criterio de bondad del modelo se ha considerado la tasa de error producida por el mismo.

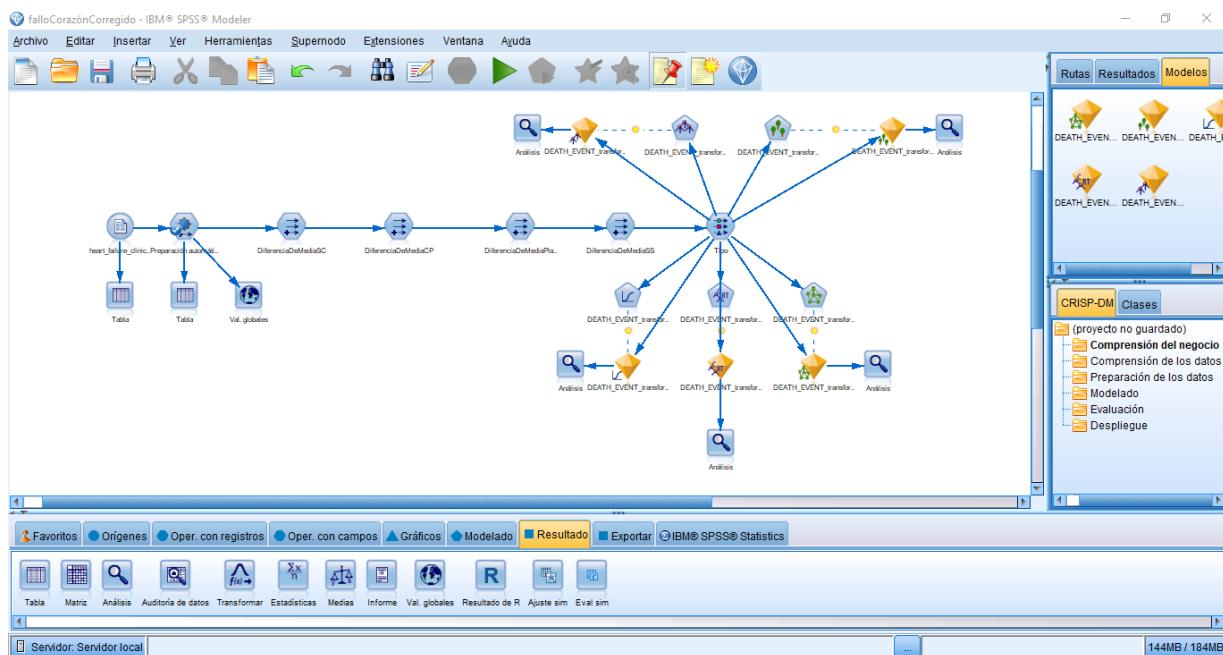


Figura 4.30: Ruta para generar modelos dentro de la herramienta IBM SPSS Modeler.

En la figura 4.30 vemos que se han generado un total de cinco diferentes modelos, los cuales se documentarán a continuación. Por otra parte, es importante mencionar que, el campo que se quitó en la sección 4.3.1 se retomó para la generación de cada modelo debido a que sí le ayudaba al modelo; además, los campos creados en la subsección 4.3.3 no se tomaron en cuenta para la generación de cada modelo debido a la nula eficiencia que tenían para aumentar la precisión del mismo.

4.4.4. Modelo Árbol de decisión C&R

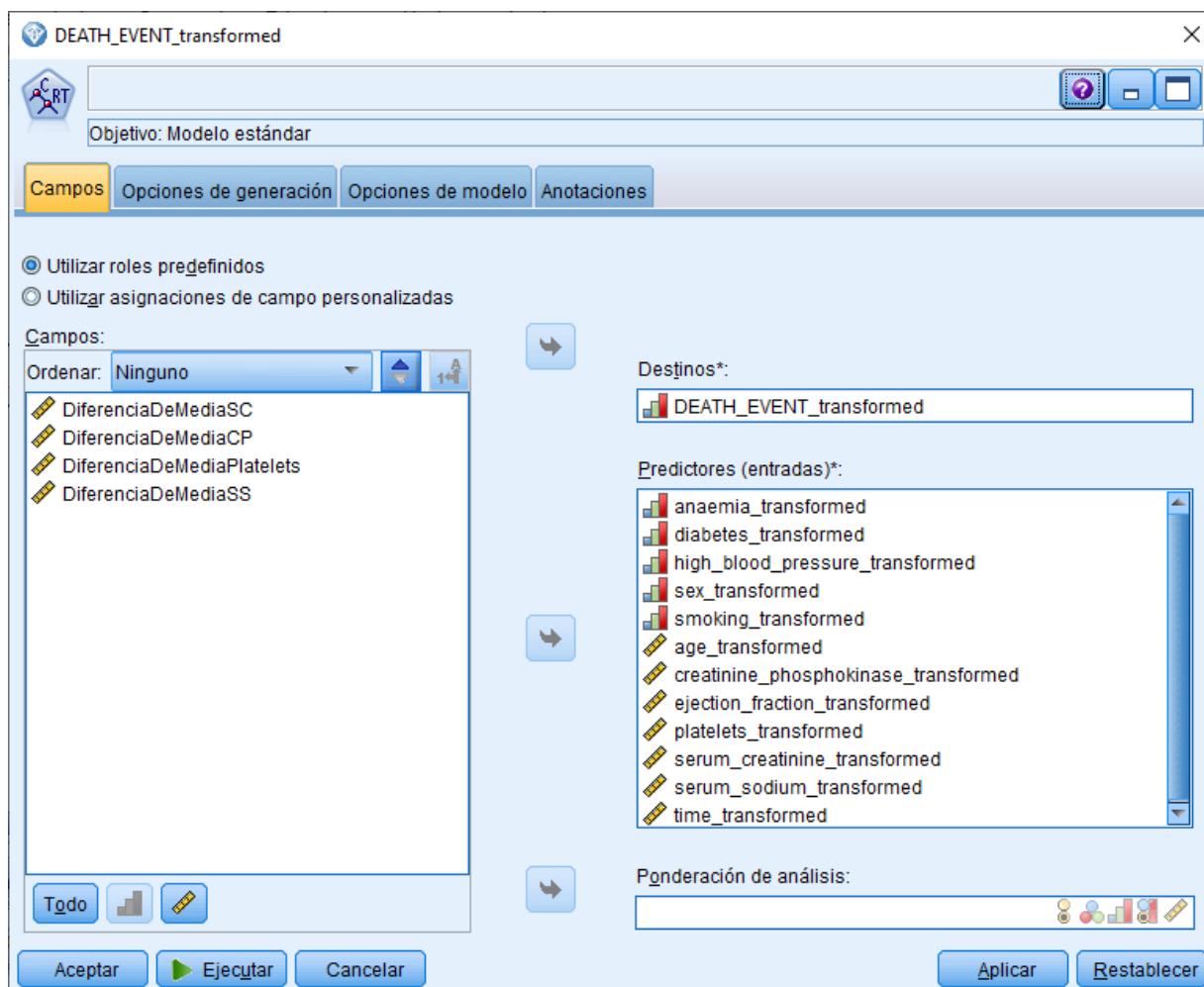


Figura 4.31: Configuración del modelo Árbol de decisión C&R - 1.

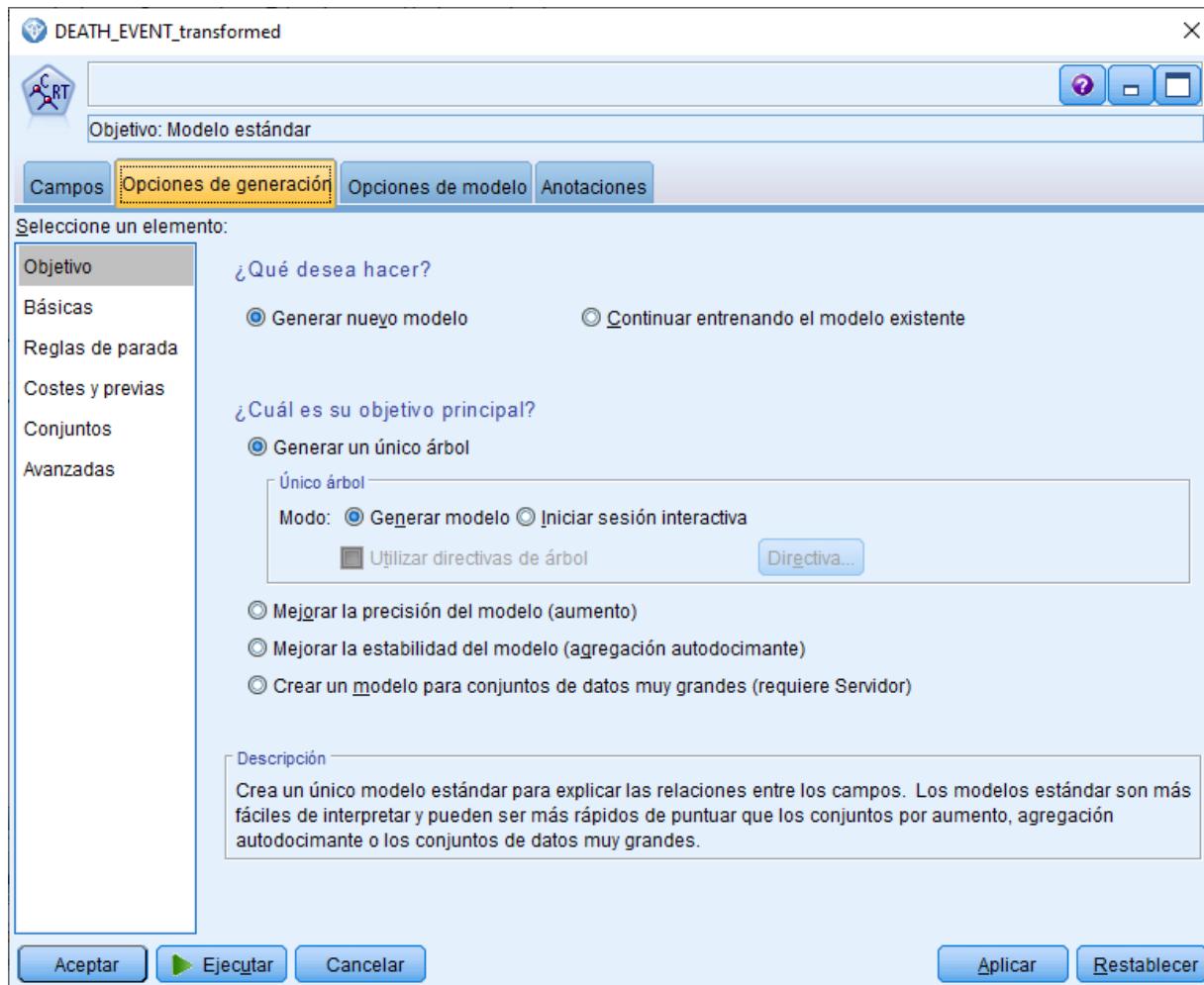


Figura 4.32: Configuración del modelo Árbol de decisión C&R - 2.

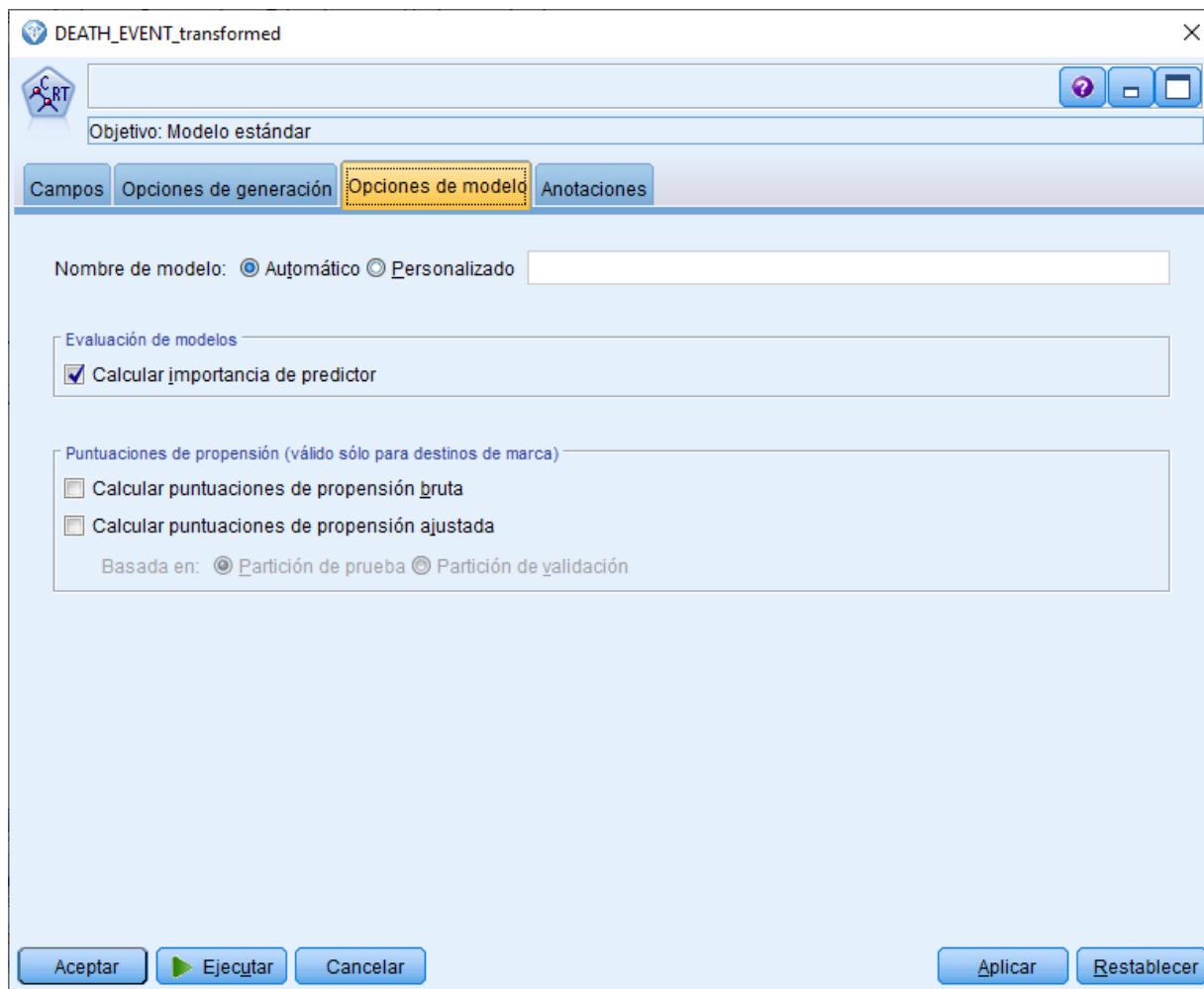


Figura 4.33: Configuración del modelo Árbol de decisión C&R - 3.

En las figuras 4.31, 4.32 y 4.33 vemos que la configuración es la predeterminada por la herramienta, para este modelo, no era necesario realizar cambios.

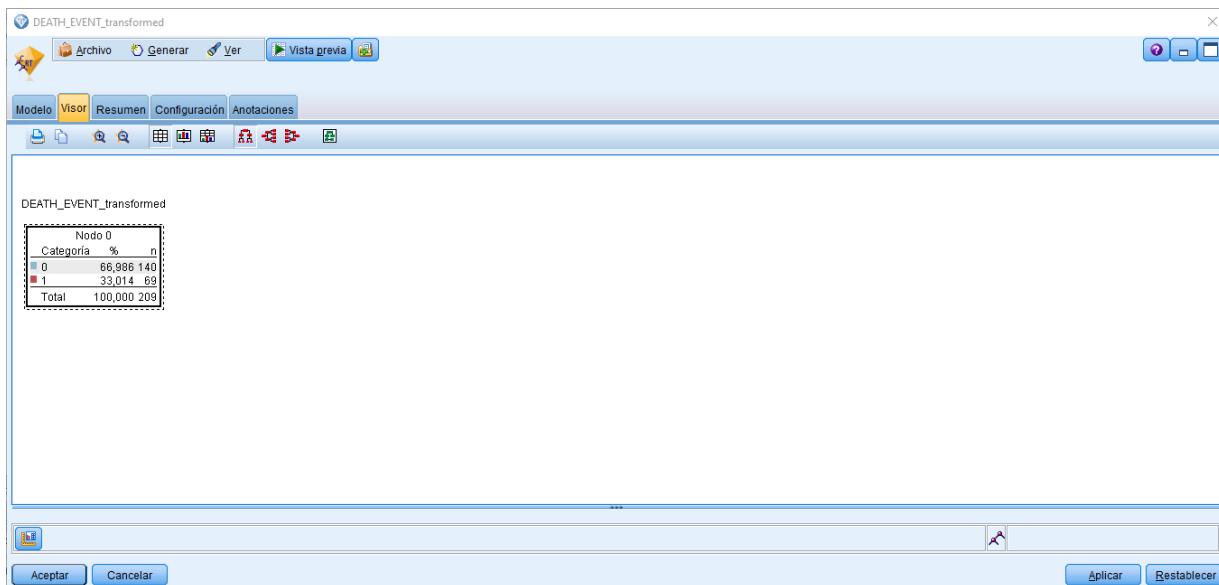


Figura 4.34: Ejecución del modelo Árbol de decisión C&R - 1.

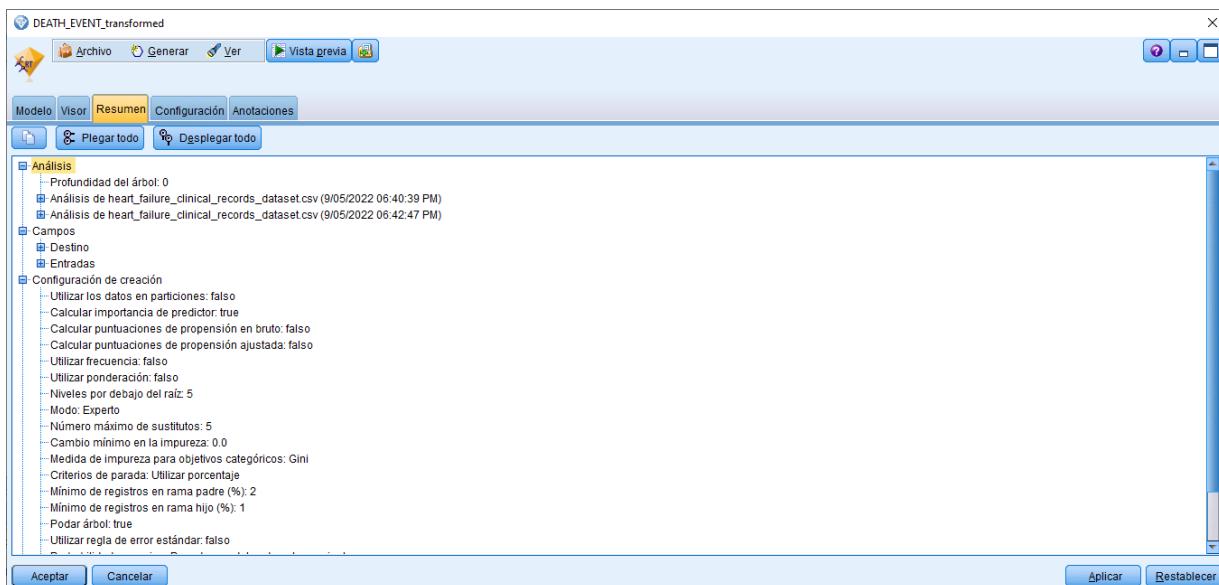


Figura 4.35: Ejecución del modelo Árbol de decisión C&R - 2.

En contraste, las figuras 4.34 y 4.35 vemos que la ejecución del modelo únicamente creó un árbol de un nodo, lo cual puede indicar que su funcionamiento no será el esperado.

4.4.5. Modelo Red Neuronal Perceptron Backpropagation

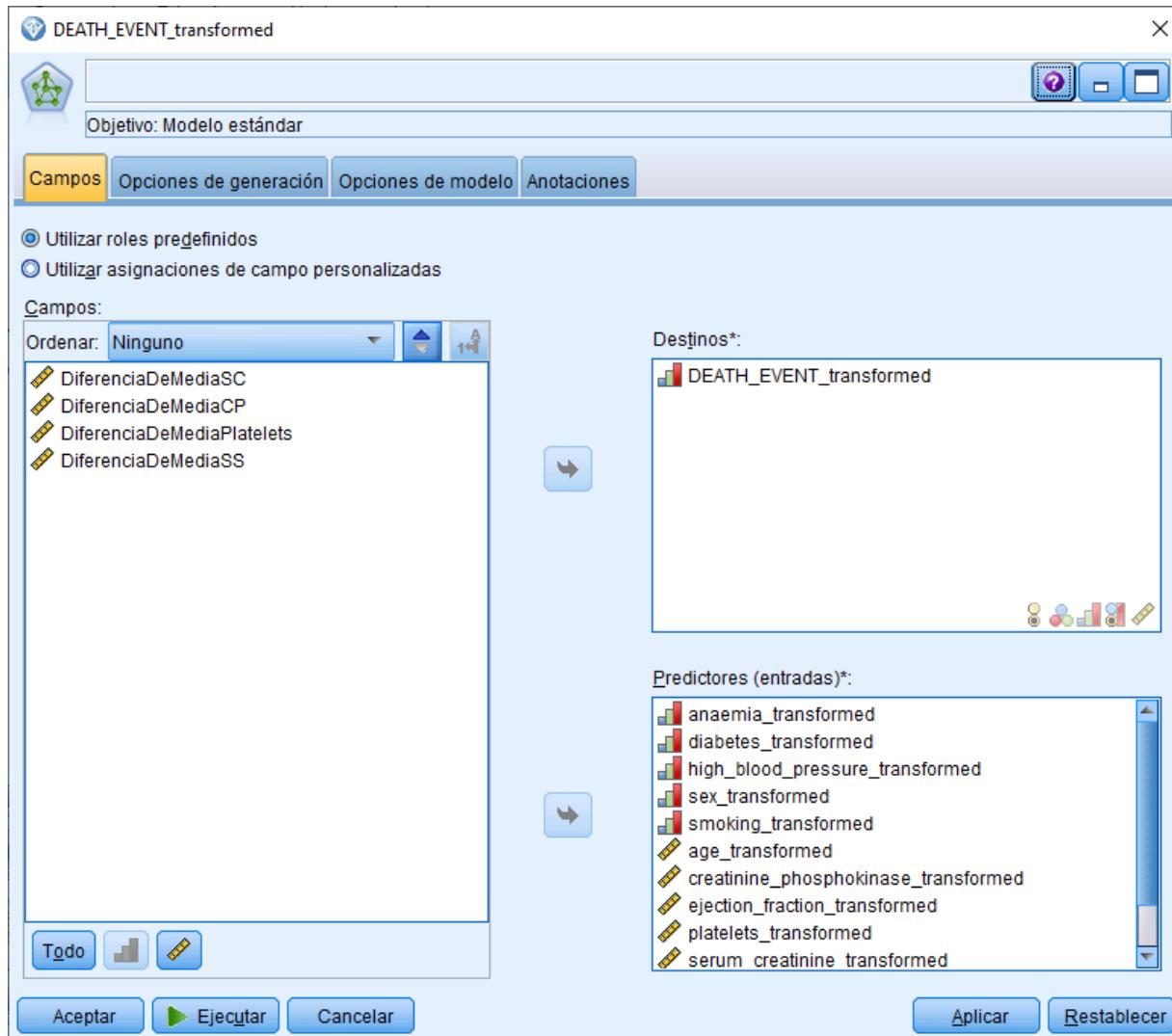


Figura 4.36: Configuración del modelo Red Neuronal Perceptron Backpropagation - 1.

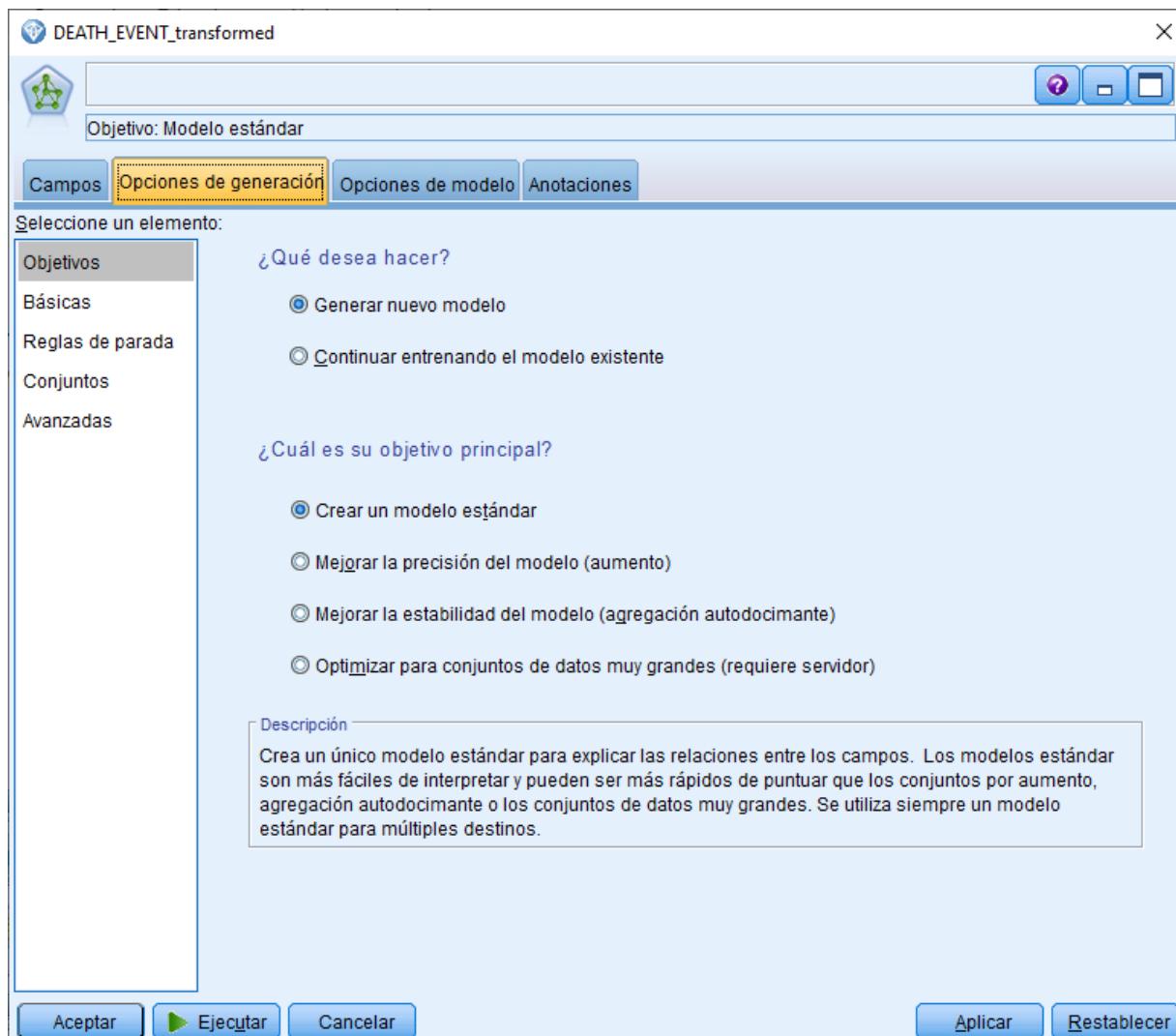


Figura 4.37: Configuración del modelo Red Neuronal Perceptron Backpropagation - 2.

