



UNIVERSIDAD AUTÓNOMA METROPOLITANA

DIVISIÓN DE CIENCIAS NATURALES E INGENIERÍA

ANÁLISIS INTELIGENTE DE DATOS
FINANCIEROS: UN CASO DE ESTUDIO EN
EVALUACIÓN DE LA SOLVENCIA PARA
OTORGAR CRÉDITOS.

T E S I S

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

PRESENTA:

ALEJANDRO LÓPEZ VÁZQUEZ

ASESOR:

DR. PEDRO PABLO GONZÁLEZ PÉREZ

Mayo, 2022



Índice general

Índice de figuras	IV
-------------------	----

Índice de tablas	VII
------------------	-----

1 Big Data	1
1.1 Definición	1
1.2 Características de los Datos a Gran Escala	1
1.2.1 Volumen	1
1.2.2 Velocidad	2
1.2.3 Variedad	2
1.2.4 Veracidad	2
1.2.5 Viabilidad	2
1.2.6 Visualización de los datos	2
1.2.7 Valor	2
1.3 Tipos de datos dentro del Big Data y de dónde provienen	3
1.3.1 Web y Social Media	3
1.3.2 Máquina a Máquina (M2M)	4
1.3.3 Gran transacción de datos	4
1.3.4 Biométrica	4
1.3.5 Generada por humanos	4
1.4 Tipos de Datos	5
1.4.1 Datos estructurados	5
1.4.2 Datos semi-estructurados	5
1.4.3 Datos no estructurados	6
1.5 Ciclo de vida genérico del análisis de datos	7
1.5.1 Etapas y principales actividades del ciclo de vida genérico del análisis de datos	8
1.6 Casos de uso de los datos a gran escala	9

2	Técnicas de Modelado	13
2.1	Redes Neuronales Artificiales.	13
2.1.1