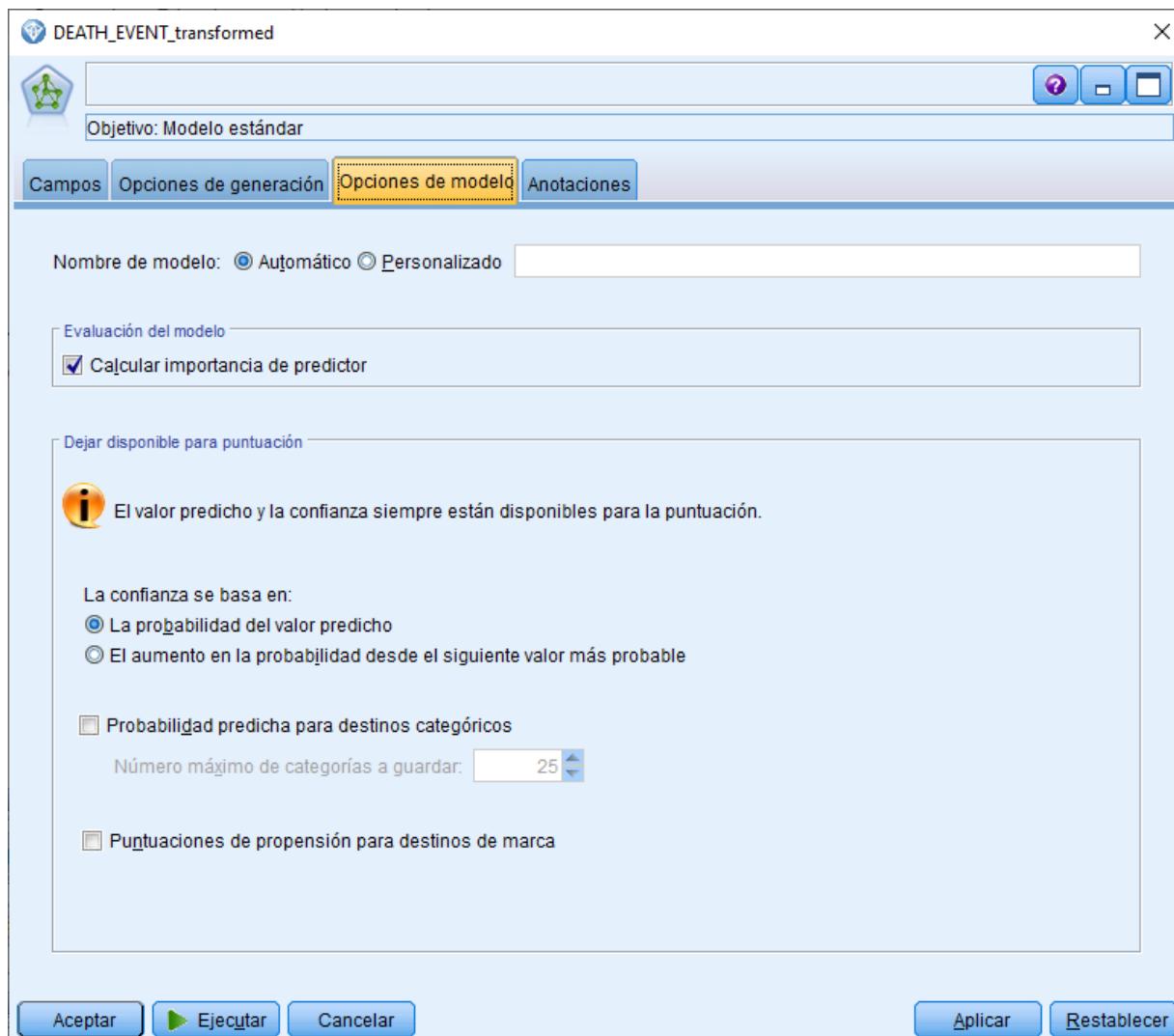


Figura 4.38: Configuración del modelo Red Neuronal Perceptron Backpropagation - 3.

Por otra parte, las figuras 4.36, 4.37 y 4.38 nos muestran la configuración predeterminada para la creación de una red neuronal Perceptron Backpropagation, para este modelo, no era necesario realizar cambios.

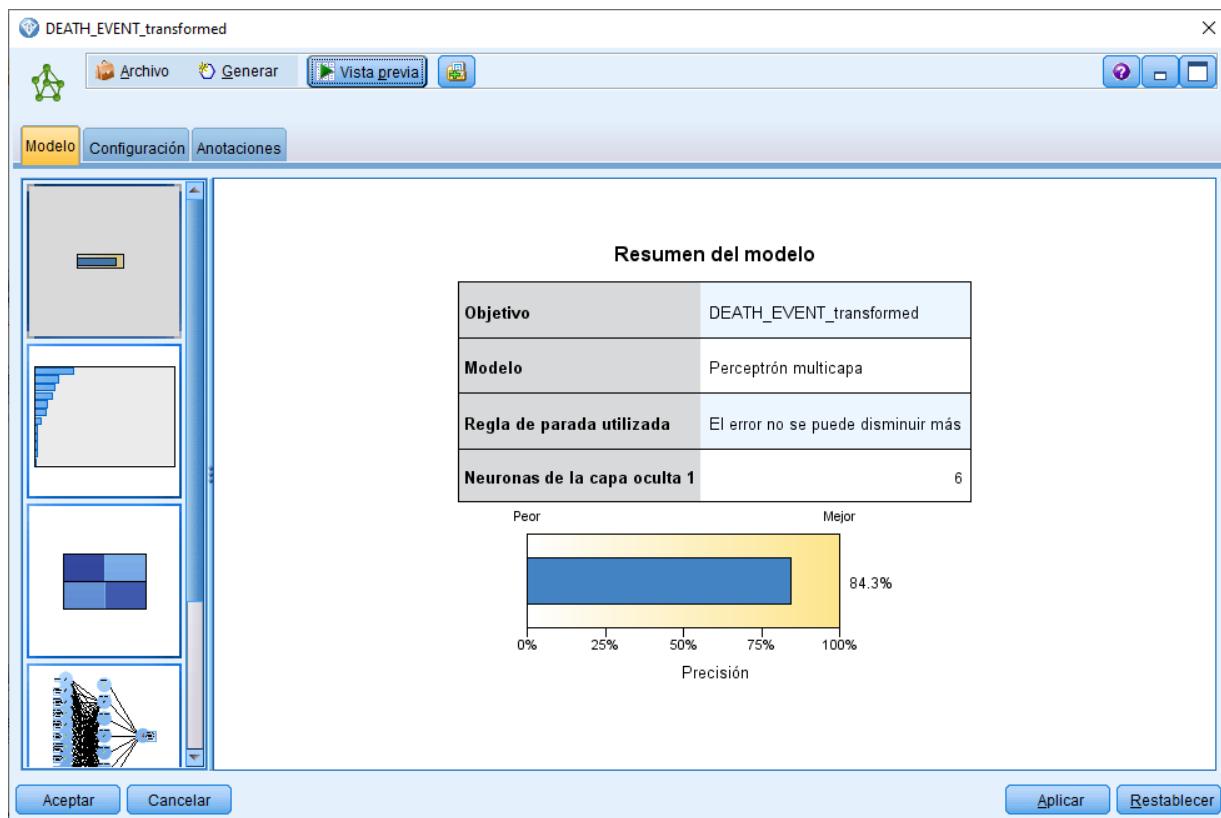


Figura 4.39: Ejecución del modelo Red Neuronal Perceptron Backpropagation - 1.

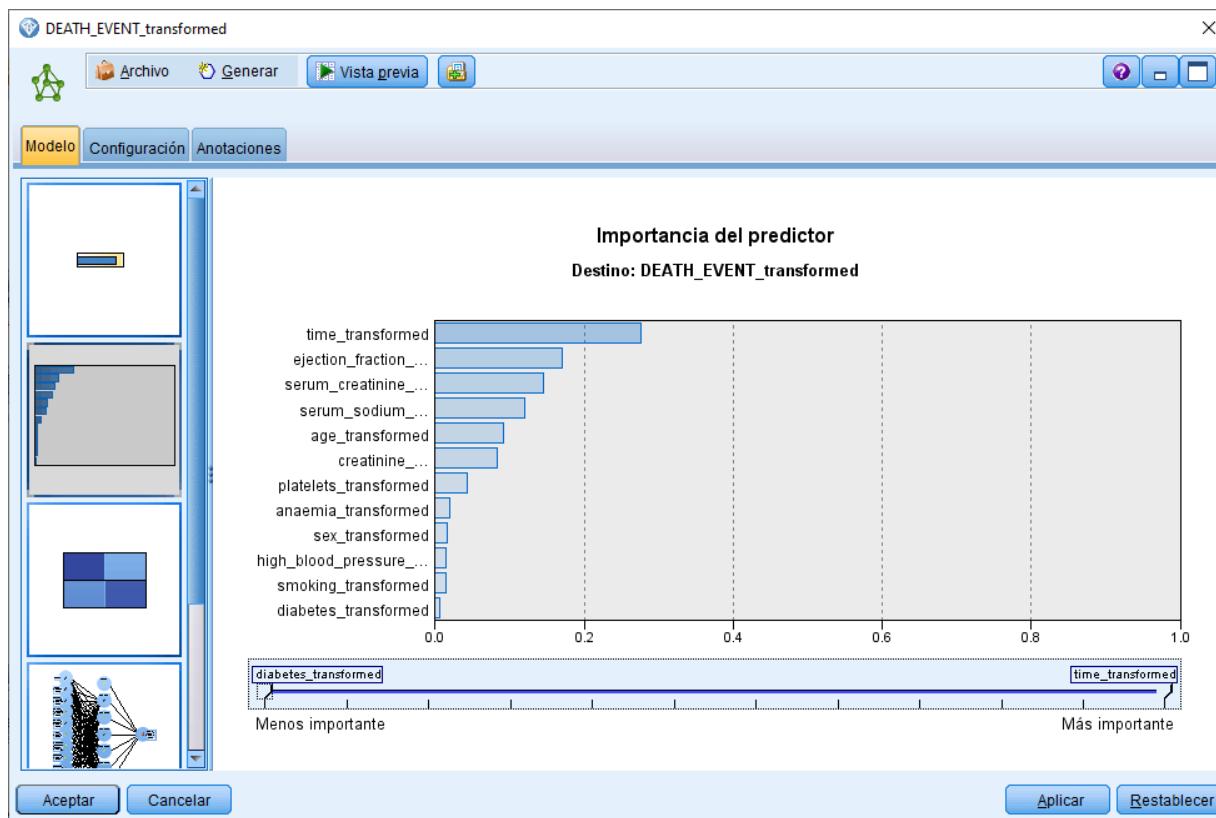


Figura 4.40: Ejecución del modelo Red Neuronal Perceptron Backpropagation - 2.

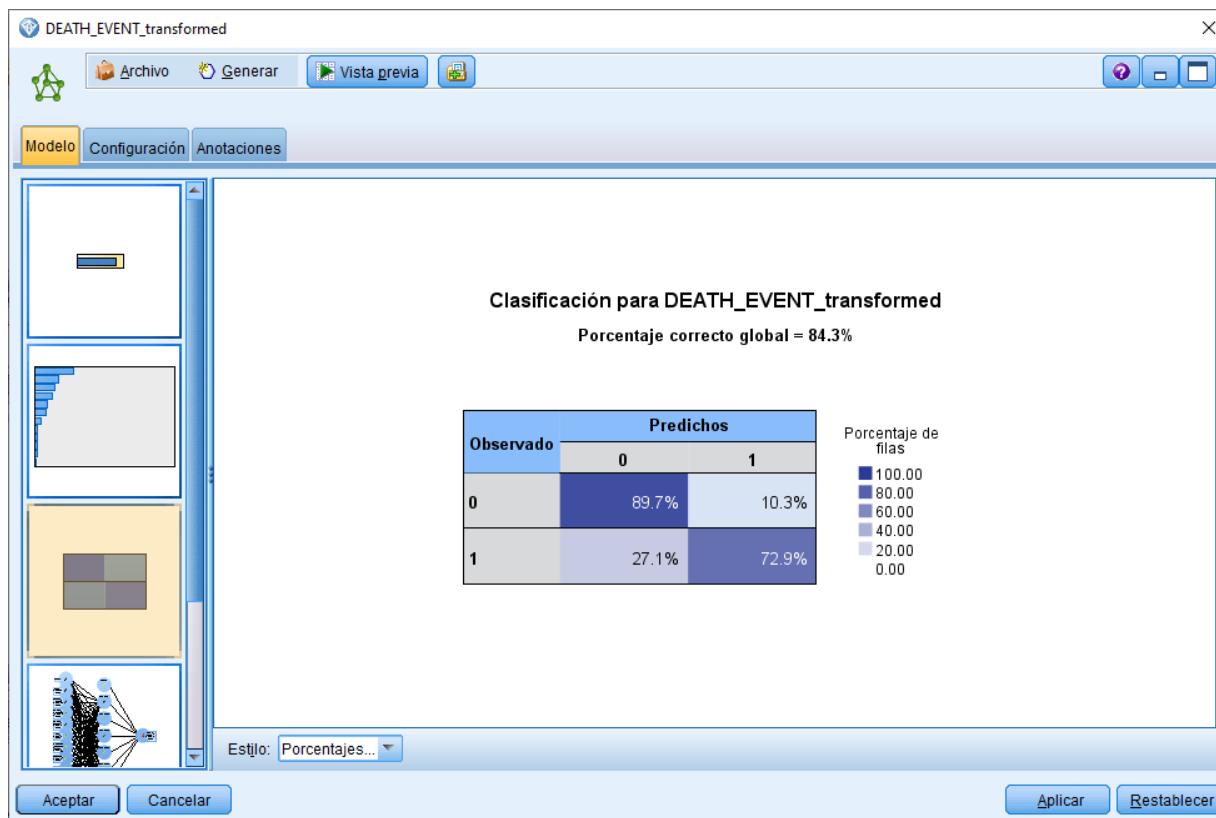


Figura 4.41: Ejecución del modelo Red Neuronal Perceptron Backpropagation - 3.

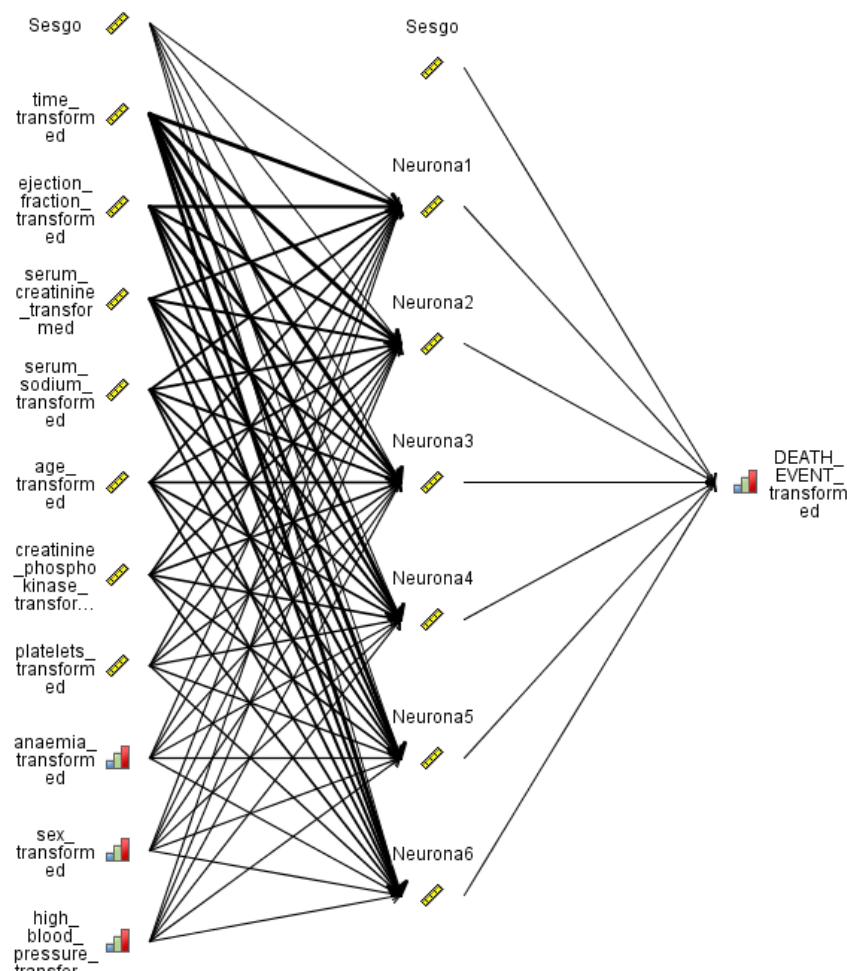


Figura 4.42: Ejecución del modelo Red Neuronal Perceptron Backpropagation - 4.

Dentro de las figuras 4.39, 4.40, 4.41 y 4.42 vemos que el resumen de la red neuronal muestra una precisión superior al 80 %, la importancia de los predictores, la clasificación de los datos de entrenamiento y el diagrama del modelo.

4.4.6. Modelo Regresión logística

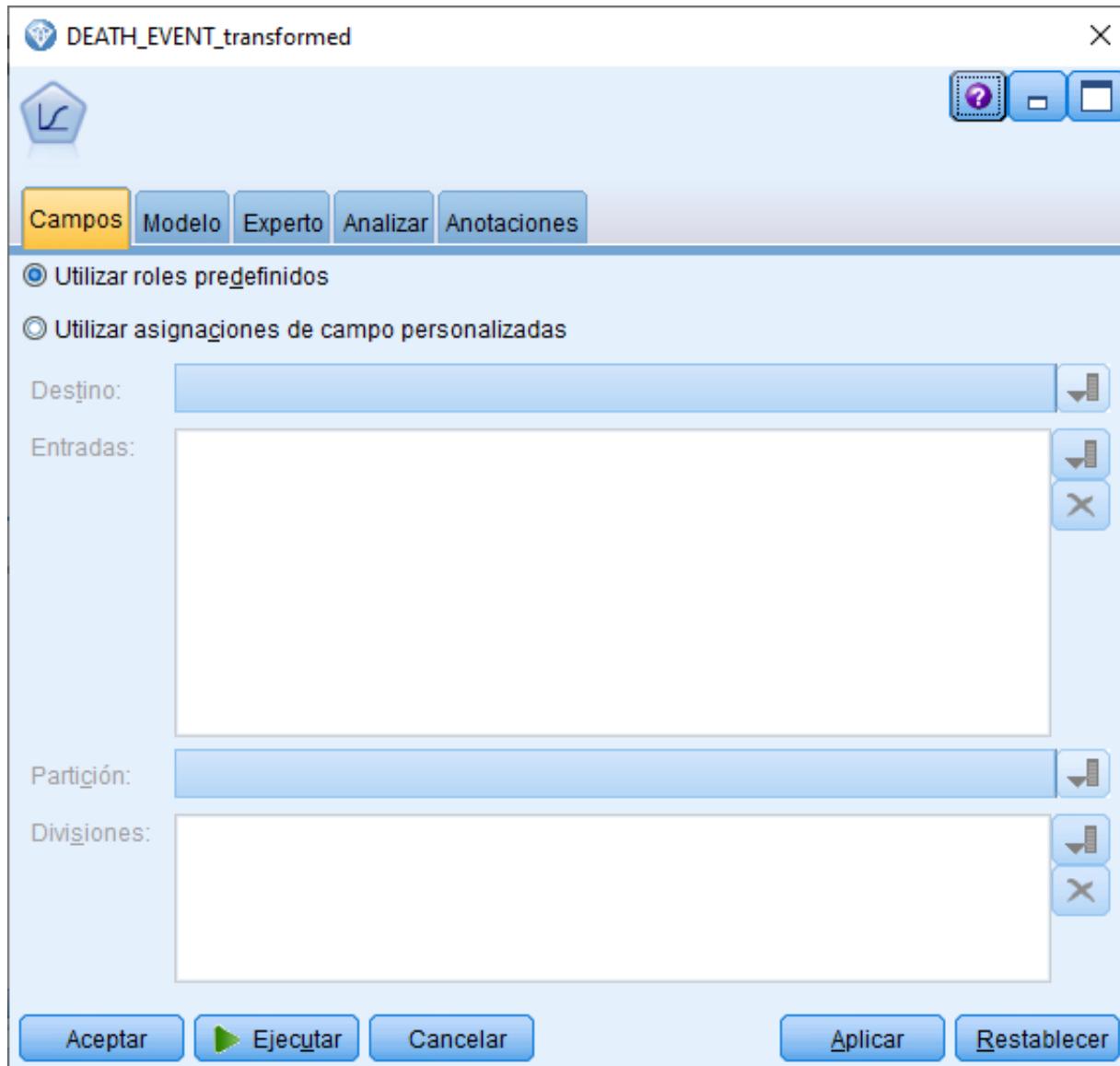


Figura 4.43: Configuración del modelo Regresión logística - 1.

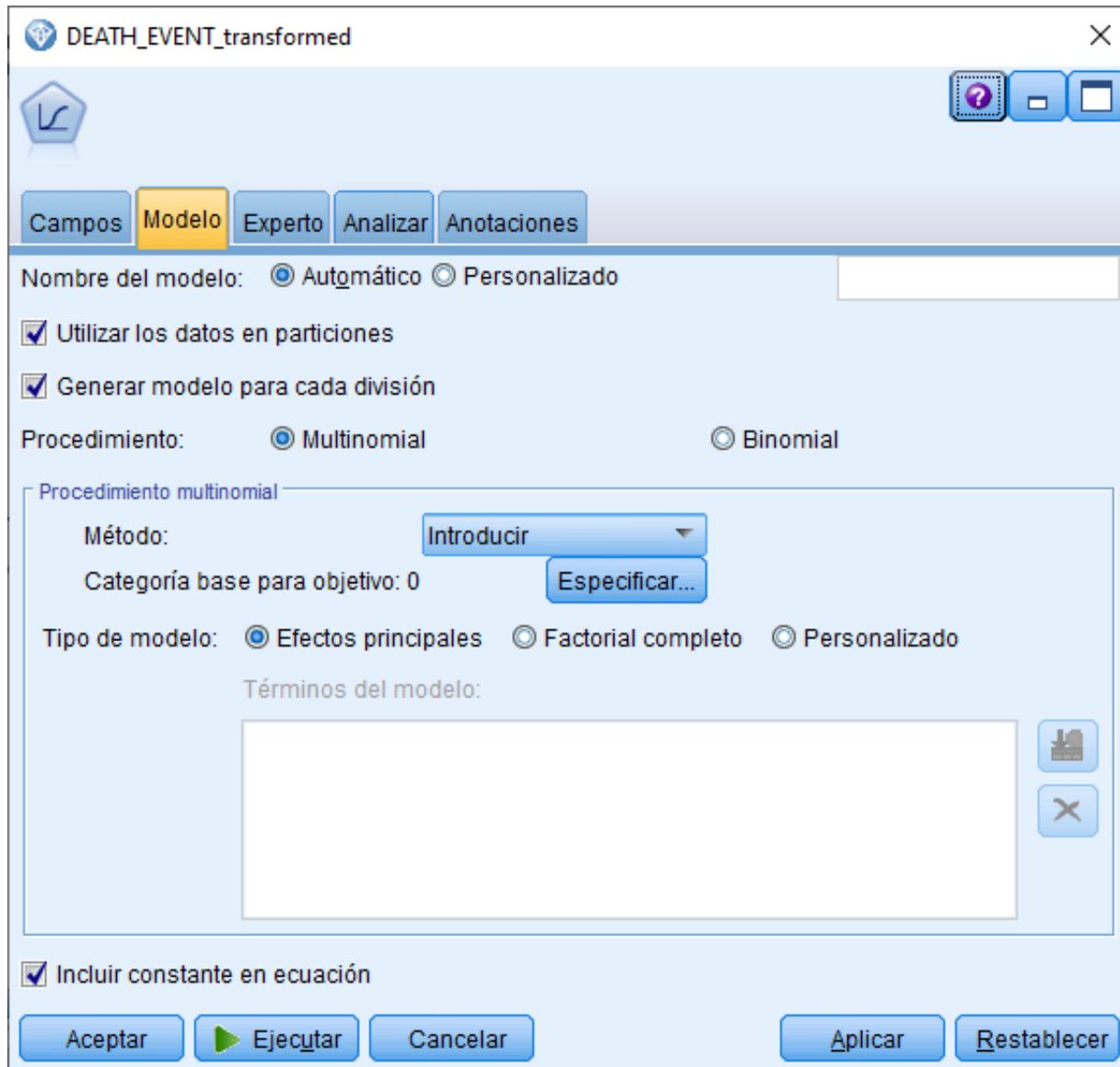


Figura 4.44: Configuración del modelo Regresión logística - 2.

En el modelo de Regresión Logística, las figuras 4.43 y 4.44 se demuestra que se ha utilizado la configuración predeterminada por la herramienta.

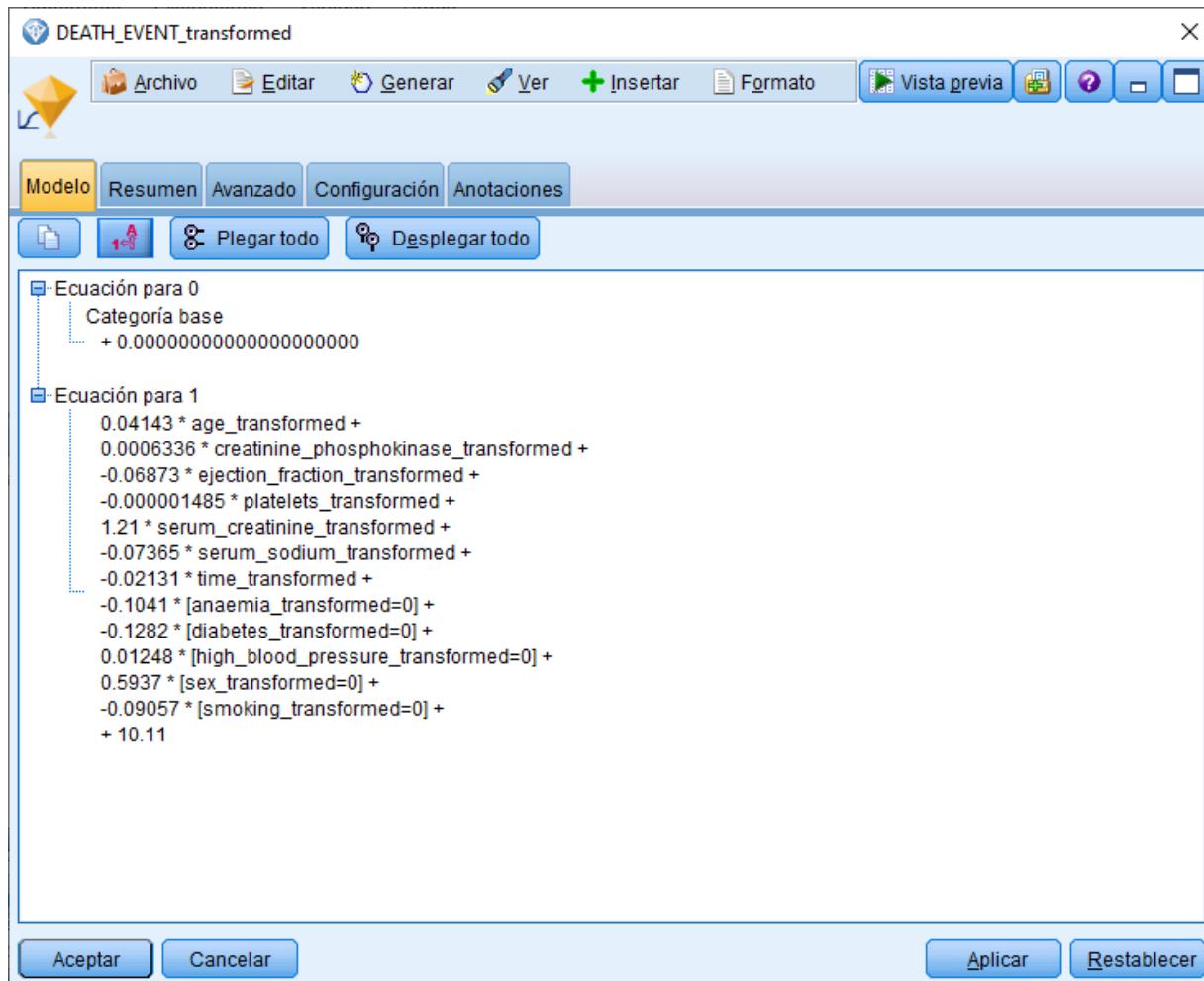


Figura 4.45: Ejecución del modelo Regresión logística - 1.

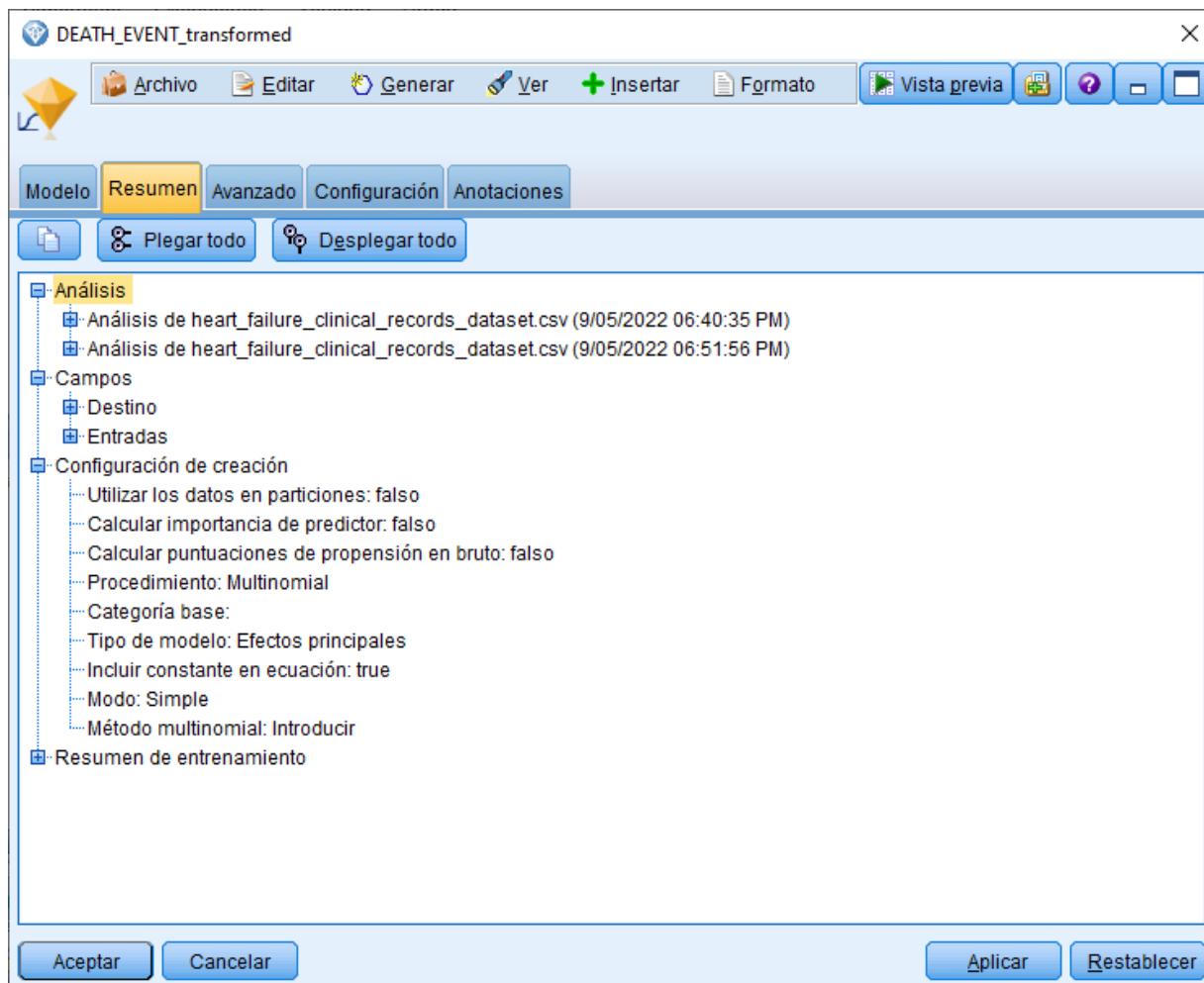


Figura 4.46: Ejecución del modelo Regresión logística - 2.

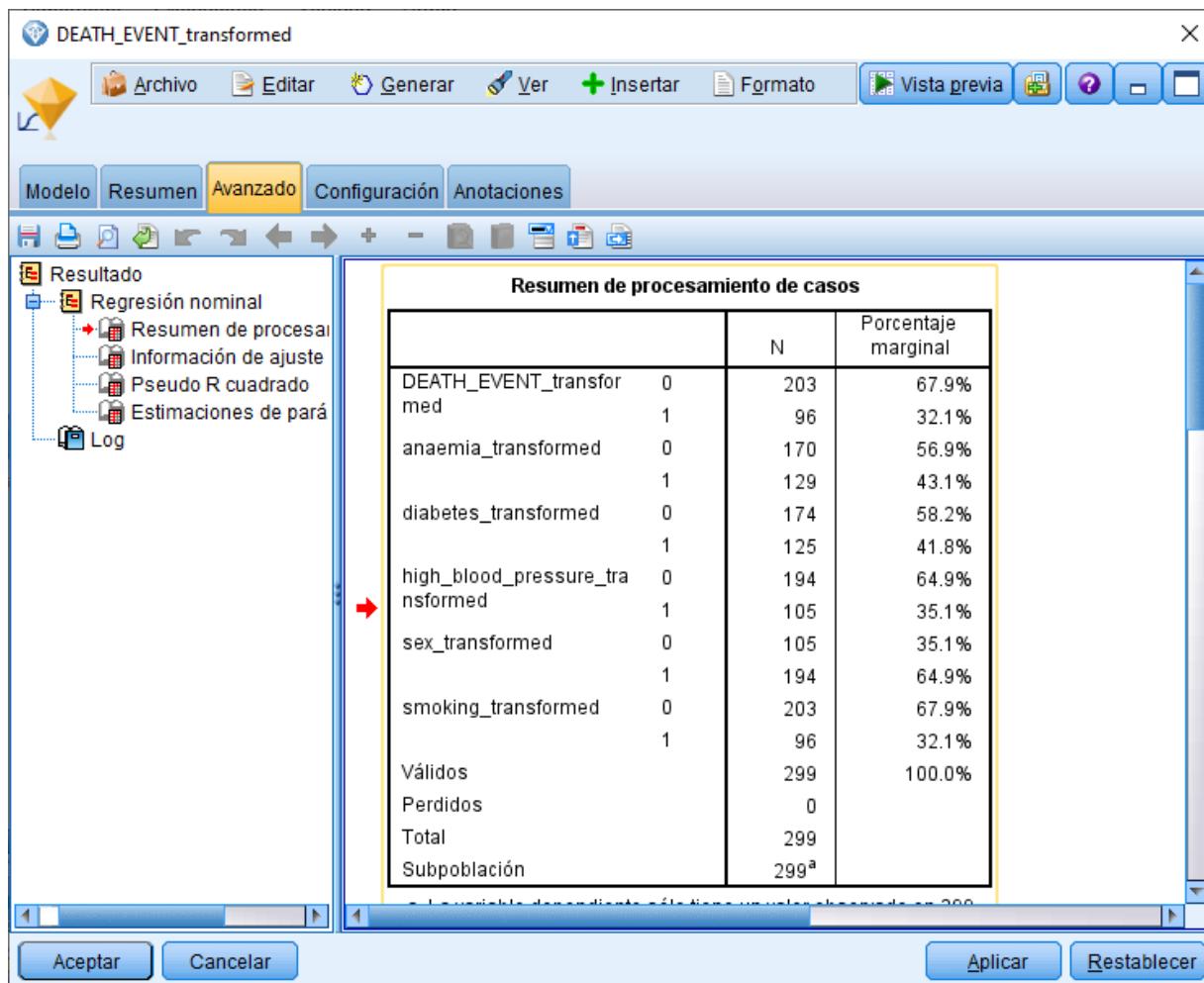


Figura 4.47: Ejecución del modelo Regresión logística - 3.

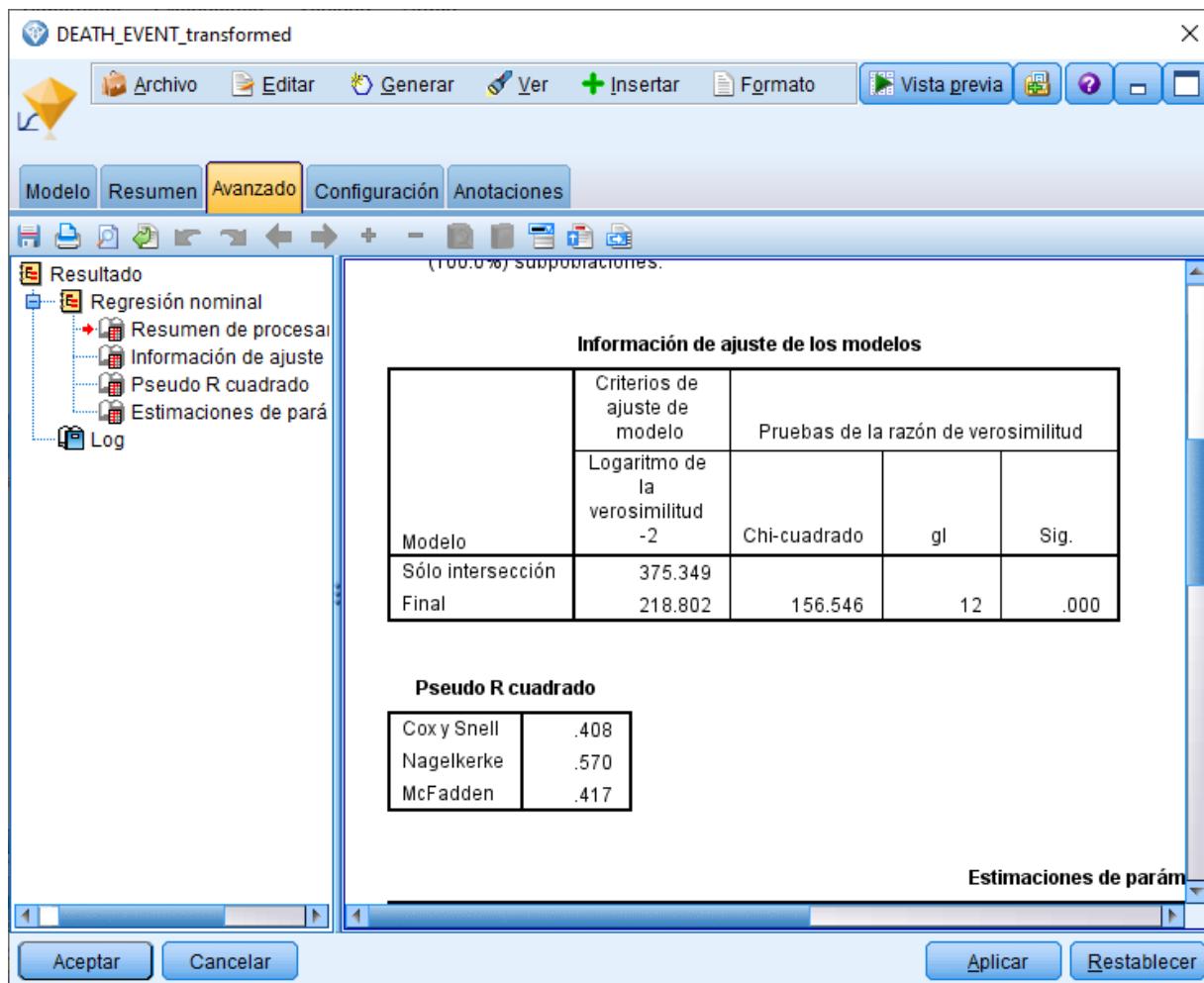


Figura 4.48: Ejecución del modelo Regresión logística - 4.

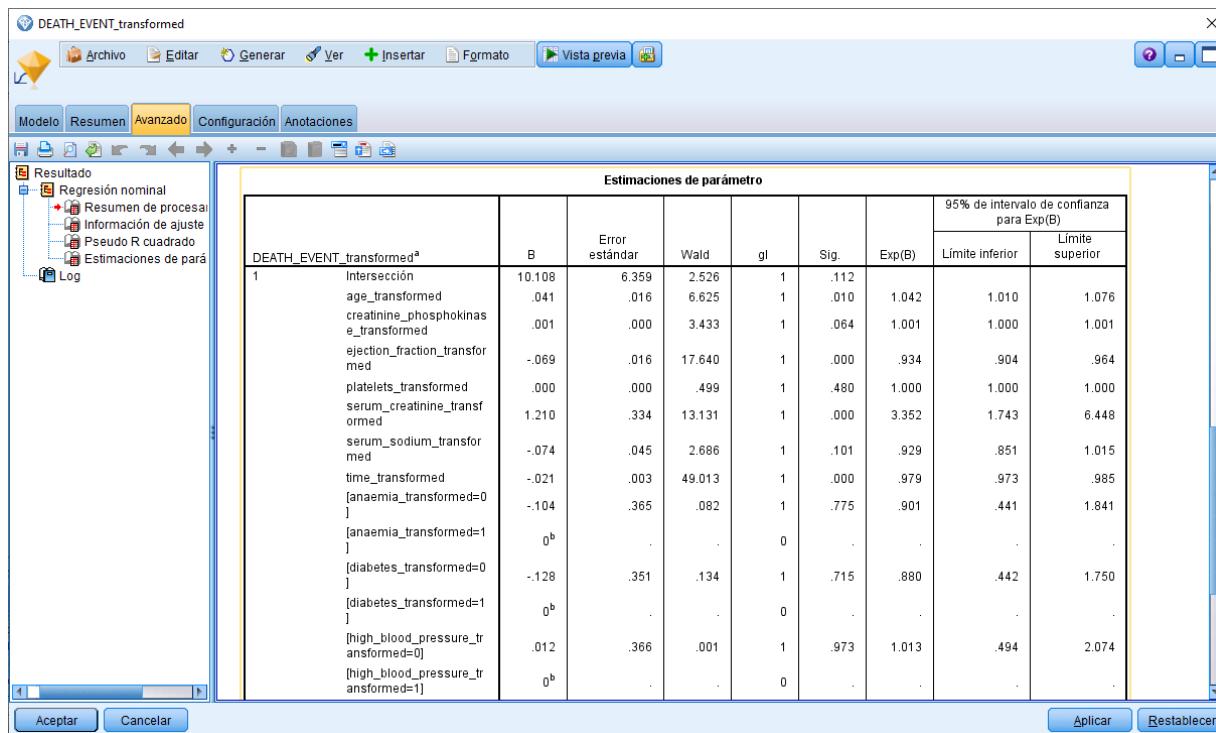


Figura 4.49: Ejecución del modelo Regresión logística - 5.

Para la ejecución de este modelo, en las figuras 4.45, 4.46, 4.47, 4.48 y 4.49 vemos el tipo de ecuación que se generó, el resumen de procesamiento de los casos, la información de ajuste de los modelos, las estimaciones de parámetro, entre otros.

4.4.7. Modelo Árbol aleatorio

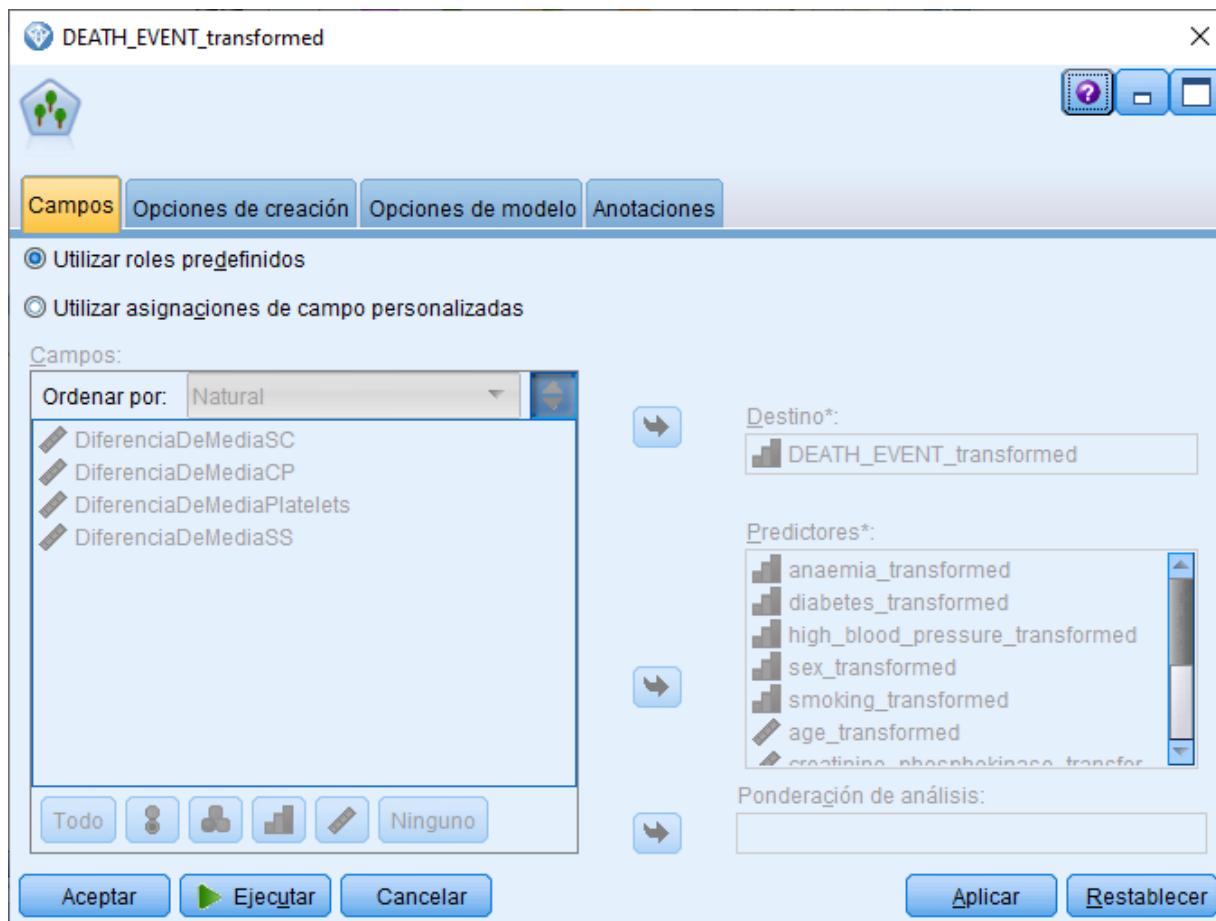


Figura 4.50: Configuración del modelo Árbol aleatorio - 1.

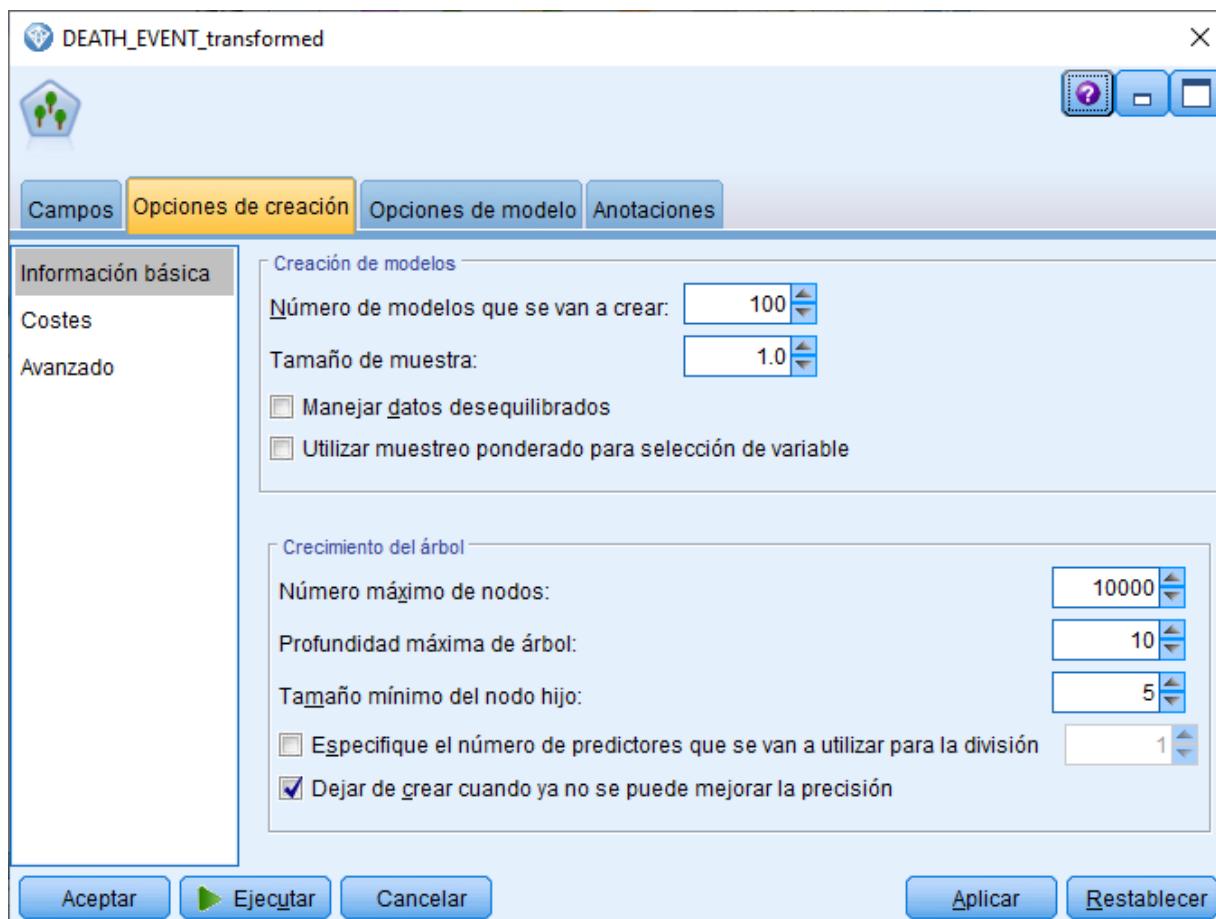


Figura 4.51: Configuración del modelo Árbol aleatorio - 2.

Para la configuración de un árbol aleatorio, las figuras 4.50 y 4.51 prueban que se ha utilizado la configuración predeterminada por la herramienta, pues no era necesario realizar cambios.

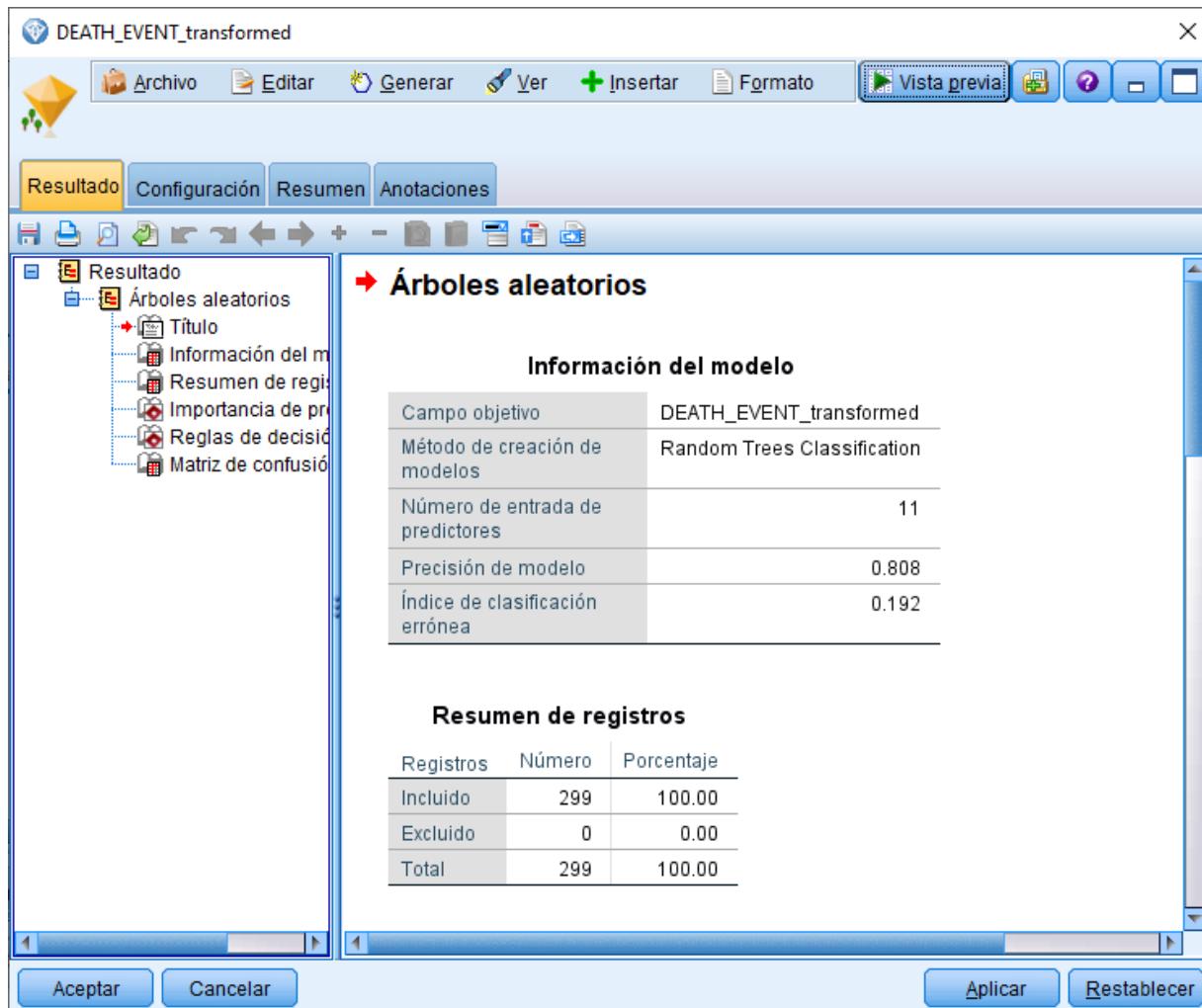


Figura 4.52: Ejecución del modelo Árbol aleatorio - 1.

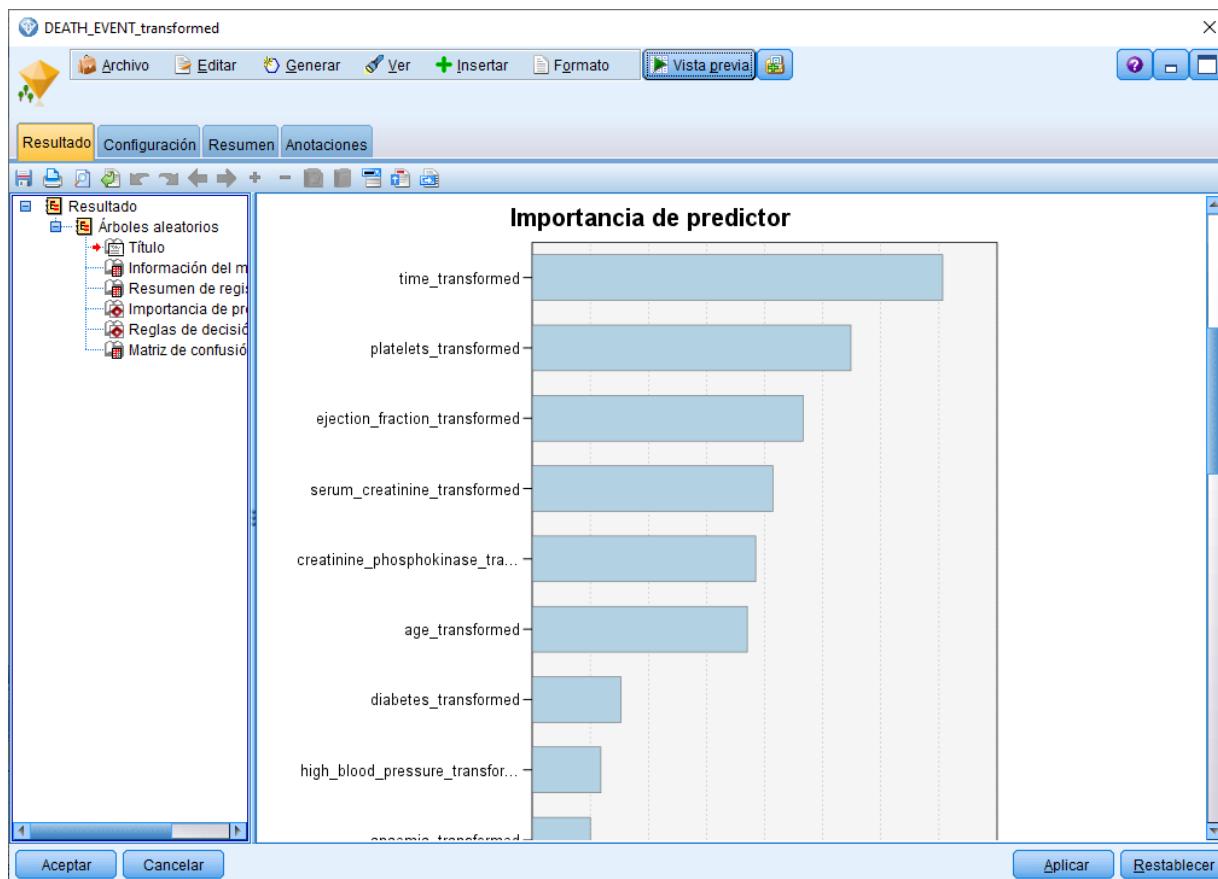


Figura 4.53: Ejecución del modelo Árbol aleatorio - 2.

Reglas de decisión principales para 'DEATH_EVENT_transformed'

Regla de decisión	Categoría más frecuente	Precisión de regla	Precisión de bosque	Índice de grado de interés
(creatinine_phosphokinase_transformed <= 1211.0) and (time_transformed > 59.0) and (anaemia_transformed > 0) and (serum_creatinine_transformed <= 1.4) and (ejection_fraction_transformed > 25.0)	0	1.000	1.000	1.000
(time_transformed > 59.0) and (age_transformed <= 75.0) and (anaemia_transformed <= 0) and (serum_creatinine_transformed <= 1.4) and (ejection_fraction_transformed > 25.0)	0	1.000	1.000	1.000
(creatinine_phosphokinase_transformed > 99.0) and (diabetes_transformed <= 0) and (time_transformed <= 59.0) and (serum_creatinine_transformed > 1.18)	1	1.000	1.000	1.000

Figura 4.54: Ejecución del modelo Árbol aleatorio - 3.

Observado	Pronosticado		Corrección de proporción
	0	1	
0	170	25	0.87
1	32	70	0.69
Corrección de proporción	0.84	0.74	0.81

Figura 4.55: Ejecución del modelo Árbol aleatorio - 4.

No obstante, en las figuras 4.52, 4.53, 4.54 y 4.55 su ejecución nos muestra la información del modelo, el resumen de los registros, importancia de los predictores, las reglas de decisión para determinar el campo destino y su respectiva matriz de confusión.

4.4.8. Modelo Árbol de decisión en Analytic Server

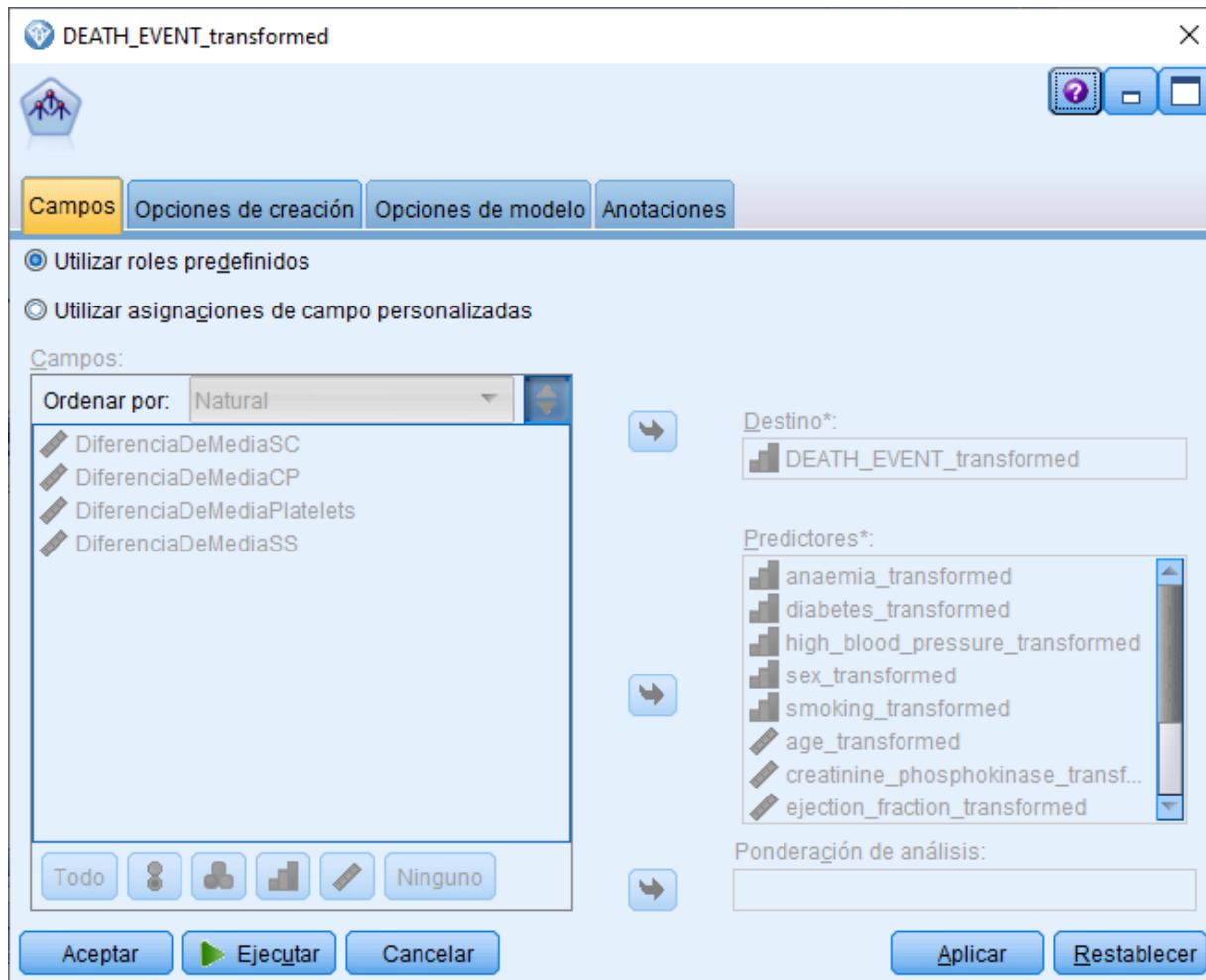


Figura 4.56: Configuración del modelo Árbol de decisión en Analytic Server - 1.

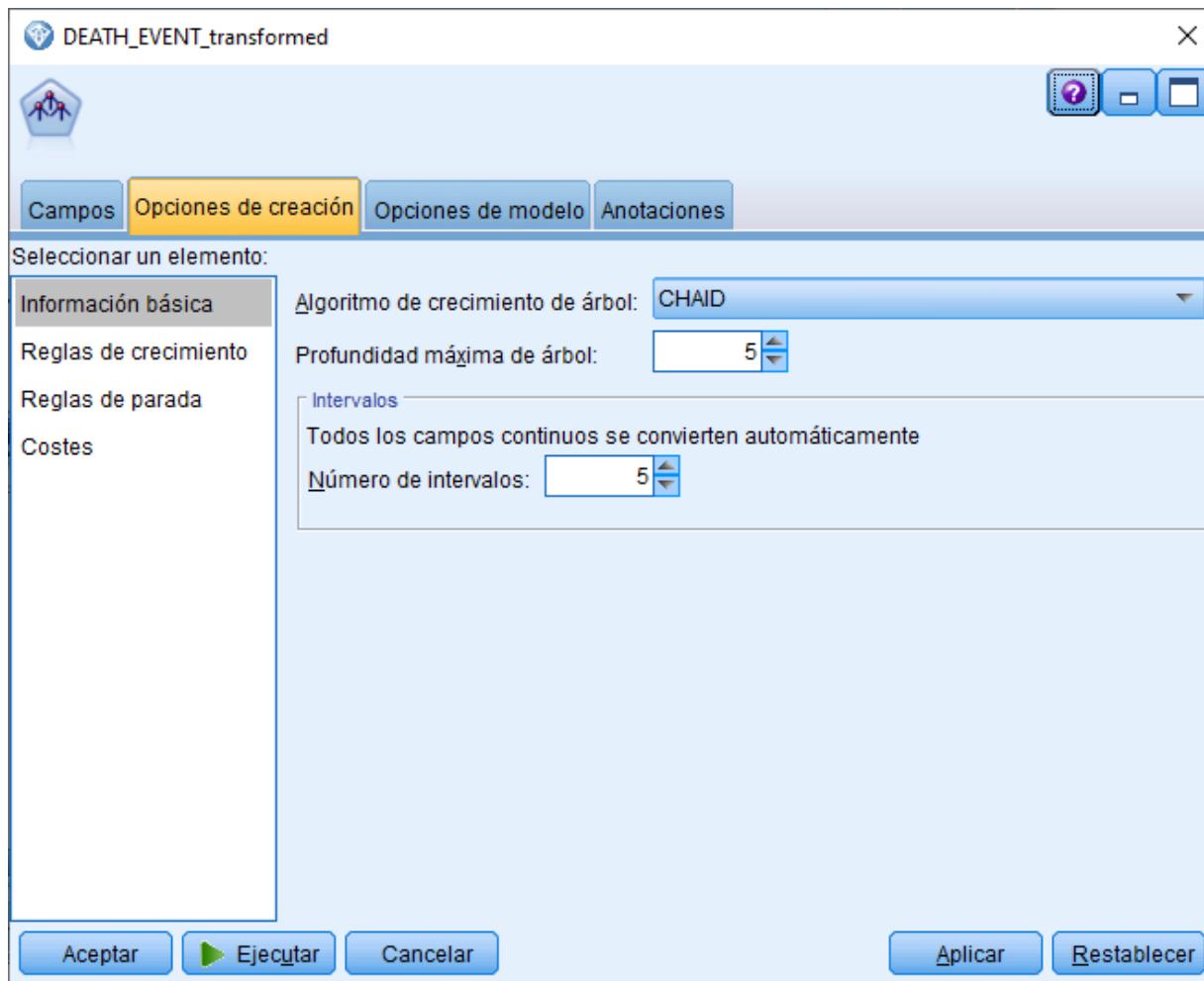


Figura 4.57: Configuración del modelo Árbol de decisión en Analytic Server - 2.

Finalmente, las figuras 4.56 y 4.57 manifiestan que la configuración es la predeterminada por la herramienta.

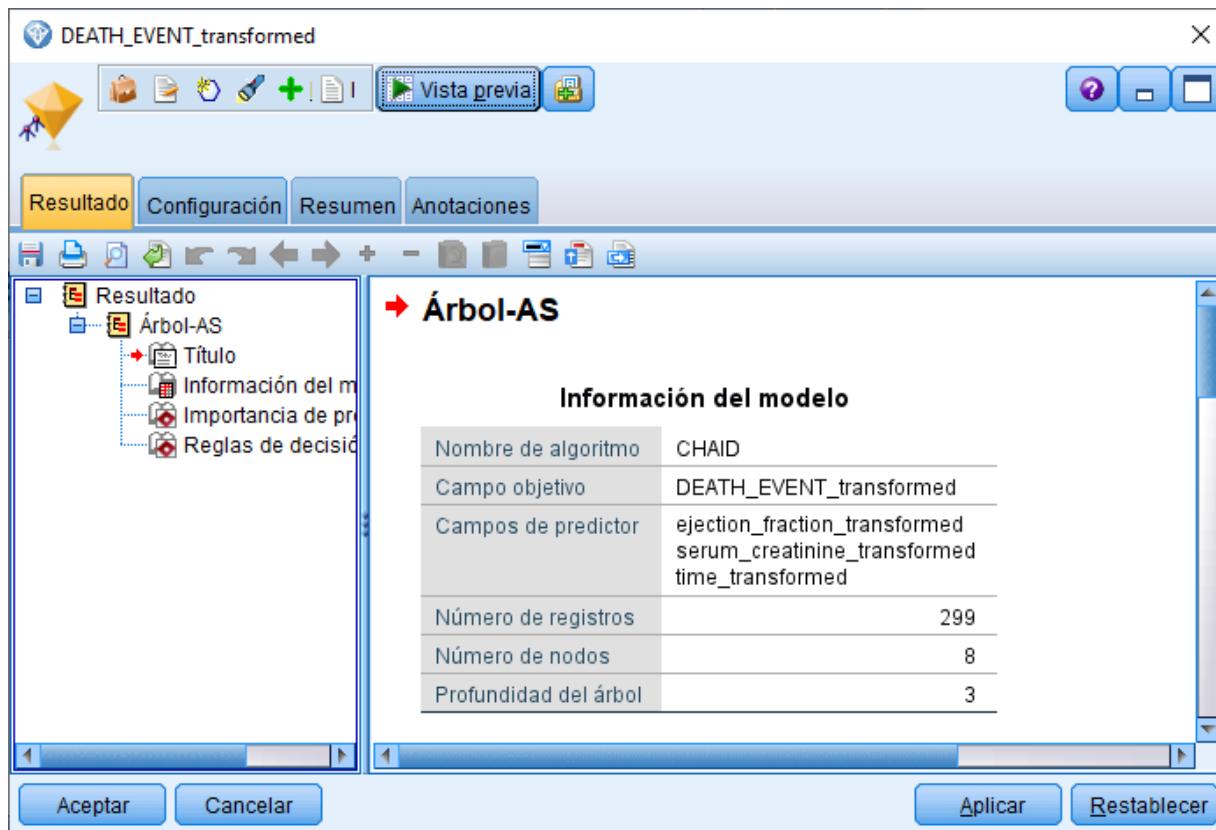


Figura 4.58: Ejecución del modelo Árbol de decisión en Analytic Server - 1.

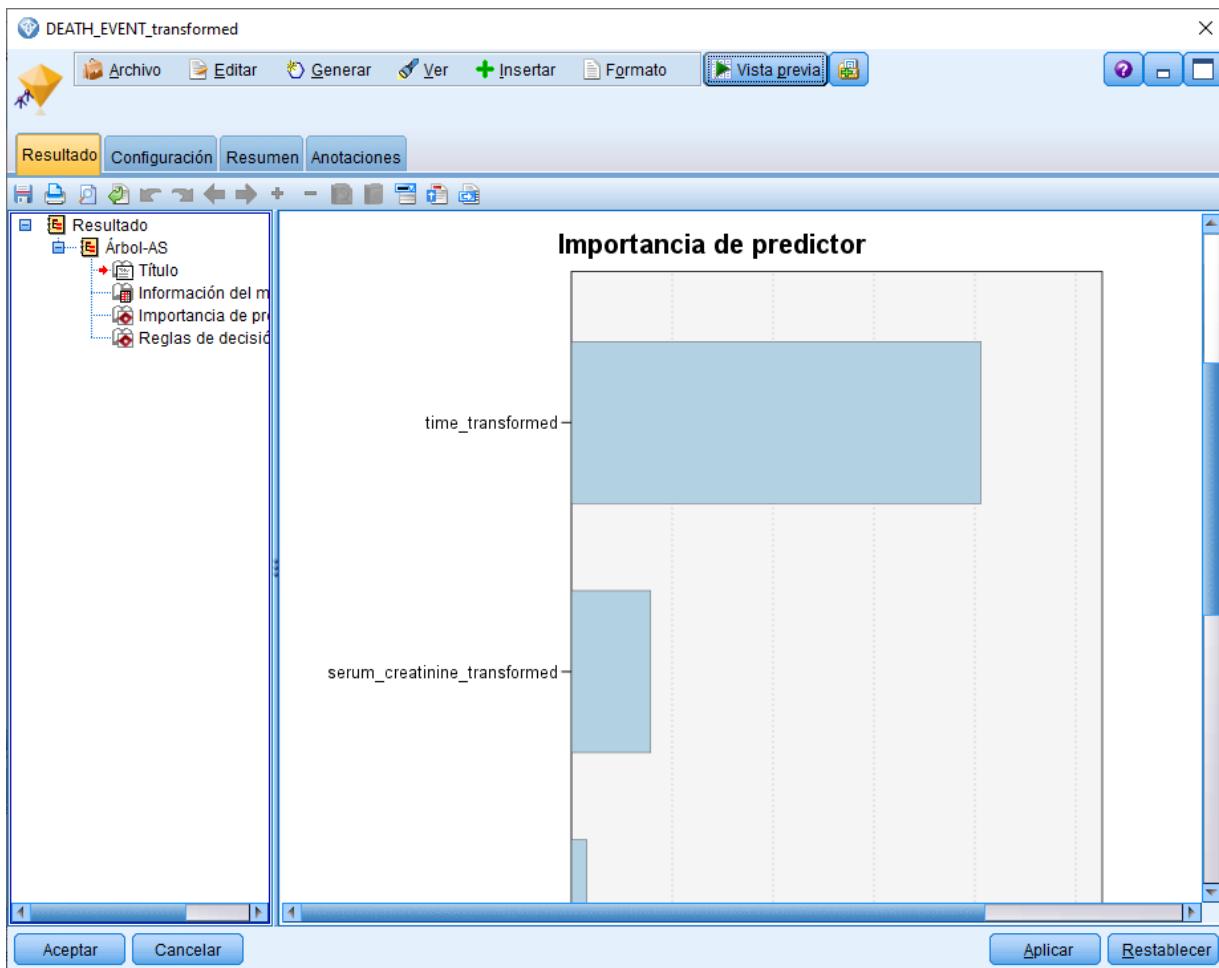


Figura 4.59: Ejecución del modelo Árbol de decisión en Analytic Server - 2.

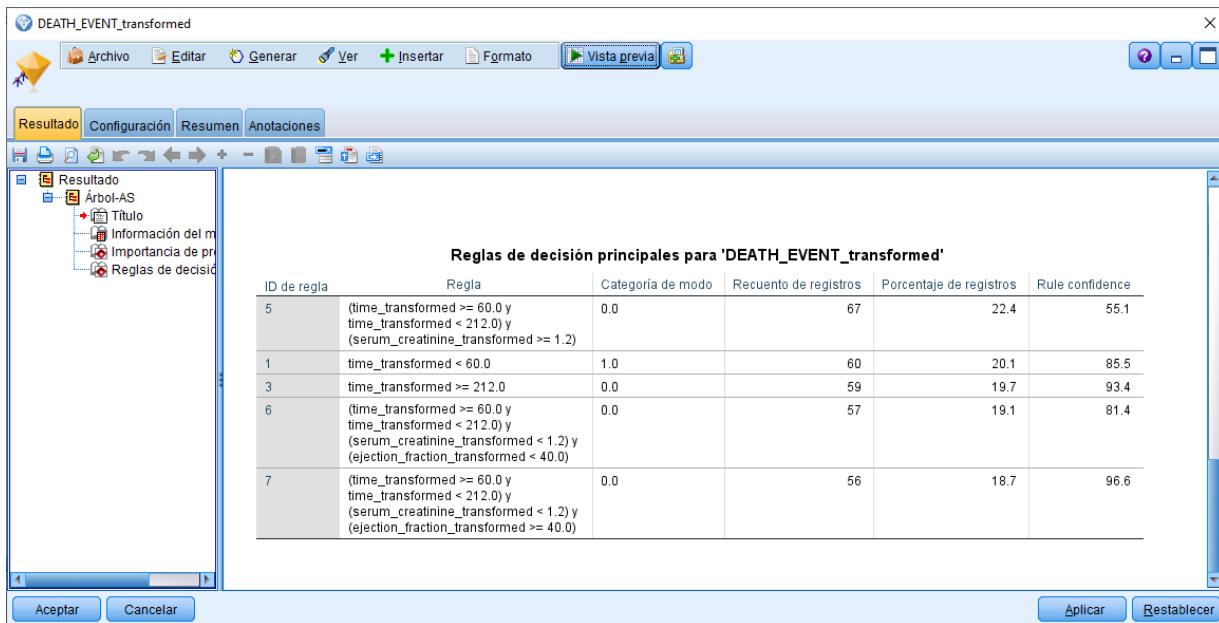


Figura 4.60: Ejecución del modelo Árbol de decisión en Analytic Server - 3.

Por otro lado, la ejecución del árbol de decisión en Analytic Server nos muestra el nombre del algoritmo utilizado, la importancia del predictor y las reglas de decisión para la determinación del campo objetivo; esto es posible apreciarlo tal y como lo muestran las figuras 4.58, 4.59 y 4.60.

4.4.9. Evaluación

En las siguientes figuras, se mostrará brevemente el análisis de la tasa de error del modelo con el fin de comparar el resultado de cada una.

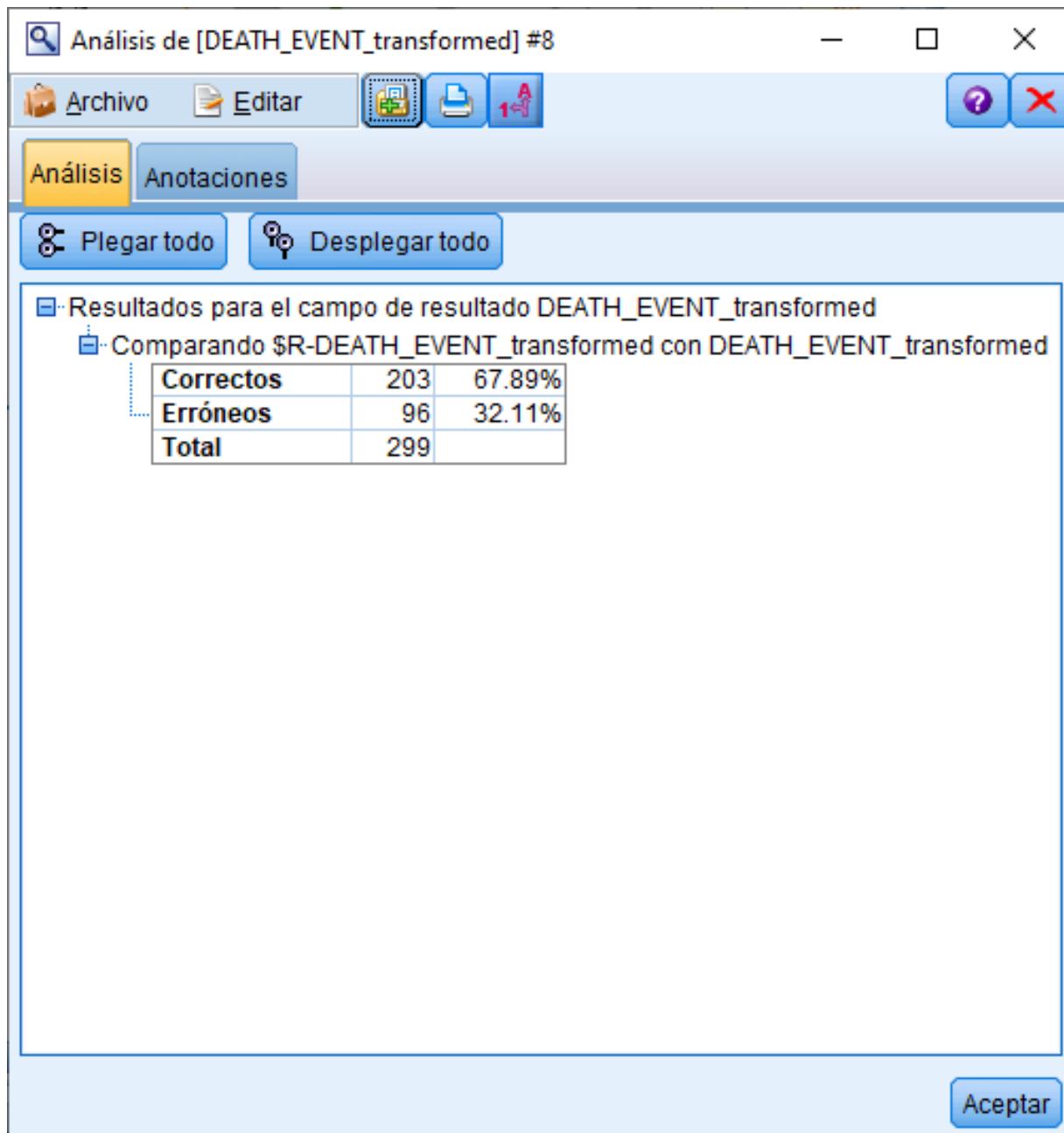


Figura 4.61: Evaluación del modelo Árbol de decisión C&R.

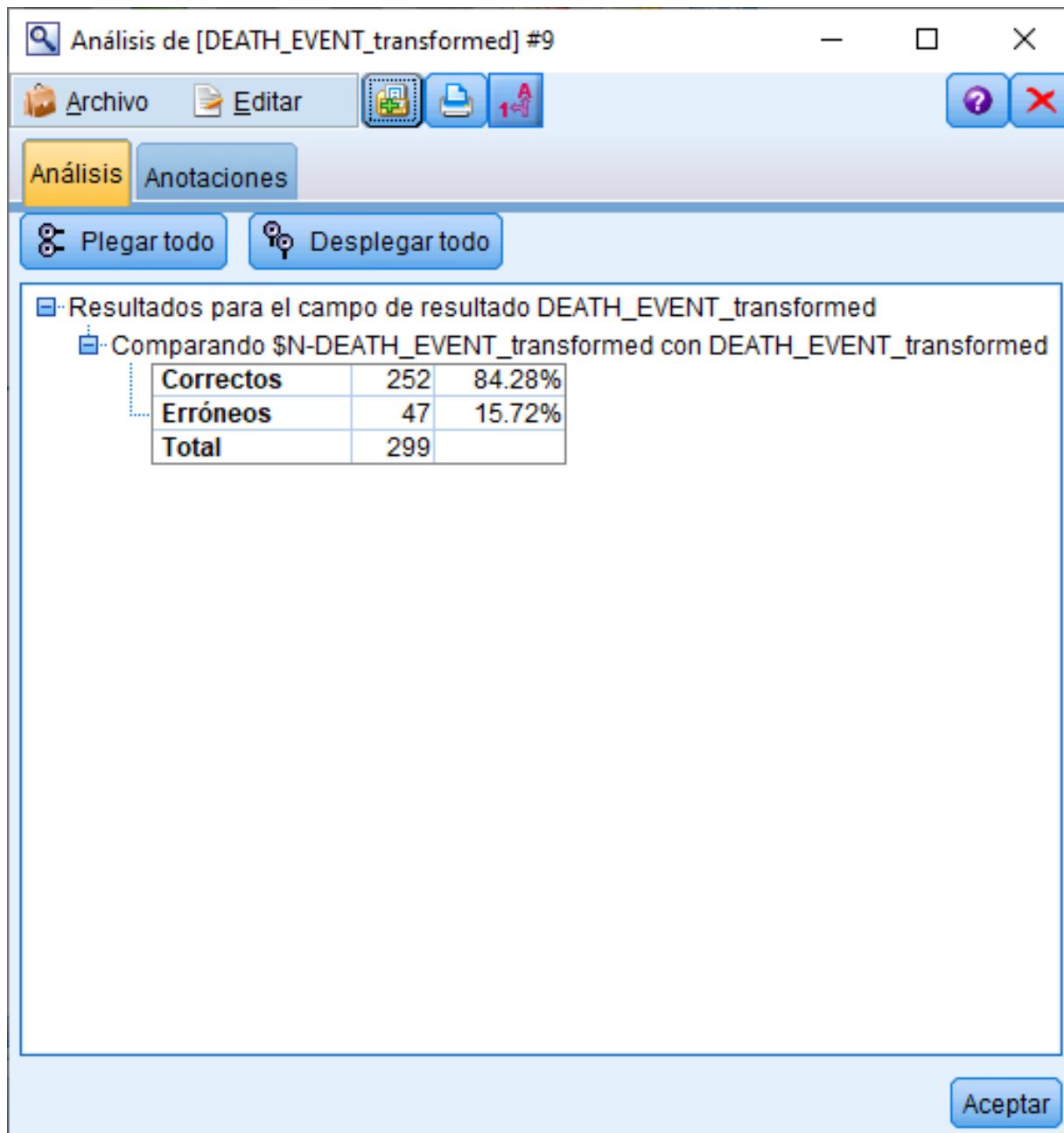


Figura 4.62: Evaluación del modelo Red Neuronal Perceptron Backpropagation.

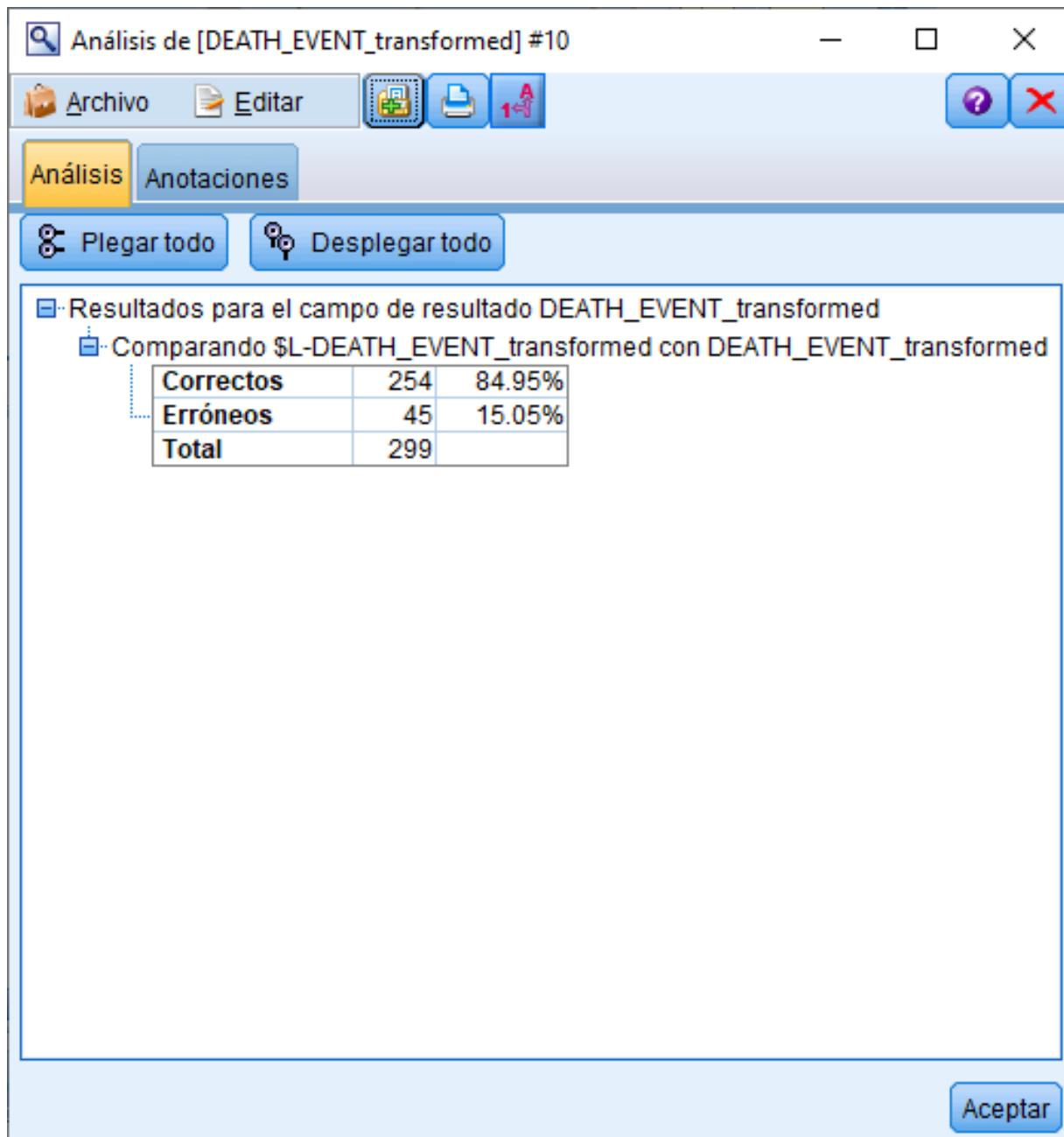


Figura 4.63: Evaluación del modelo Regresión logística.

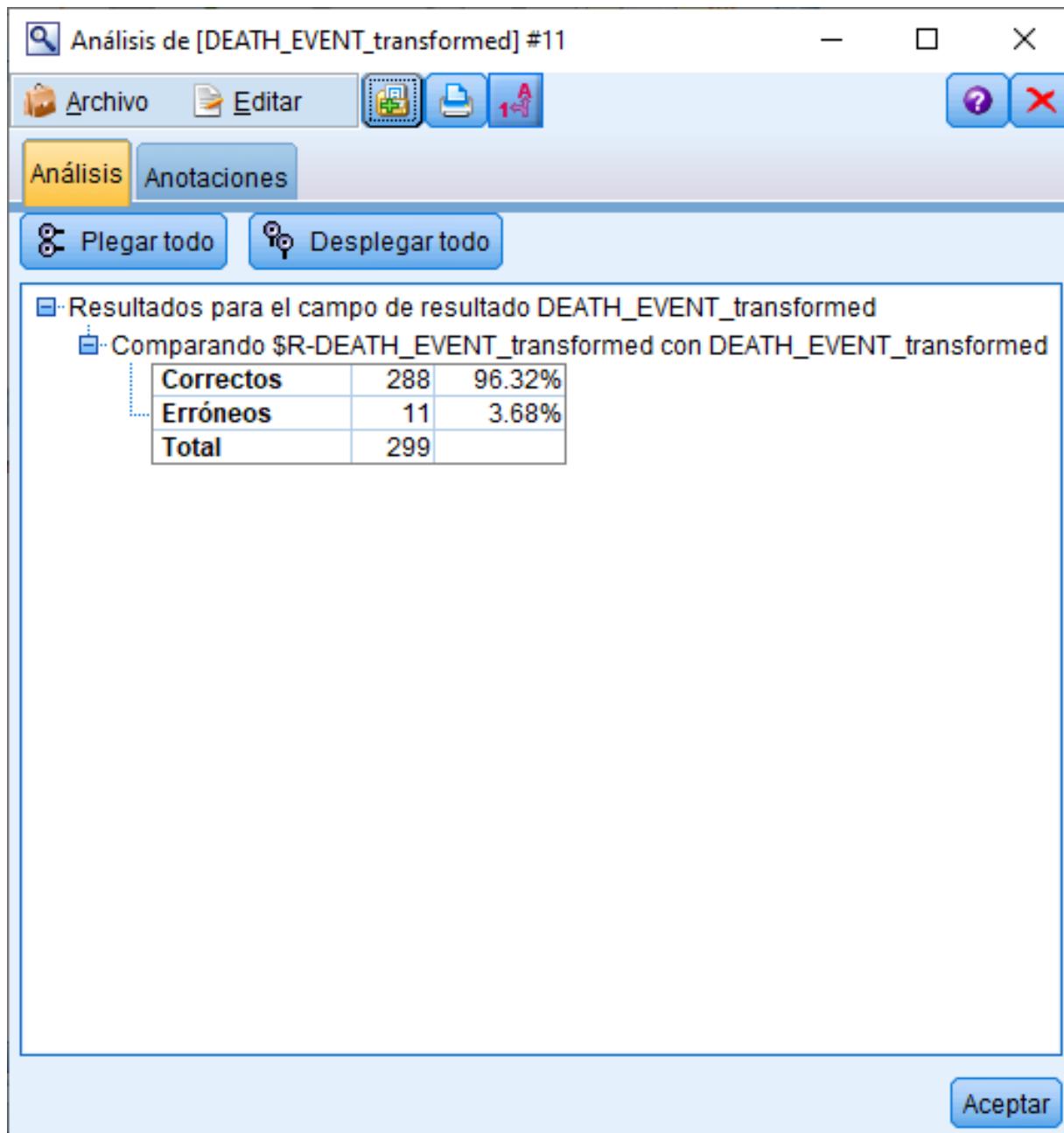


Figura 4.64: Evaluación del modelo Árbol aleatorio.

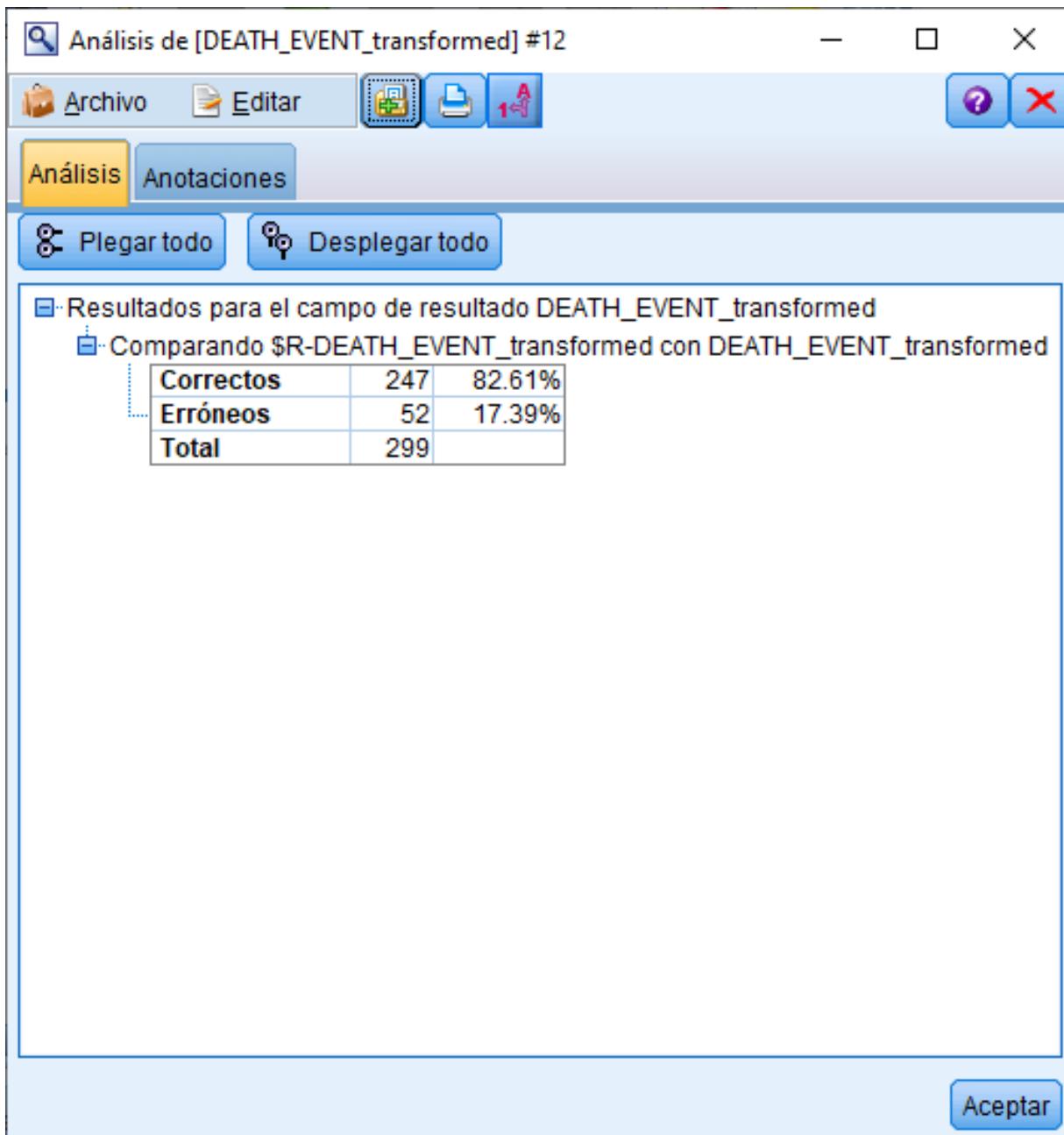


Figura 4.65: Evaluación del modelo Árbol de decisión en Analytic Server.

Como observamos en la figura 4.64, la evaluación del árbol de decisión obtuvo la mejor tasa de error, un 28.43 % más que la evaluación del Árbol de decisión C&R que se encuentra en la figura 4.61, el cual obtuvo la peor tasa de error.

Para comprobar el funcionamiento de cada modelo, se creó una nueva ruta en la herramienta *IBM SPSS Modeler* que nos permitiera ingresar datos a cada modelo y estos comenzarán a realizar predicciones, esto se puede apreciar en la siguiente figura.

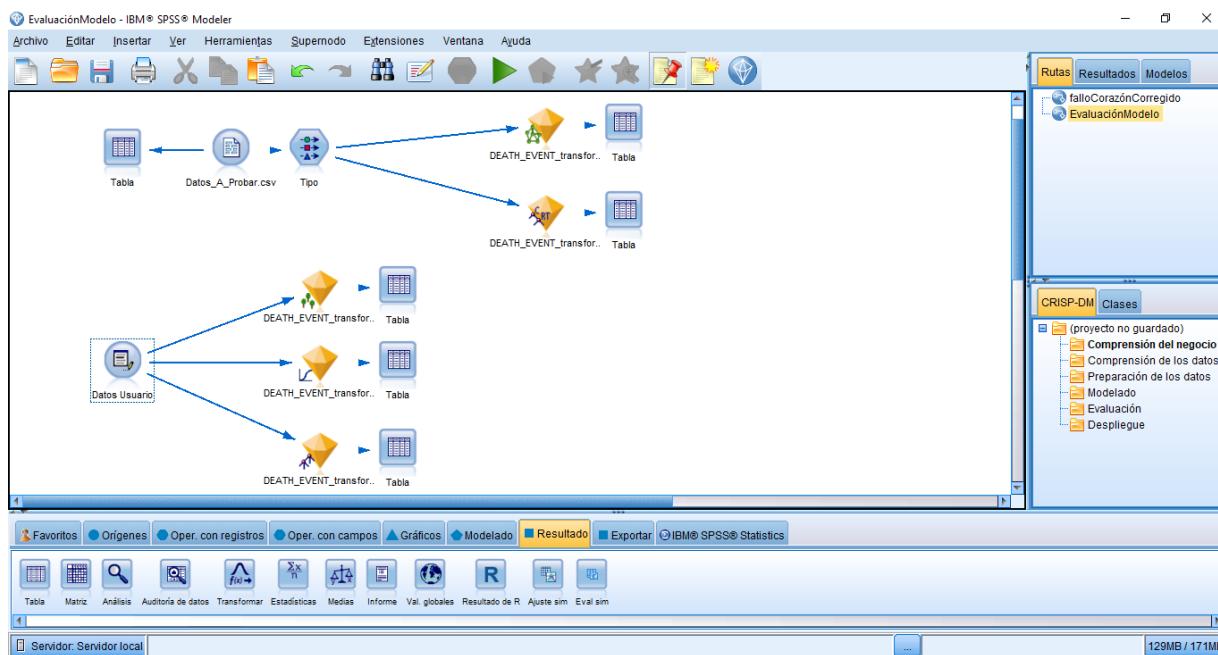


Figura 4.66: Ruta para evaluar los modelos en la herramienta IBM SPSS Modeler.

Dentro de la figura 4.66, vemos que hacemos uso de dos funcionalidades para probar los modelos, la primera es cargar datos mediante un archivo de tipo ".csv" y la segunda mediante el ingreso manual de los datos. A continuación, se mostrará el proceso para utilizar cada funcionalidad con las respectivas predicciones de los modelos.

The screenshot shows the 'Tabla' (Table) view in IBM SPSS Modeler. It displays a data grid with 5 rows and 12 columns. The columns are labeled: 'anaemia_transformed', 'diabetes_transformed', 'high_blood_pressure_transformed', 'sex_transformed', 'smoking_transformed', 'age_transformed', 'creatinine_phosphokinase_transformed', 'ejection_fraction_transformed', 'platelets_transformed', and two unnamed columns at the end. Row 1 is highlighted in yellow. The bottom right corner of the table view has a 'Aceptar' (Accept) button.

	anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transformed	smoking_transformed	age_transformed	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed		
1	0	0	0	1	1	85	1909.003	35	2		
2	0	1	0	1	1	65	224.000	50	1		
3	0	0	0	1	1	69	582.000	20	2		
4	1	0	0	1	1	60	47.000	20	2		
5	0	0	1	0	1	70	92.000	60	3		

Figura 4.67: Registros seleccionados para comprobar el funcionamiento de los modelos.

Tal y como se muestra en la figura 4.67, se importó un archivo que contiene cinco registros del data set con el que se trabajó a lo largo del proyecto. Es importante preservar el mismo nombre y orden de los campos para que cada modelo cumpla con su funcionamiento.

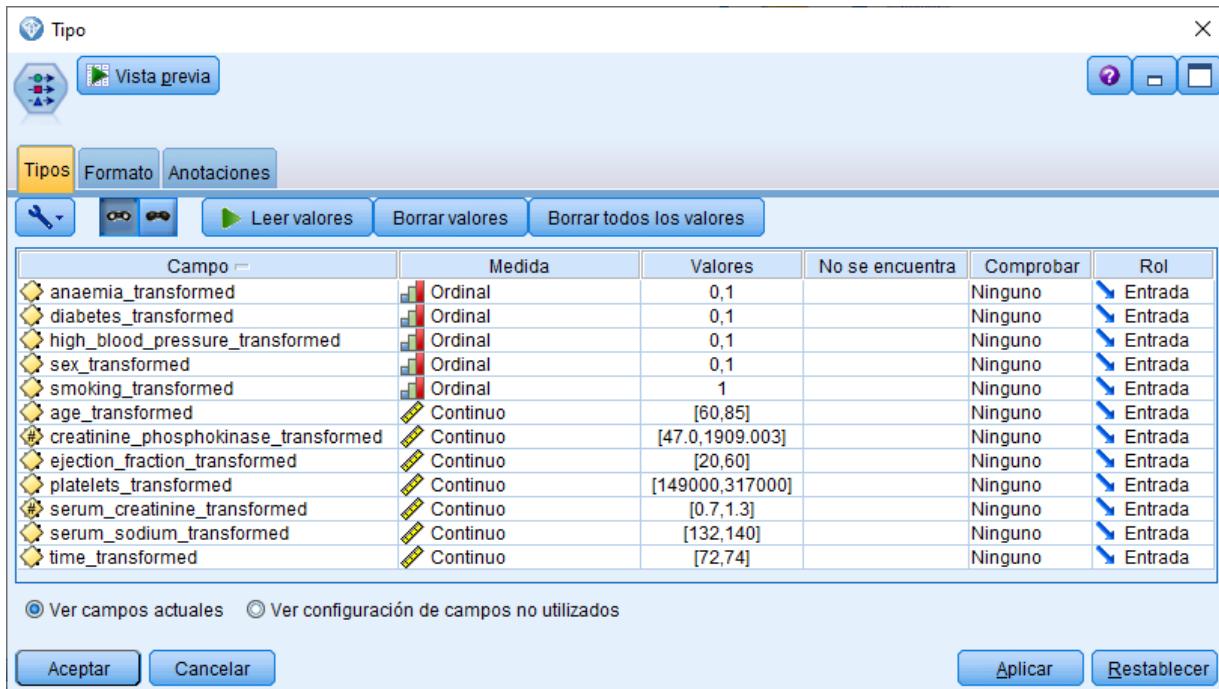


Figura 4.68: Cambio de la medida de los campos.

Posteriormente en la figura 4.68 nos encargamos de cambiar la medida de cada campo, esto porque se deben preservar las medidas con las que se crearon los modelos.

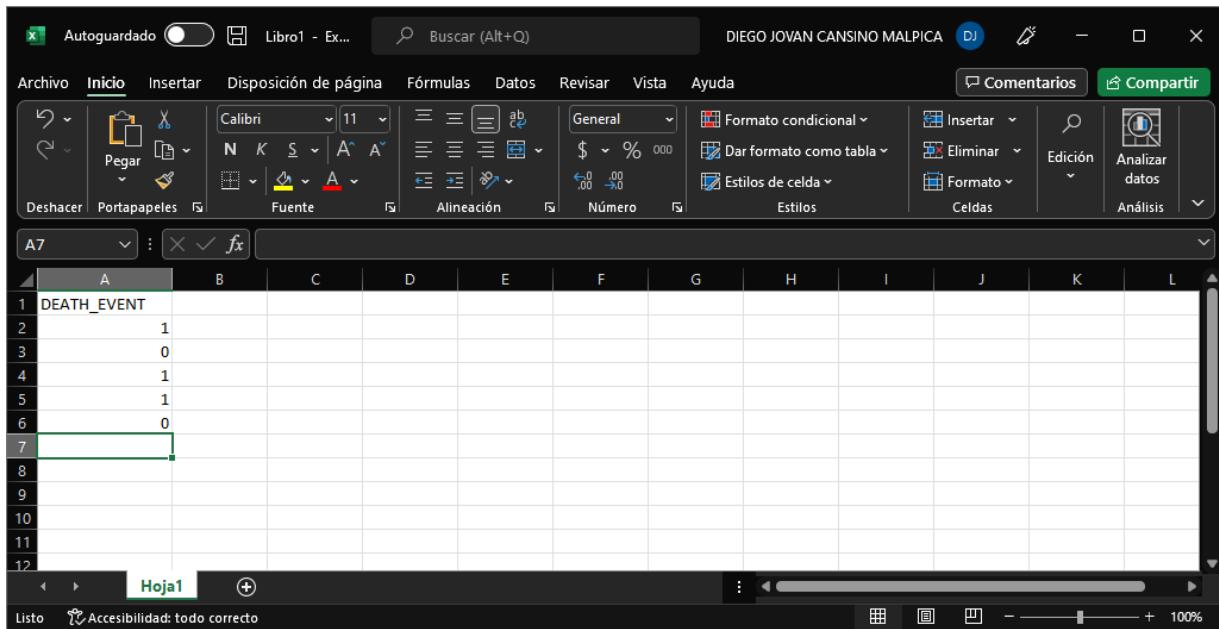


Figura 4.69: Valores destino de los registros seleccionados.

Para verificar las predicciones de los modelos, en la figura 4.69 podemos apreciar la salida que debe tener cada registro que ingresamos a la herramienta *IBM SPSS Modeler*.

	id	med	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed	serum_creatinine_transformed	serum_sodium_transformed	time_transformed	\$N-DEATH_EVENT_transformed
1	85		1909.003	35	243000	1.000	132	72	1
2	65		224.000	50	149000	1.300	137	72	0
3	69		582.000	20	266000	1.200	134	73	1
4	60		47.000	20	204000	0.700	139	73	1
5	70		92.000	60	317000	0.800	140	74	0

Figura 4.70: Predicción del modelo Red Neuronal Perceptron Backpropagation.

La figura 4.70 muestra la predicción que realizó el modelo Red Neuronal Perceptron Backpropagation, el cual a pesar de tener cerca del 84 % de efectividad en su evaluación, acertó la predicción de los cinco registros.

	id	med	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed	serum_creatinine_transformed	serum_sodium_transformed	time_transformed	\$R-DEATH_EVENT_transformed
1	85		1909.003	35	243000	1.000	132	72	0
2	65		224.000	50	149000	1.300	137	72	0
3	69		582.000	20	266000	1.200	134	73	0
4	60		47.000	20	204000	0.700	139	73	0
5	70		92.000	60	317000	0.800	140	74	0

Figura 4.71: Predicción del modelo Árbol de decisión C&R.

En contraste, la figura 4.70 muestra la predicción que realizó el modelo Árbol de decisión C&R, el cual solamente acertó dos de los cinco registros a predecir.

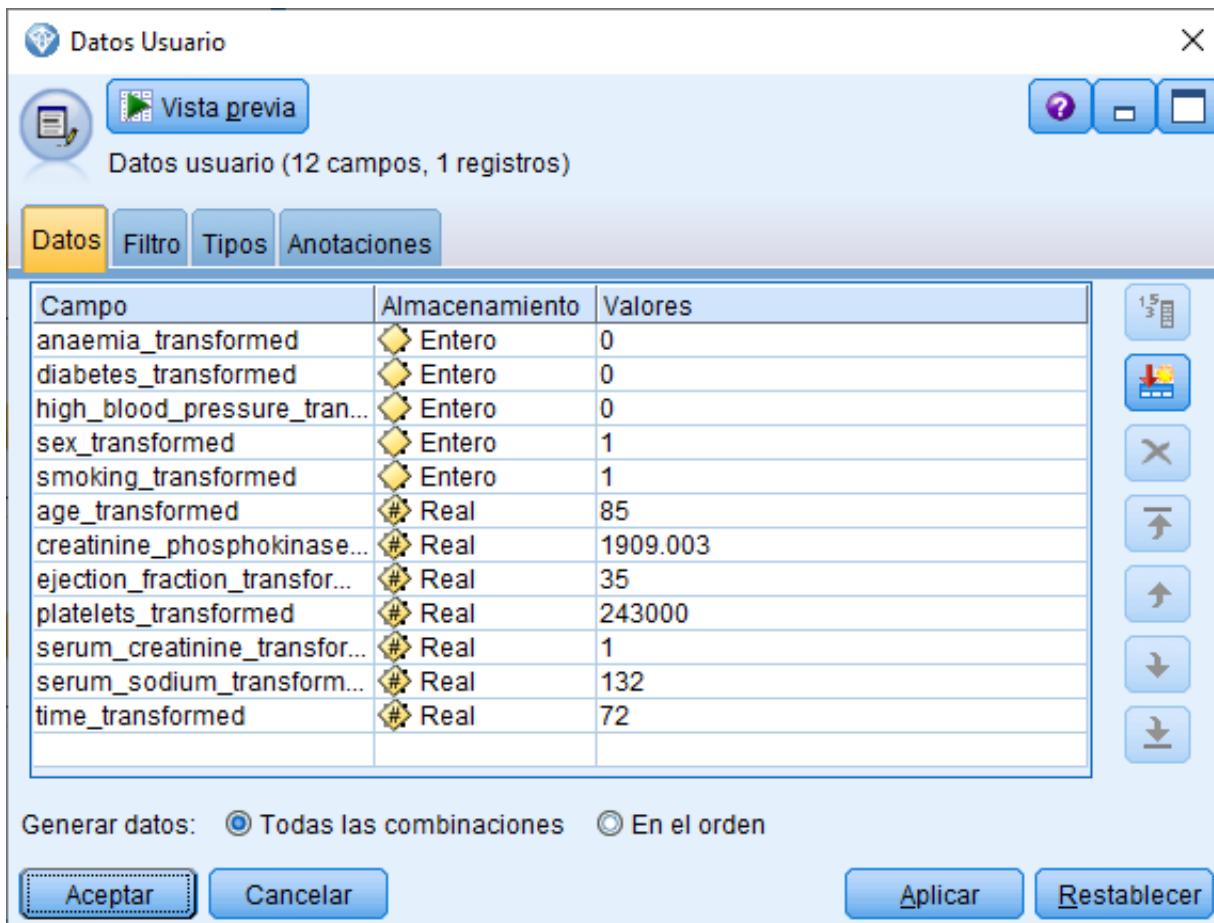


Figura 4.72: Configuración ingresada en la funcionalidad 'Datos Usuario'.

La segunda funcionalidad para realizar predicciones con los modelos es mediante 'Datos Usuario', donde debemos ingresar cada uno de los datos en el mismo orden que el data set original, con el mismo nombre de los campos y con el mismo tipo de almacenamiento. En este caso, en la figura 4.72 se ingresaron los datos del primer registro que se aprecia en la figura 4.67.

hed	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed	serum_creatinine_transformed	serum_sodium_transformed	time_transformed	\$R-DEATH_EVENT_transformed
1	000	1909.003	35.000	243000.000	1.000	132.000	72.000

Figura 4.73: Predicción del modelo Árbol aleatorio.

La figura 4.73 muestra la predicción que realizó el modelo Árbol aleatorio, el cual acertó con la predicción del registro.

	med	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed	serum_creatinine_transformed	serum_sodium_transformed	time_transformed	\$L-DEATH_EVENT_transformed
1	5.000	1909.003	35.000	243000.000	1.000	132.000	72.000	1

Figura 4.74: Predicción del modelo Regresión logística.

Por otra parte, en la figura 4.74 vemos la predicción que realizó el modelo Regresión logística, el cual acertó con la predicción del registro.

	med	creatinine_phosphokinase_transformed	ejection_fraction_transformed	platelets_transformed	serum_creatinine_transformed	serum_sodium_transformed	time_transformed	\$R-DEATH_EVENT_transformed
1	0.000	1909.003	35.000	243000.000	1.000	132.000	72.000	0

Figura 4.75: Predicción del modelo Árbol de decisión en Analytic Server.

Finalmente, dentro de la figura 4.75 se comprueba que el modelo Árbol de decisión en Analytic Server fue el que peor desempeño obtuvo para este caso de estudio, pues no acertó con la predicción esperada.

4.5. Caso de estudio - "Breast Cancer"

Para comprobar los resultados que se obtuvieron en la subsección 4.4.9, se decidió seleccionar un nuevo data set que me permitiera utilizar los modelos:

- Red Neuronal Perceptron Backpropagation.
- Árbol de decisión aleatorio.
- Regresión logística.

El data set seleccionado fue extraído del repositorio [UCI Machine Learning](#), el cual incluye 286 instancias, donde cada instancia se describen mediante 9 atributos, algunos de los cuales son lineales y otros son nominales.

El propósito de este data set es predecir si se debe irradiar un tumor de cáncer de mama o no.

La ruta creada dentro de la herramienta *IBM SPSS Modeler* se encuentra en la siguiente figura.

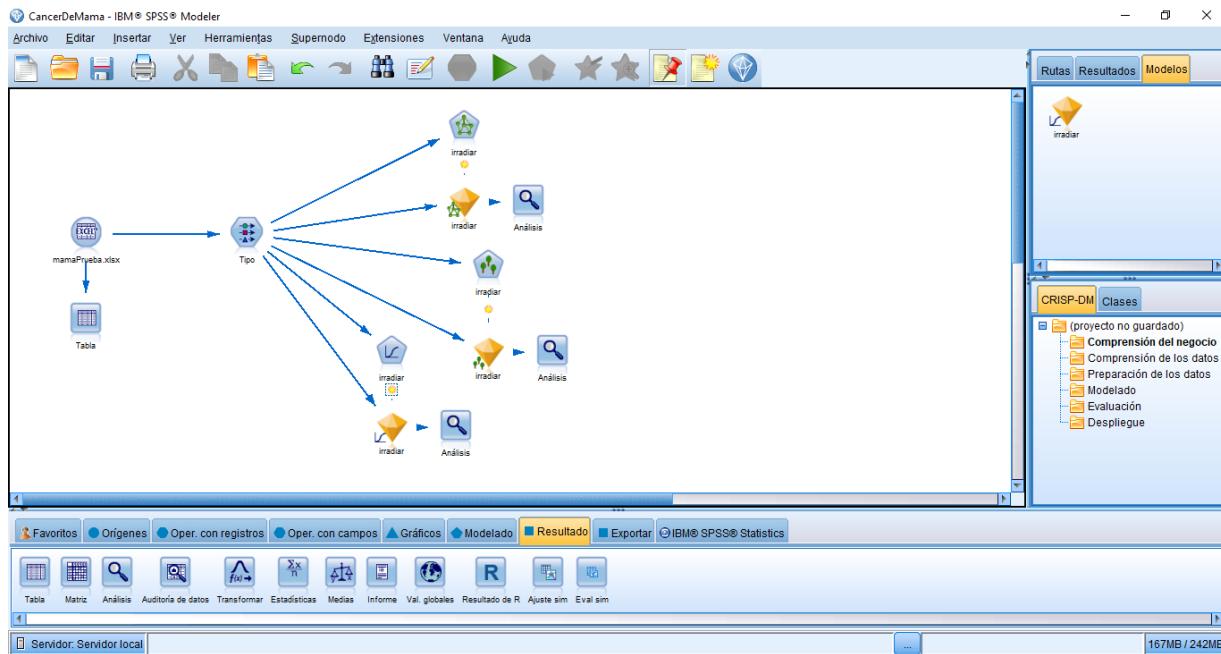


Figura 4.76: Ruta para evaluar los modelos con el data set "Breast Cancer".

En la figura 4.76 podemos apreciar que para la creación de los modelos no hicimos una limpieza de los datos, derivación de nuevos campos u otras funcionalidades debido a que se tiene como propósito comprobar la evaluación de los modelos, no mejorar la precisión de cada uno.

Tabla (10 campos, 286 registros) #2

Archivo Editar Generar

Clase edad menopausia tamaño del tumor inv-nodos tapas de nodos deg-malig mama cuádriceps mamario irradiar

1 no-recurrence-events 30-39 premeno 30-34 0-2 no 3.000 left left_low no
 2 no-recurrence-events 40-49 premeno 20-24 0-2 no 2.000 right right_up no
 3 no-recurrence-events 40-49 premeno 20-24 0-2 no 2.000 left left_low no
 4 no-recurrence-events 60-69 ge40 15-19 0-2 no 2.000 right left_up no
 5 no-recurrence-events 40-49 premeno 0-4 0-2 no 2.000 right right_low no
 6 no-recurrence-events 60-69 ge40 15-19 0-2 no 2.000 left left_low no
 7 no-recurrence-events 50-59 premeno 25-29 0-2 no 2.000 left left_low no
 8 no-recurrence-events 60-69 ge40 20-24 0-2 no 1.000 left left_low no
 9 no-recurrence-events 40-49 premeno 50-54 0-2 no 2.000 left left_low no
 10 no-recurrence-events 40-49 premeno 20-24 0-2 no 2.000 right left_up no
 11 no-recurrence-events 40-49 premeno 0-4 0-2 no 3.000 right central no
 12 no-recurrence-events 50-59 ge40 25-29 0-2 no 2.000 left left_low no
 13 no-recurrence-events 60-69 lt40 10-14 0-2 no 1.000 left right_up no
 14 no-recurrence-events 50-59 ge40 25-29 0-2 no 3.000 left right_up no
 15 no-recurrence-events 40-49 premeno 30-34 0-2 no 3.000 left left_up no
 16 no-recurrence-events 60-69 lt40 30-34 0-2 no 1.000 left left_low no
 17 no-recurrence-events 40-49 premeno 15-19 0-2 no 2.000 left left_low no
 18 no-recurrence-events 50-59 premeno 30-34 0-2 no 3.000 left left_low no
 19 no-recurrence-events 60-69 ge40 30-34 0-2 no 3.000 left left_low no
 20 no-recurrence-events 50-59 ge40 30-34 0-2 no 1.000 right right_up no
 21 no-recurrence-events 50-59 ge40 40-44 0-2 no 2.000 left left_low no
 22 no-recurrence-events 60-69 ge40 15-19 0-2 no 2.000 left left_low no
 23 no-recurrence-events 30-34 premeno 25-29 0-2 no 2.000 right left_low no
 24 no-recurrence-events 50-59 premeno 40-44 0-2 no 2.000 left left_up no
 25 no-recurrence-events 50-59 premeno 35-39 0-2 no 2.000 right left_up no
 26 no-recurrence-events 40-49 premeno 25-29 0-2 no 2.000 left left_up no
 27 no-recurrence-events 50-59 premeno 20-24 0-2 no 1.000 left left_low no
 28 no-recurrence-events 60-69 ge40 25-29 0-2 no 3.000 right left_up no
 29 no-recurrence-events 40-49 premeno 40-44 0-2 no 2.000 right left_low no
 30 no-recurrence-events 60-69 ge40 30-34 0-2 no 2.000 left left_low no
 31 no-recurrence-events 50-59 ge40 40-44 0-2 no 3.000 right left_up no
 32 no-recurrence-events 50-59 premeno 15-19 0-2 no 2.000 right left_low no
 33 no-recurrence-events 50-59 premeno 10-14 0-2 no 3.000 left left_low no
 34 no-recurrence-events 50-59 ge40 10-14 0-2 no 1.000 right left_up no
 35 no-recurrence-events 50-59 ge40 10-14 0-2 no 1.000 left left_up no
 36 no-recurrence-events 30-39 premeno 30-34 0-2 no 2.000 left left_up no

Aceptar

Figura 4.77: Visualización del data set "Breast Cancer".

Tipo Vista previa

Tipos Formato Anotaciones

Campo Medida Valores / No se encuentra Comprobar Rol

Clase	Marca	recurrence-events/no-recurrence-events	Ninguno	Entrada
edad	Nominal	"20-29";"30-39";"40-49";"50-59";"60-69";"70-79"	Ninguno	Entrada
menopausia	Nominal	ge40,lt40,premeno	Ninguno	Entrada
tamaño del tumor	Nominal	"0-4";"10-14";"15-19";"20-24";"25-29";"30-34";"35-39";"40-44";"45-49";"5-9";"50-54"	Ninguno	Entrada
inv-nodos	Nominal	"0-2";"12-14";"15-17";"24-26";"3-5";"6-8";"9-11"	Ninguno	Entrada
tapas de nodos	Nominal	"?";no,yes	Ninguno	Entrada
deg-malig	Continuo	[1.0;3.0]	Ninguno	Entrada
mama	Marca	right:left	Ninguno	Entrada
cuádriceps mamario	Nominal	"?",central,left_low,left_up,right_low,right_up	Ninguno	Entrada
irradiar	Marca	yes/no	Ninguno	Destino

Ver campos actuales Ver configuración de campos no utilizados

Aceptar Cancelar Aplicar Restablecer

Figura 4.78: Medidas y valores de los campos del data set "Breast Cancer".

Las figuras 4.77 y 4.78 nos muestran la tabla del data set dentro de la herramienta *IBM SPSS Modeler* al igual que las medidas y valores de los campos. Es importante destacar que se ha cambiado el rol del campo "irradiar", pues será el campo a predecir.

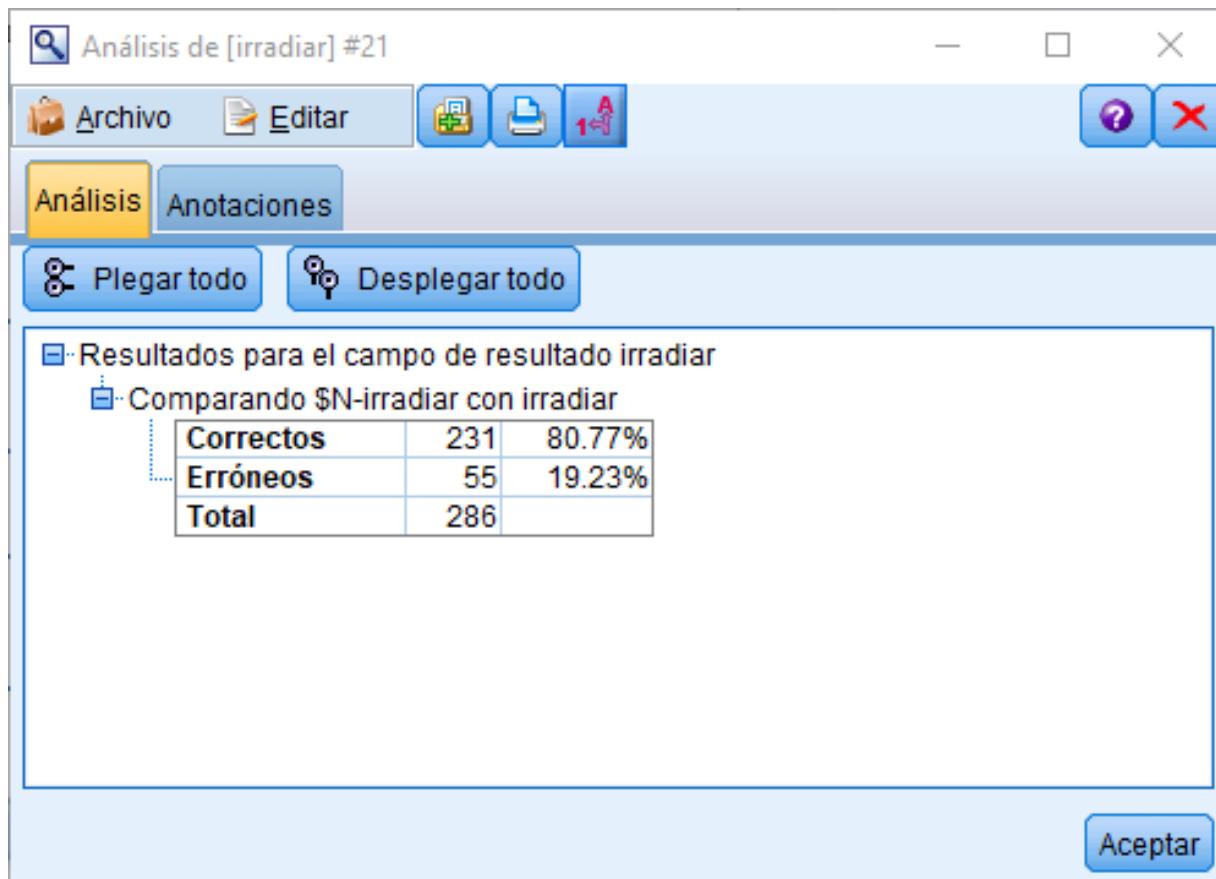


Figura 4.79: Evaluación del modelo Red Neuronal Perceptron Backpropagation.

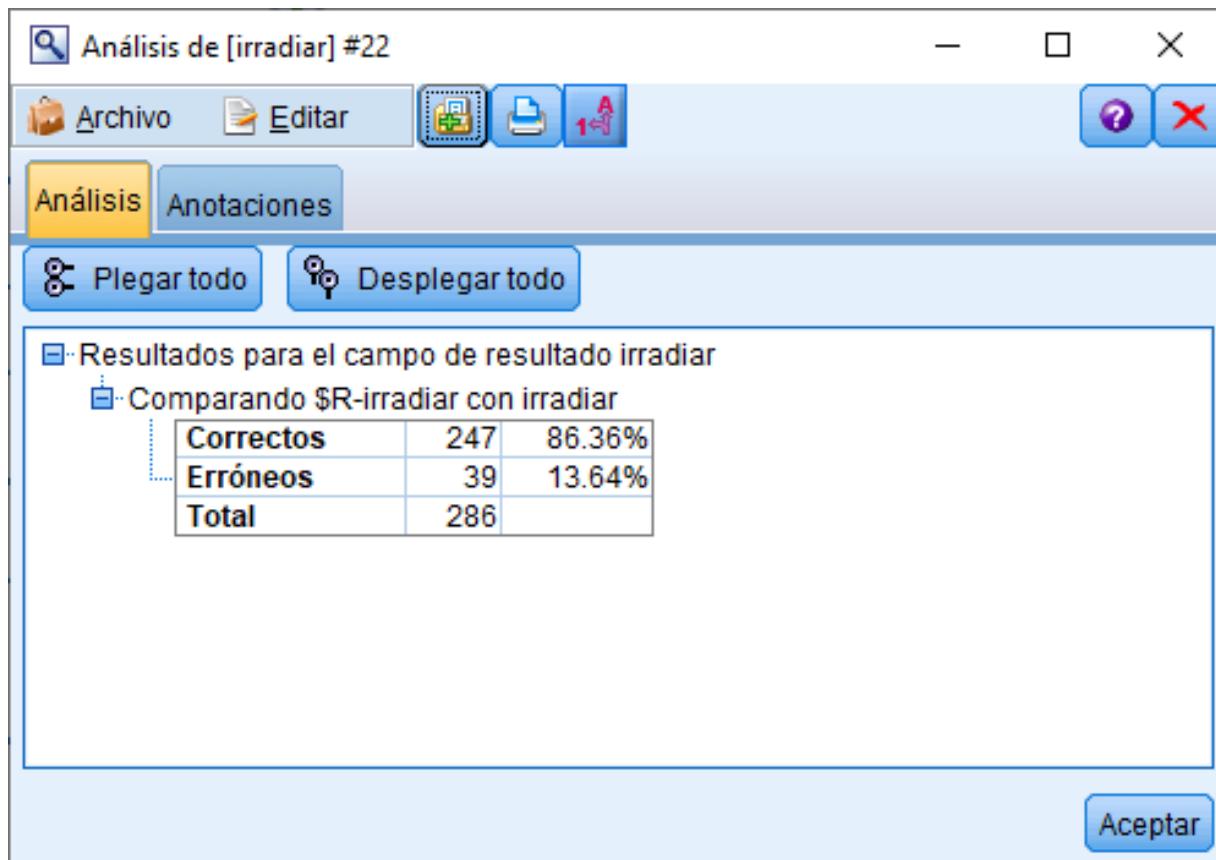


Figura 4.80: Evaluación del modelo Árbol aleatorio.

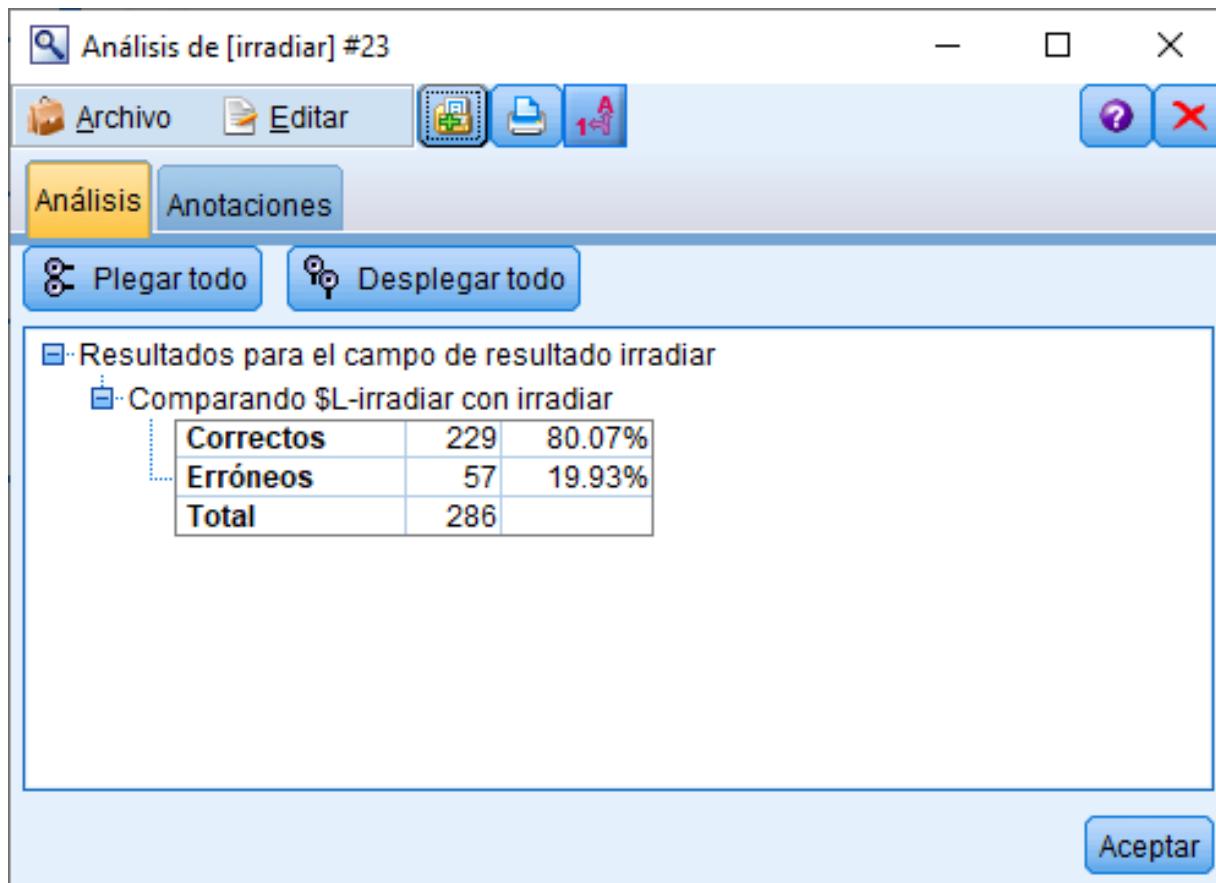


Figura 4.81: Evaluación del modelo Regresión logística.

En las figuras 4.79, 4.80 y 4.81 podemos comprobar que, nuevamente el modelo Árbol aleatorio ha sido el que mejores resultados obtuvo.

Capítulo 5

Herramienta Intelligent Data Analysis Tool

5.1. Descripción de la herramienta

A lo largo de este trabajo, se creó la herramienta Intelligent Data Analysis Tool. Esta herramienta fue diseñada con el propósito de gestionar un data set, realizar la limpieza de los datos contenidos en el data set y mostrar gráficas que nos ayuden a comprender de manera más clara los datos. Además, tiene la capacidad de implementar un modelo de aprendizaje automático en base al data set seleccionado.

Con lo antes mencionado, podemos entender que es una alternativa a la aplicación *IBM SPSS Modeler*, donde tendremos la ventaja de utilizar las configuraciones que se crean pertinentes para obtener buenos resultados.

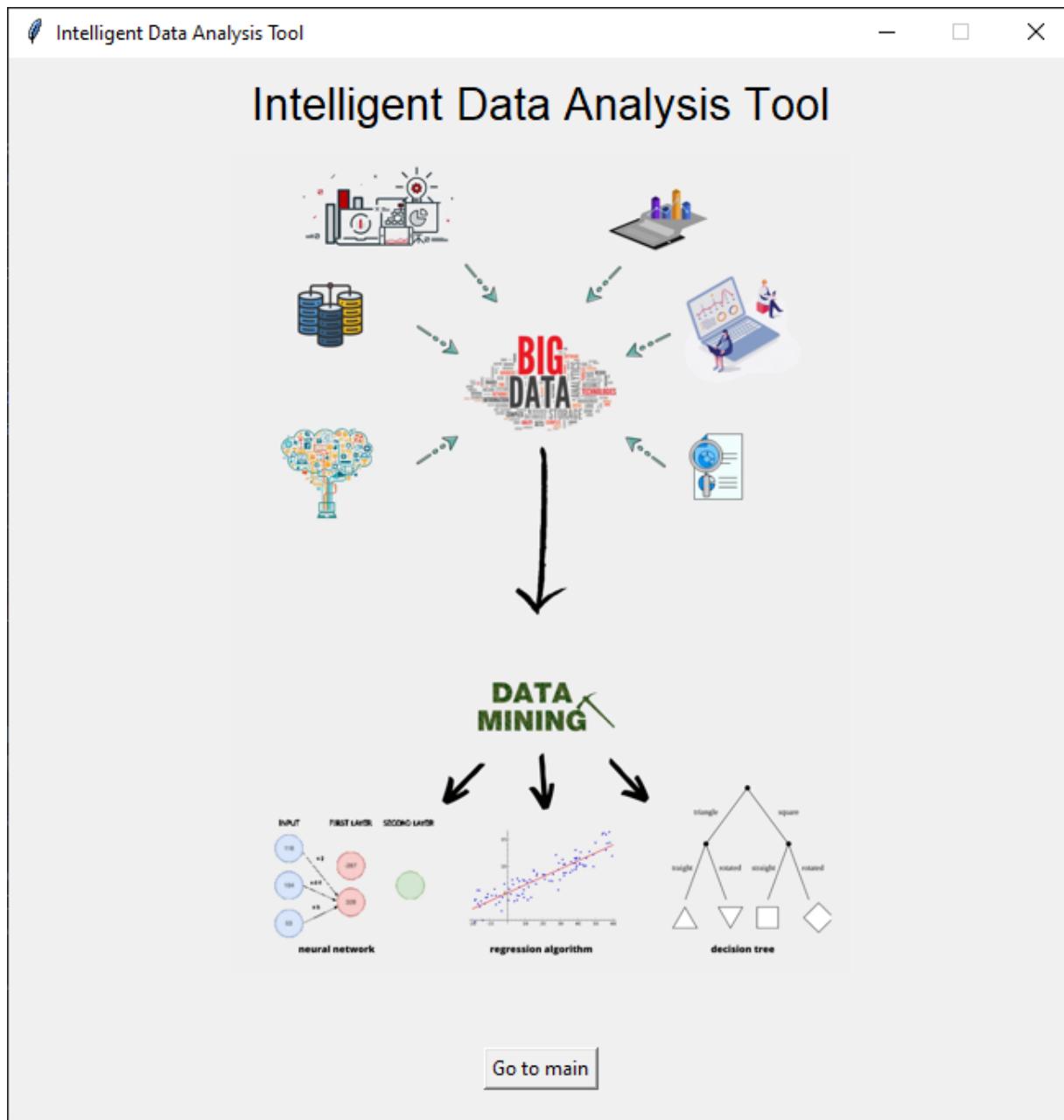


Figura 5.1: Interfaz introductoria de la aplicación.

La herramienta cuenta con una interfaz introductoria en la cual al presionar el botón "Go to main" se podrá acceder a la interfaz principal, esto se puede apreciar en la figura 5.1.

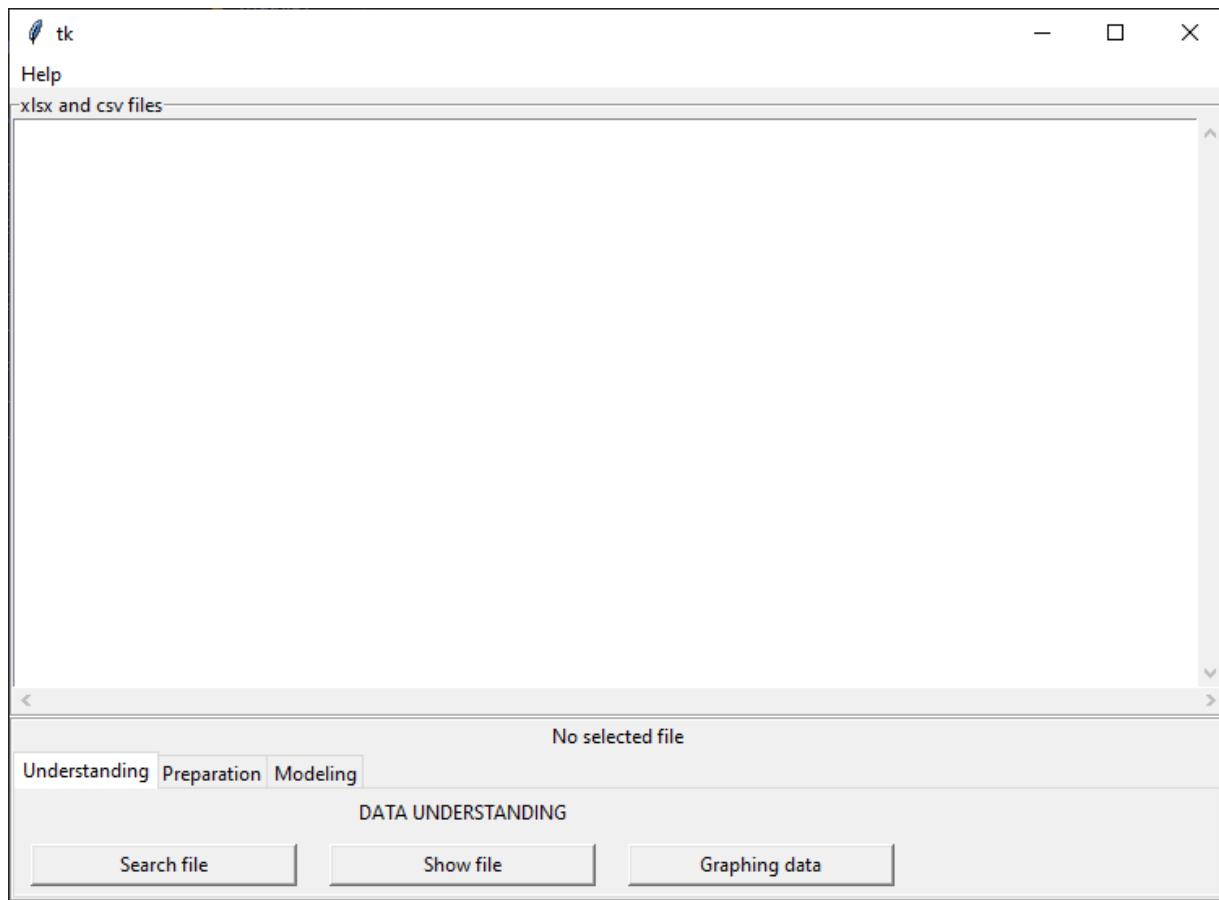


Figura 5.2: Interfaz principal de la aplicación.

Por otra parte, la interfaz principal que se visualiza en la figura 5.2 consta de un visualizador de datos y un menú de pestañas que nos proporcionan distintas funcionalidades.

anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transformed
1.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	1.0	0.0	1.0
0.0	0.0	0.0	0.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	1.0
0.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	0.0
1.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0
1.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0

C:/Users/User/Documents/DH.csv

Understanding Preparation Modeling

DATA UNDERSTANDING

Search file Show file Graphing data

Figura 5.3: "Show file" dentro de la aplicación.

La pestaña "Understanding" cuenta con tres botones, el botón "Search File" es para buscar el data set con el que se desea trabajar, una vez seleccionado se deberá presionar el botón de "Show File" para mostrar los datos, el resultado de esta funcionalidad la vemos en la figura 5.3.

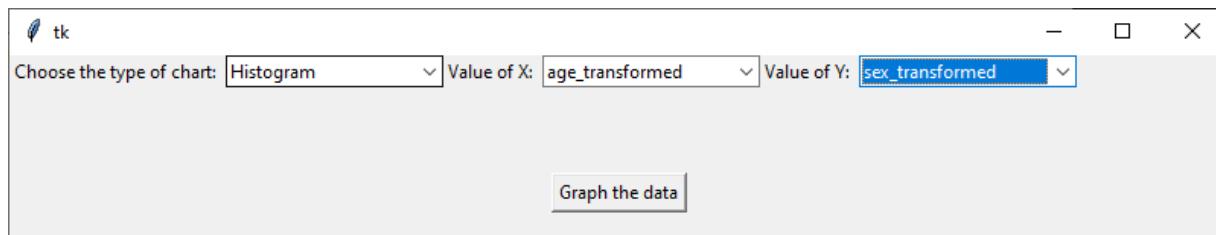


Figura 5.4: Interfaz para realizar gráficos.

Si seleccionamos el botón de "Graphing data", se desplegará la ventana emergente que tenemos en la figura 5.4, la cual nos permite trabajar con distintos tipos de gráficos, así como el ajuste del eje X y Y del gráfico.

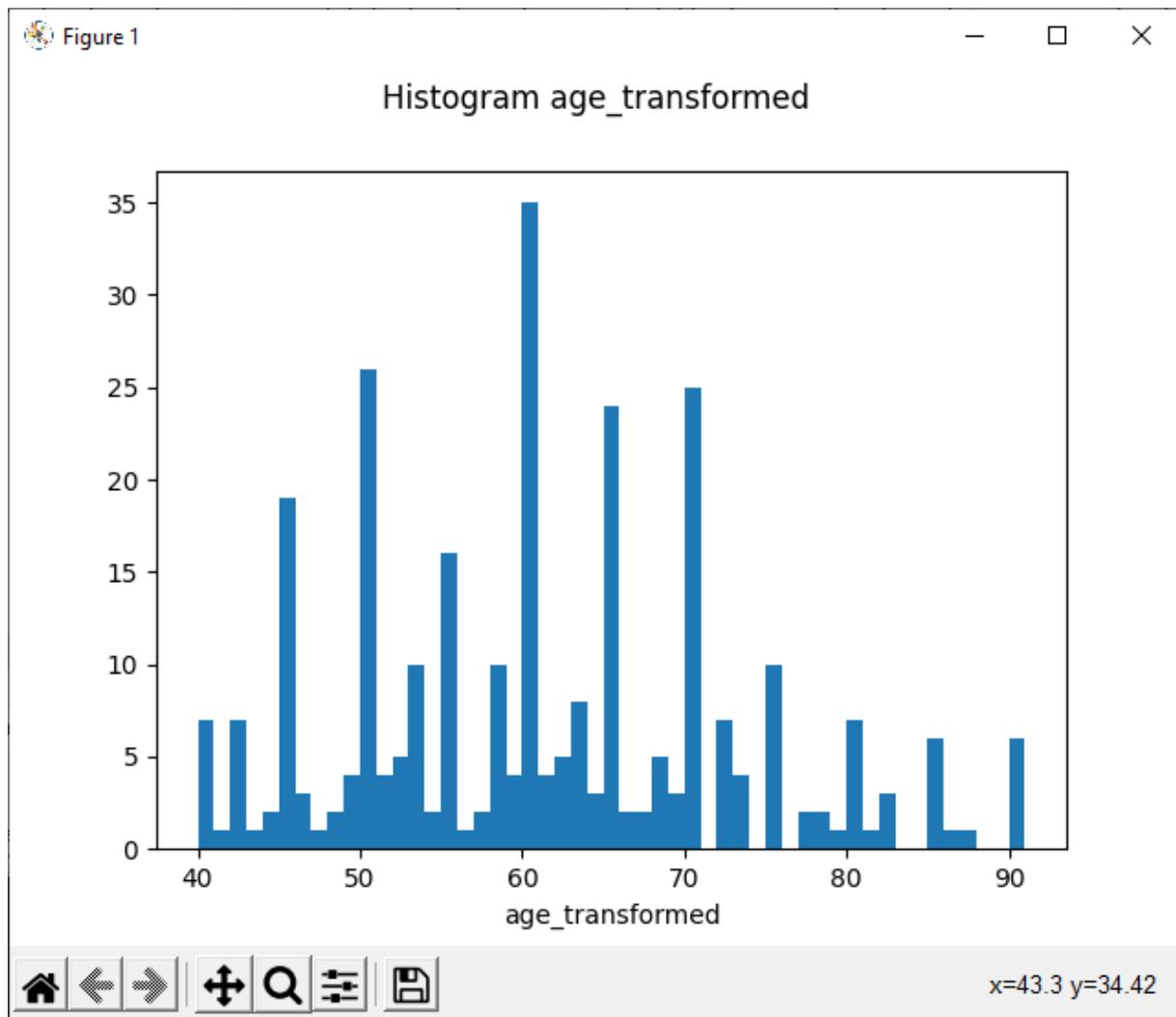


Figura 5.5: Gráfico creado con la aplicación.

Como se muestra en la figura 5.5, al presionar el botón "Graph the data" la aplicación creará el gráfico seleccionado.

The screenshot shows a window titled "tk" with a menu bar containing "Help" and "xlsx and csv files". The main area displays a data grid with four columns: "anaemia_transformed", "diabetes_transformed", "high_blood_pressure_transformed", and "sex_transform". The data consists of 15 rows of binary values (0.0 or 1.0). Below the grid, the file path "C:/Users/User/Documents/DH.csv" is shown. At the bottom, there are four tabs: "Understanding", "Preparation" (which is selected and highlighted in blue), and "Modeling". Under the "Preparation" tab, the heading "DATA PREPARATION" is centered, followed by four buttons: "FFILL", "BFILL", "None", and "All".

anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transform
1.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	1.0	0.0	1.0
0.0	0.0	0.0	0.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	1.0
0.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	0.0
1.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0
1.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0

Figura 5.6: Pestaña "Preparation" de la aplicación.

Continuando con la interfaz principal de la aplicación, la pestaña "Preparation" cuenta con cuatro botones, los cuales son opciones de limpieza de los datos, esto lo podemos visualizar en la figura 5.6.

- **Botón FFILL:** Rellena los campos nulos con el valor del registro siguiente.
- **Botón BFILL:** Rellena los campos nulos con el valor del registro anterior.
- **Botón None:** Si se encuentra datos nulos, elimina todo el registro.
- **All:** Esta opción combina las tres técnicas anteriores.

The screenshot shows a window titled "tk" with a menu bar containing "Help" and "xlsx and csv files". The main area displays a table with four columns: "anaemia_transformed", "diabetes_transformed", "high_blood_pressure_transformed", and "sex_transform". The data consists of 15 rows of binary values (0.0 or 1.0). Below the table, the file path "C:/Users/User/Documents/DH.csv" is shown. At the bottom, there is a navigation bar with tabs: "Understanding", "Preparation", and "Modeling" (which is highlighted), followed by a "MODELING" section containing four buttons: "Artificial Neural Network" (selected), "Linear Regression", "Decision tree", and "KNN".

anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed	sex_transform
1.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	1.0	0.0	1.0
0.0	0.0	0.0	0.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	1.0
0.0	0.0	1.0	1.0
1.0	0.0	0.0	1.0
1.0	0.0	1.0	1.0
1.0	0.0	1.0	0.0
1.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0
1.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0

Figura 5.7: Pestaña "Modeling" de la aplicación.

En contraste, la pestaña "Modeling" nos permite hacer la implementación de algún modelo como Red Neuronal Artificial, Regresión Lineal, Árbol de Decisión o algoritmo KNN, cada uno representado por los botones que se encuentran en la figura 5.7.

Para esta versión sólo se tiene parcialmente implementado el modelo de Red Neuronal, permitiéndonos predecir únicamente campos binarios.

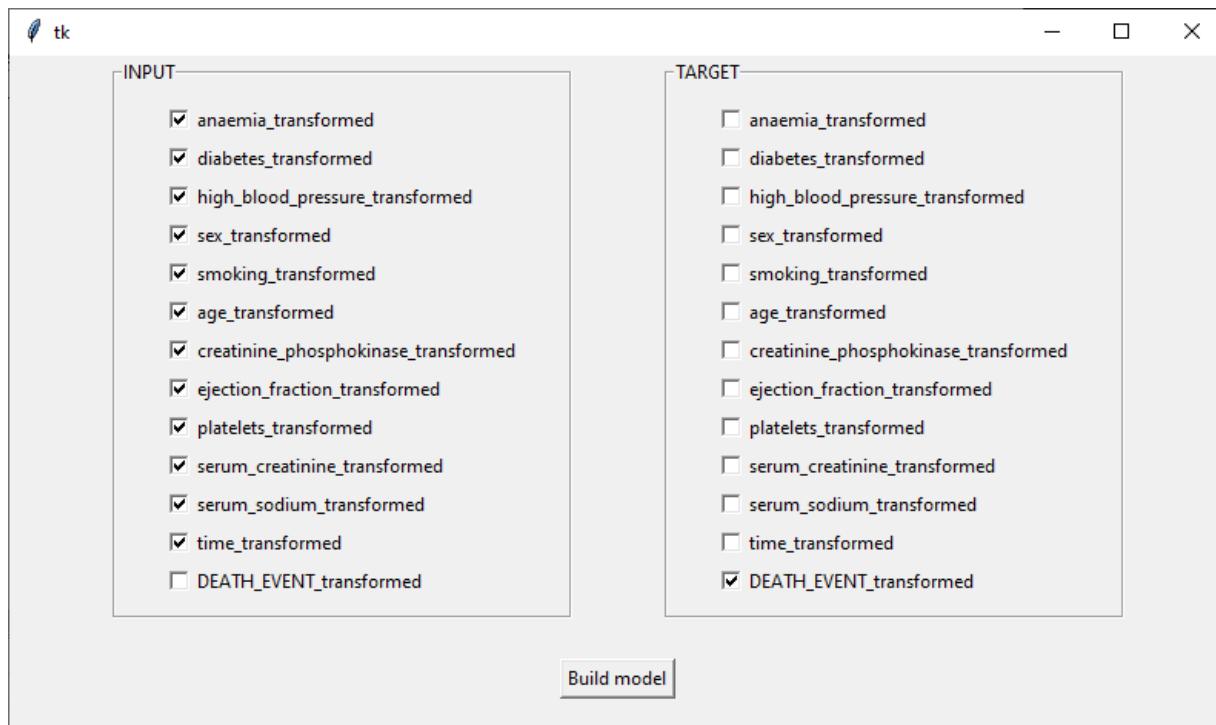


Figura 5.8: Interfaz para seleccionar los INPUT y TARGET.

Al seleccionar el modelo "Artificial Neural Network" se abrirá la ventana emergente que se muestra en la figura 5.8, la cual cuenta con dos secciones, en la sección "INPUT" se deberán seleccionar todos los campos de entrada para el modelo, en la sección "TARGET" se seleccionará el campo destino.

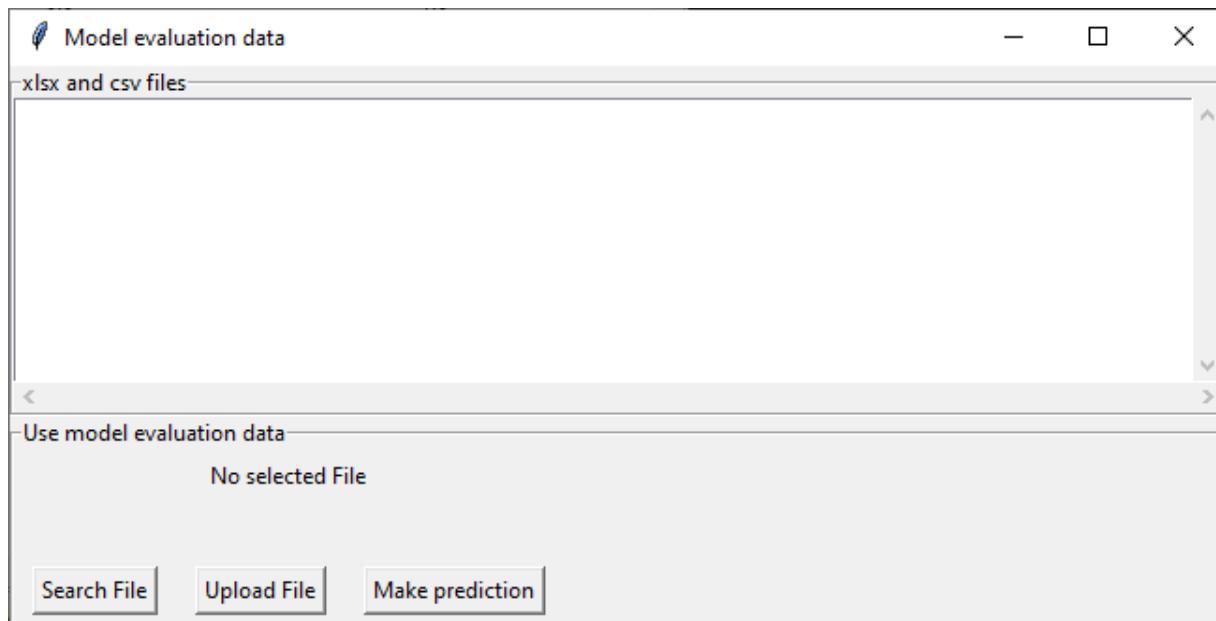


Figura 5.9: Interfaz para cargar los datos de evaluación.

Al presionar el botón "Build Model" nuevamente se abrirá una ventana emergente como la que se visualiza en la figura 5.9, en la cual se deberá seleccionar los datos de evaluación para el modelo.

The screenshot shows a window titled "Model evaluation data". At the top, there is a section labeled "xlsx and csv files" containing a table with three columns: "anaemia_transformed", "diabetes_transformed", and "high_blood_pressure_transformed". The table has five rows of data. Below this is a section labeled "Use model evaluation data" which contains a file path "C:/Users/User/Documents/DPrueba.csv". At the bottom of the window are three buttons: "Search File", "Upload File", and "Make prediction".

anaemia_transformed	diabetes_transformed	high_blood_pressure_transformed
0.0	0.0	1.0
0.0	0.0	0.0
0.0	0.0	0.0
1.0	0.0	0.0
1.0	1.0	0.0

Figura 5.10: Datos de evaluación dentro de la aplicación.

Al cargar los datos dentro de la herramienta tal y como se visualiza en la figura 5.10, es posible comenzar con la predicción al presionar el botón "Make prediction".

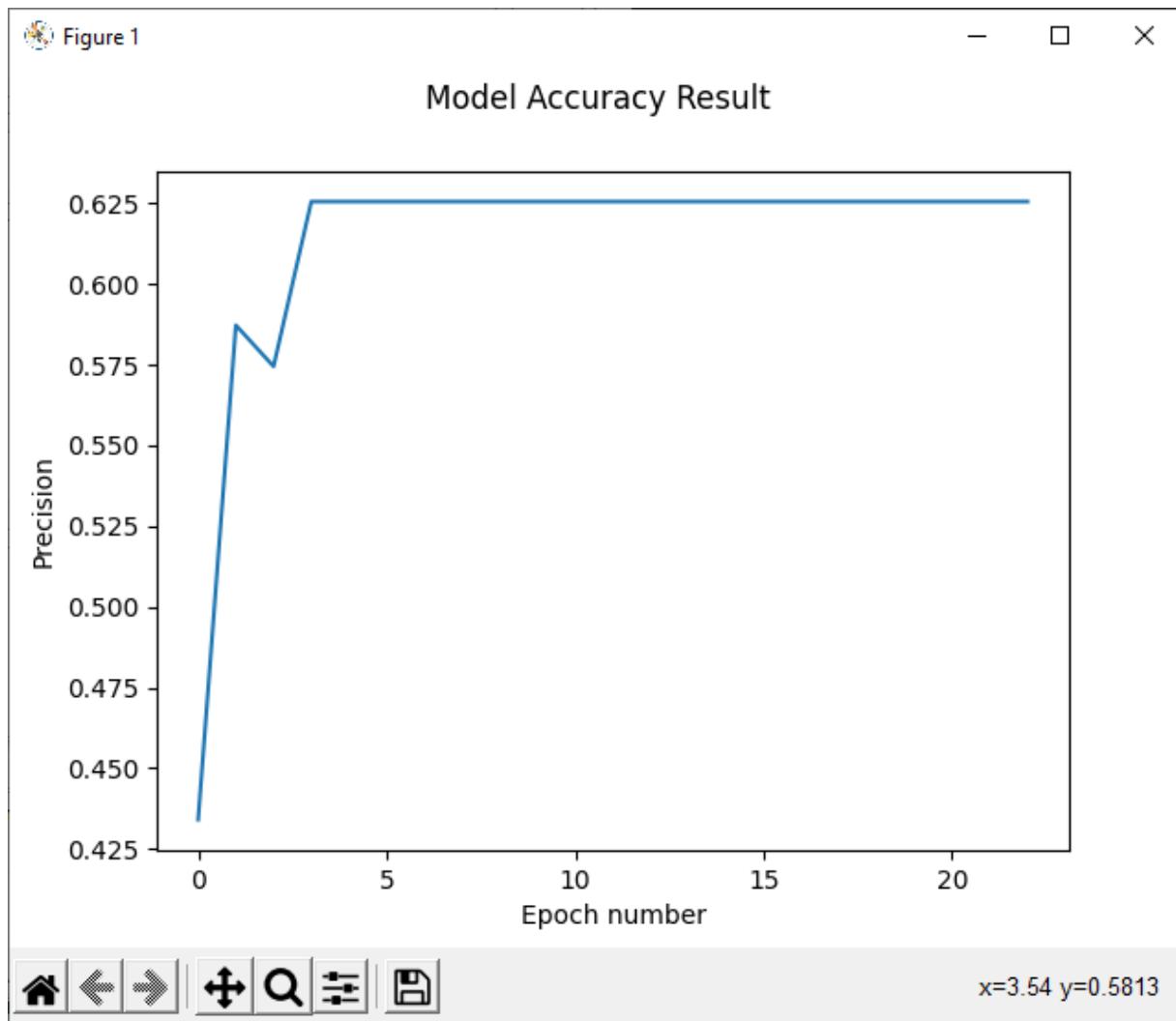


Figura 5.11: Gráfica de la precisión del modelo.

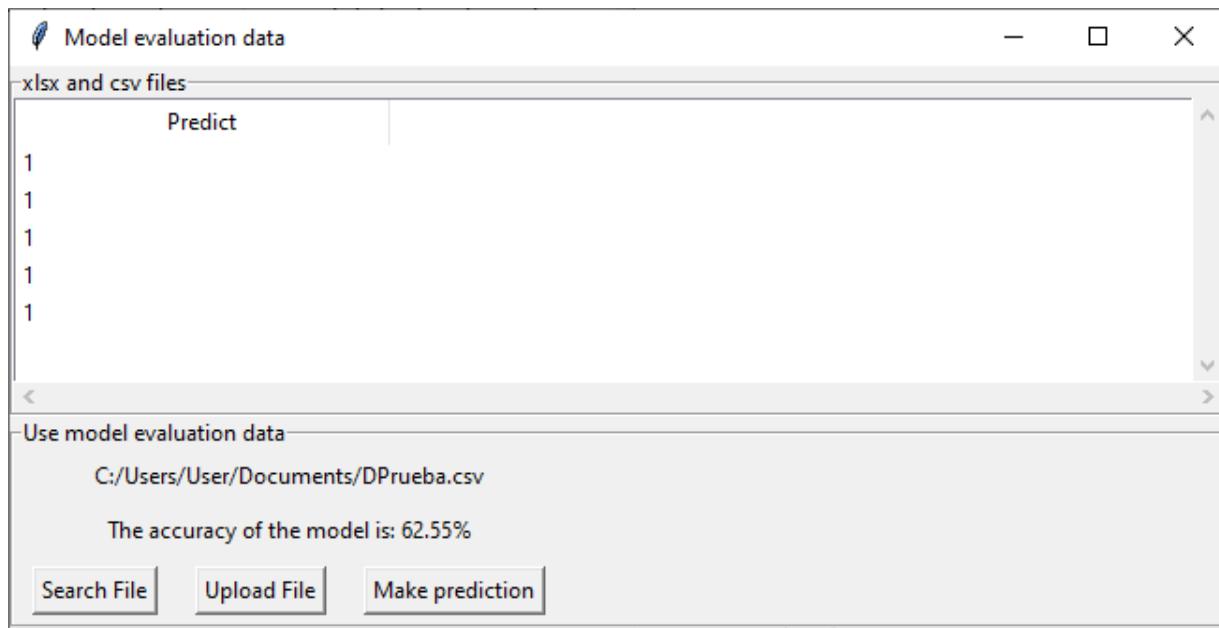


Figura 5.12: Predicción del modelo.

Finalmente, al realizar la predicción se mostrará el gráfico que se encuentra en la figura 5.11 sobre la evolución en la precisión del modelo así como la predicción del modelo que se encuentra en la figura 5.12.

Como se destaca en las figuras 5.11 y 5.12, el desempeño del modelo no es óptimo, pues tiene 62.55 % de precisión; sin embargo, esto no significa que la herramienta no funcione, simplemente la configuración del modelo o los datos no fue correcta. En la siguiente sección, se mostrará la predicción del modelo con un data set diferente para comprobar su correcto funcionamiento.

5.2. Comprobación de su funcionamiento

En las siguientes figuras se mostrará la predicción del modelo de un nuevo data set con el fin de demostrar el funcionamiento de la herramienta.

The screenshot shows a window titled "tk" with a menu bar containing "Help" and a section labeled "xlsx and csv files". The main area displays a data grid with four columns: "duration", "amount", "employment_duration", and "installment_rate". The data consists of 18 rows of numerical values. Below the grid, the file path "C:/Users/ALEX/Desktop/DatosHerramienta.csv" is shown. At the bottom, there are three buttons: "Search file", "Show file", and "Graphing data". Above the buttons, the text "DATA UNDERSTANDING" is displayed.

duration	amount	employment_duration	installment_rate
18.0	1049.0	2.0	4.0
9.0	2799.0	3.0	2.0
12.0	841.0	4.0	2.0
12.0	2122.0	3.0	3.0
12.0	2171.0	3.0	4.0
10.0	2241.0	2.0	1.0
8.0	3398.0	4.0	1.0
6.0	1361.0	2.0	2.0
18.0	1098.0	1.0	4.0
24.0	3758.0	1.0	1.0
11.0	3905.0	3.0	2.0
30.0	6187.0	4.0	1.0
6.0	1957.0	4.0	1.0
48.0	7582.0	1.0	2.0
18.0	1936.0	4.0	2.0
6.0	2647.0	3.0	2.0

Figura 5.13: "Show file" dentro de la aplicación - 2.

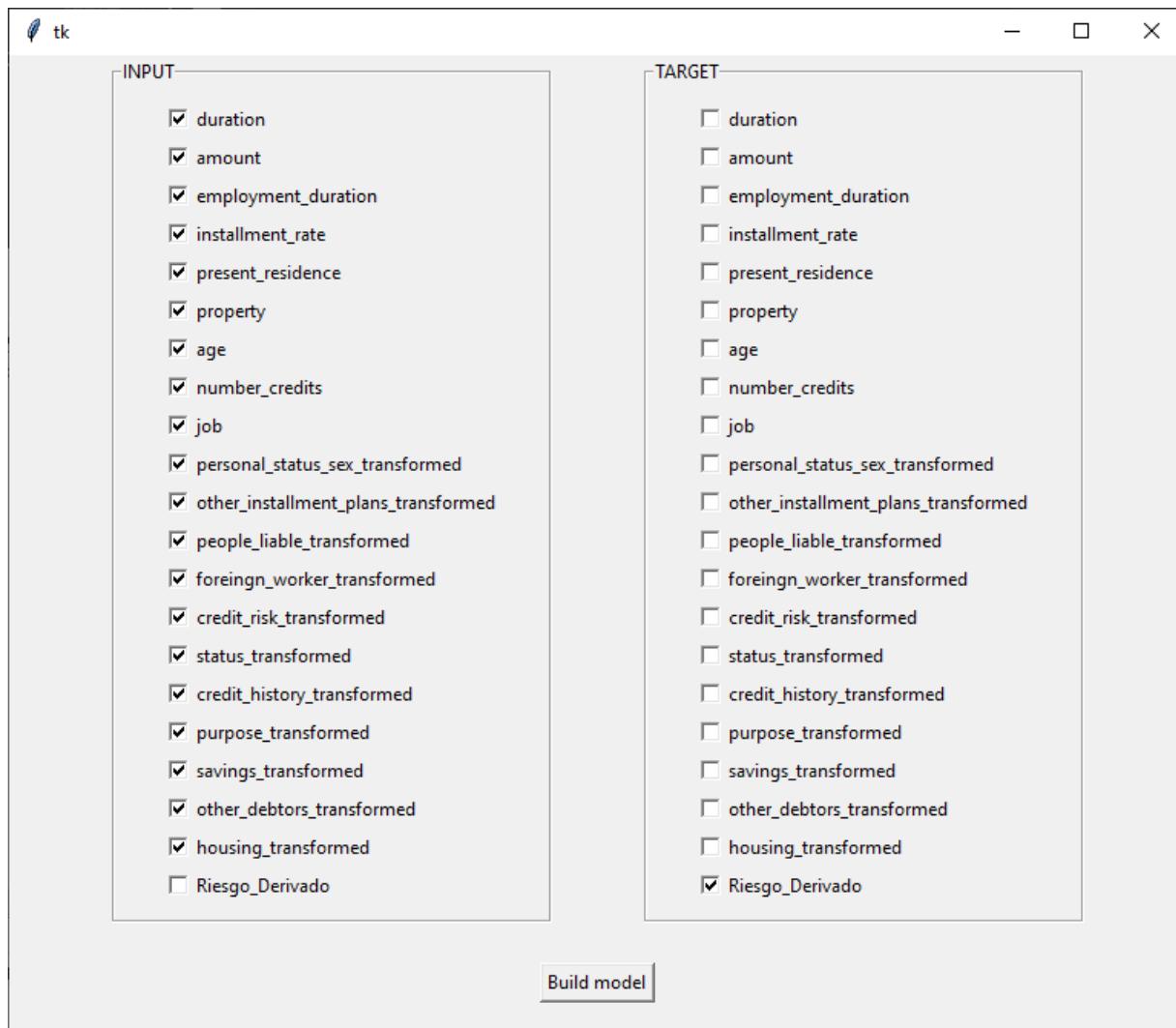


Figura 5.14: Interfaz para seleccionar los INPUT y TARGET - 2.

Model evaluation data

xlsx and csv files

duration	amount	employment_duration
12	691	5
42	4370	4
36	2746	5
24	4110	5
18	2462	3
12	1282	3
-	----	-

Use model evaluation data
C:/Users/ALEX/Desktop/Datos Pruebas.xlsx

Search File Upload File Make prediction

The screenshot shows a user interface for a data analysis tool. At the top, there's a title 'Model evaluation data' with a small icon. Below it is a section titled 'xlsx and csv files' containing a table with three columns: 'duration', 'amount', and 'employment_duration'. The table has several rows of data. At the bottom of the interface, there's a section titled 'Use model evaluation data' with a file path 'C:/Users/ALEX/Desktop/Datos Pruebas.xlsx'. Below this are three buttons: 'Search File', 'Upload File', and 'Make prediction'.

Figura 5.15: Datos de evaluación dentro de la aplicación - 3.

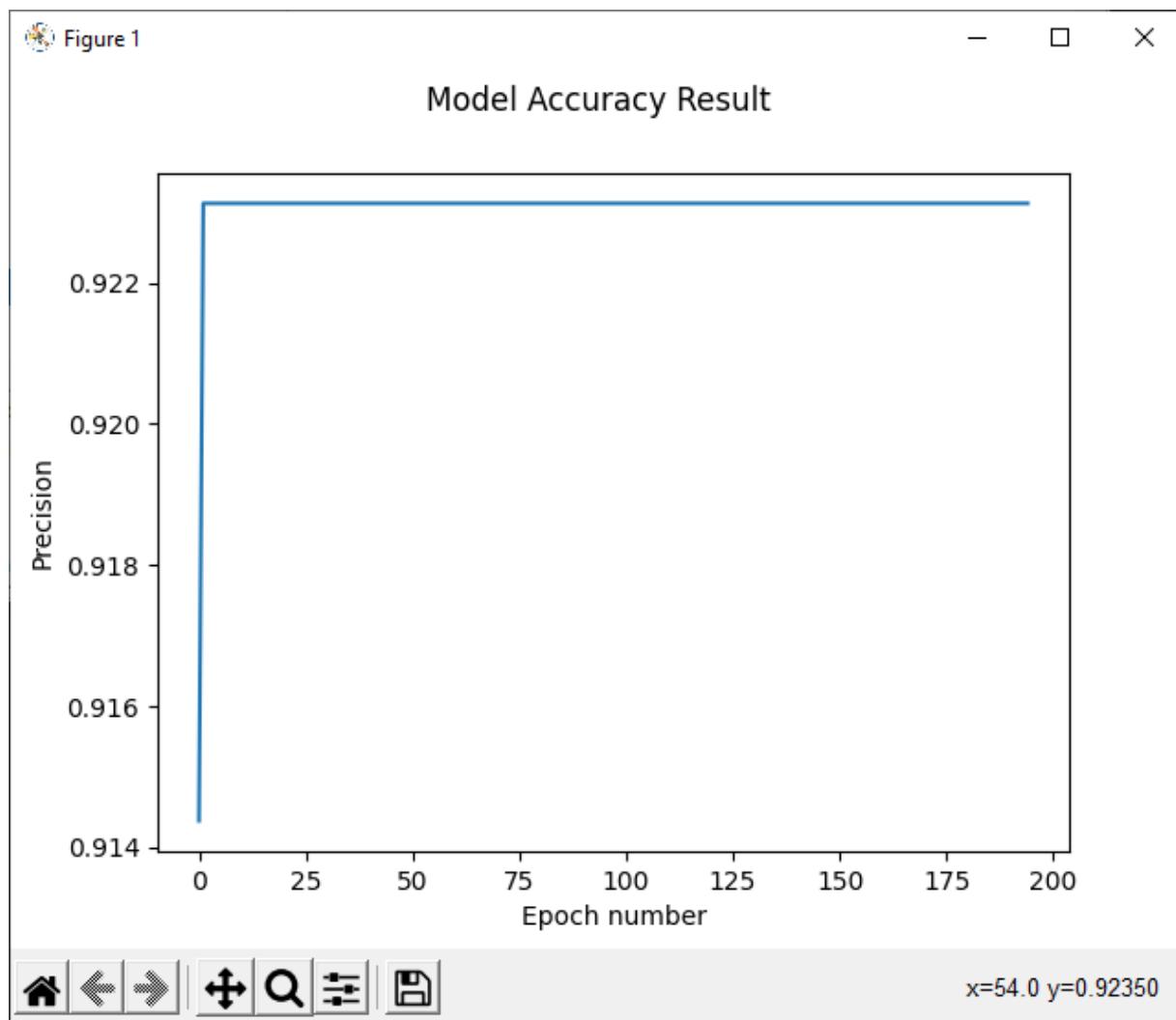


Figura 5.16: Gráfica de la precisión del modelo - 2.

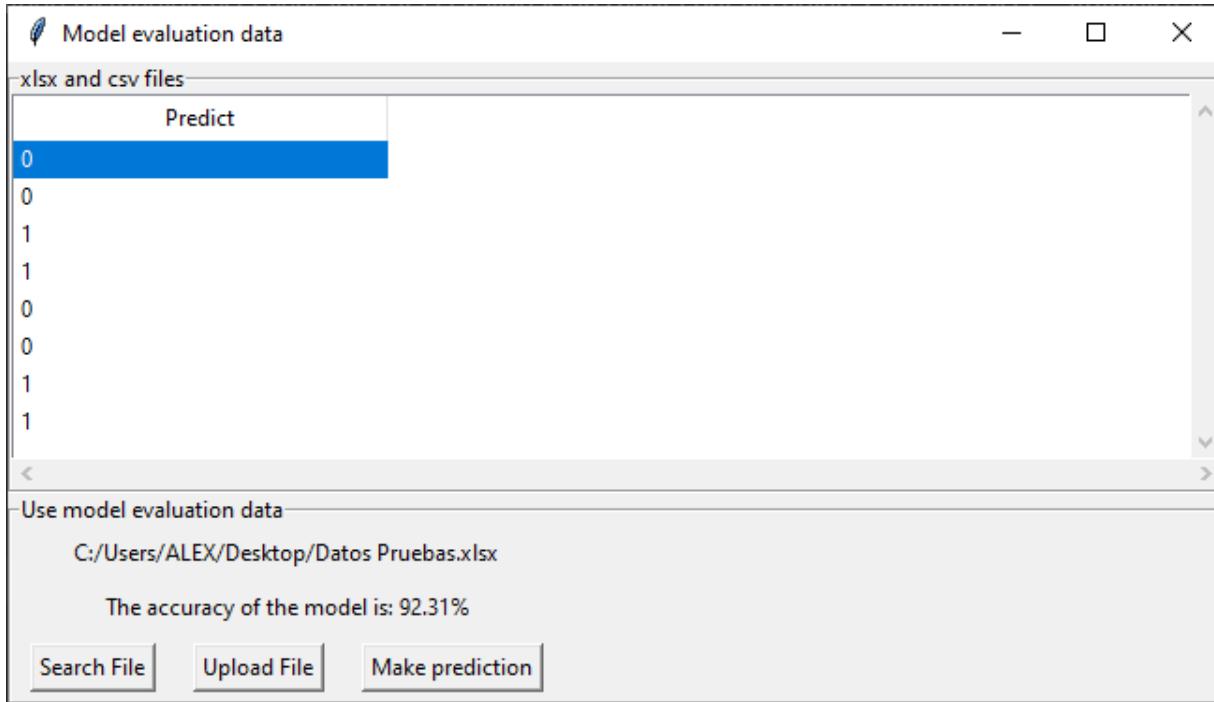


Figura 5.17: Predicción del modelo - 2.

Como se puede apreciar en las figuras 5.13, 5.14, 5.15, 5.16 y 5.17 al trabajar con el nuevo data set y ejecutar el modelo, se tiene como resultado una precisión superior a la que se muestra en la figura 5.14, logrando 92.31 % de efectividad en su precisión, acertando el 100 % de las predicciones esperadas.

El proyecto se encuentra en un [repositorio de GitHub](#), el cual cuenta con una breve explicación de su funcionamiento y las librerías que fueron utilizadas para su correcto funcionamiento al igual que su proceso de instalación.

5.3. Futuros trabajos

Como se mencionó anteriormente, para esta primera versión no se consideraron múltiples funcionalidades de limpieza de datos y tampoco se logró implementar un modelo diferente a la Red Neuronal artificial, por lo que se plantea que la siguiente versión contenga:

- Normalización de los datos.
- Transformación de los datos.
- Creación de árboles de decisión.
- Creación del modelo KNN.
- Gestor de configuración de los modelos.

- Visualización del resumen de los modelos.
- Integración de elementos visuales -iconos- en los botones para mejorar la comprensión de su funcionalidad.

Conclusiones

A lo largo de este trabajo que fue desarrollado durante los proyectos terminales I, II y III, se realizó el estado del arte sobre la aplicación de la minaría de datos en los big data, específicamente hablando, dentro del área de ciencias de la vida, demostrando cómo es que la aplicación de estas metodologías puede beneficiar casos reales de múltiples sectores y mejorar la calidad de vida de los seres humanos.

Al emplear herramientas como *IBM SPSS Modeler* y la herramienta *Intelligent Data Analysis Tool*, se demuestra que aún cuando se utilice el mismo modelo, el resultado no será el mismo gracias a que cada una de ellas implementa una configuración diferente.

El hecho de que la herramienta que se ha desarrollado no presentara un mayor porcentaje de precisión que la aplicación *IBM SPSS Modeler* no quiere decir que no funcione, demuestra que la configuración de los datos que se utilizó, la configuración de la red neuronal artificial e incluso las librerías seleccionadas no fueron las adecuadas para este caso de estudio.

Con lo antes mencionado, puedo destacar que la gran diferencia de usar una aplicación que está establecida en el mercado como lo es *IBM SPSS Modeler* a una aplicación que se ha desarrollado desde cero es la libertad de utilizar la configuración que sea necesaria, pues esto permitirá utilizar toda la capacidad de cómputo, seleccionar los algoritmos que se crean pertinentes, definir la estructura del modelo a nuestra conveniencia, entre muchos otros factores.

En concreto, puedo afirmar que independientemente de la metodología que se esté utilizando para la gestión de los big data, es importe utilizar una metodología cuyo flujo de actividades nos permita trabajar constantemente con la limpieza de los datos y la evaluación del modelado. A lo largo del proyecto, logré mejorar la predicción de mi modelo gracias a que continuamente trabajaba en la limpieza de los datos, pues cuando consideras que no hay valores atípicos, registros nulos o campos con poca importancia, puedes restringirte a mejores resultados.

No obstante, he observado que dentro del área de ciencias de la vida no es viable aceptar un porcentaje de precisión menor al 95 %, pues se trata de una vida humana la que se encuentra en riesgo.

Finalmente, considero que es posible mejorar las predicciones que se obtuvieron durante la elaboración de este proyecto al agregar más campos de los pacientes que sean de tipo continuo, pues cuando un data set contiene demasiados campos categóricos o nominales, es más complicado para el modelo que pueda efectuar una precisión óptima.

Bibliografía

- [1] IBM, “Analítica de Big Data.” [Online]. Available: <https://www.ibm.com/mx-es/analytics/hadoop/big-data-analytics>
- [2] Oracle, “¿Qué es el big data? | Oracle México.” [Online]. Available: <https://www.oracle.com/mx/big-data/what-is-big-data/>
- [3] J. Porway and L. Pierson, *Data Science for Dummies*, 2nd ed. For Dummies, 2017.
- [4] L. Joyanes, *Big Data - Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega, 2022.
- [5] IBM, “5 tipos de datos en el Big Data - Instituto Europeo de Posgrado.” [Online]. Available: <https://www.iep.edu.es/5-tipos-de-datos-en-el-big-data/>
- [6] A. Panesar, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes (English Edition)*, 1st ed. Apress, 2019.
- [7] IBM, “Conceptos básicos de ayuda de CRISP-DM.” [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview>
- [8] J. J. Espinosa Zúñiga, “Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública,” *Ingeniería Investigación y Tecnología*, vol. 21, no. 1, pp. 1–13, 2020.
- [9] F. Berzal, *Redes Neuronales & Deep Learning - Volumen 1: Entrenamiento de redes neuronales artificiales [formato 8.5"x 11"] (Spanish Edition)*. Independently published, 2019.
- [10] IBM, “El modelo de redes neuronales.” [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model>
- [11] X. Basogain, *REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES*, 1st ed. Escuela Superior de Ingeniería de Bilbao, 2011.
- [12] R. E. B. Martínez, N. C. Ramírez, H. G. A. Mesa, I. R. Suárez, M. d. C. G. Trejo, P. P. León, and S. L. B. Morales, “Árboles de decisión como herramienta en el diagnóstico médico,” *Revista médica de la Universidad Veracruzana*, vol. 9, no. 2, pp. 19–24, 2009.

- [13] L. Rokach and O. Maimon, “Top-down induction of decision trees classifiers - a survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.
- [14] MATLAB, “¿Qué es la regresión lineal?” [Online]. Available: <https://la.mathworks.com/discovery/linear-regression.html>
- [15] V. Berlanga-Silvente and R. Vilà-Baños, “How to get a binary logistic regression model with spss,” *Revista d'Innovació i Recerca en Educació*, vol. 7, no. 2, p. 105, 2014.
- [16] “Ordinary Least Squares regression (OLS).” [Online]. Available: [https://www.xlstat.com/es/soluciones/funciones/ordinary-least-squares-regression-ols#:~:text=Ordinary%20Least%20Squares%20regression%20\(OLS\)%20is%20a%20common%20technique%20for,simple%20or%20multiple%20linear%20regression.](https://www.xlstat.com/es/soluciones/funciones/ordinary-least-squares-regression-ols#:~:text=Ordinary%20Least%20Squares%20regression%20(OLS)%20is%20a%20common%20technique%20for,simple%20or%20multiple%20linear%20regression.)
- [17] Merriam-Webster, “medicine.” [Online]. Available: <https://www.merriam-webster.com/dictionary/medicine>
- [18] Instituto Nacional del Cáncer, “Diccionario de cáncer del NCI.” [Online]. Available: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/diagnostico>
- [19] ——, “Diccionario de cáncer del NCI.” [Online]. Available: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/tecnica-diagnostica>
- [20] fisterra.com, “Metodología investigación: Pruebas diagnósticas: Sensibilidad y especificidad.” [Online]. Available: <https://www.fisterra.com/fichas/interior.asp?idArbol=8&idTipoFicha=8&urlseo=pruebas-diagnosticas-sensibilidad-especificidad>
- [21] Universidad Teletón, “Tipos de diagnóstico.” [Online]. Available: https://gc.scalahed.com/recursos/files/r161r/w18964w/handout_s10.pdf
- [22] Instituto Nacional del Cáncer, “Diccionario de cáncer del NCI.” [Online]. Available: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/oncologia>
- [23] ——, “Diccionario de cáncer del NCI.” [Online]. Available: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/cardilogia>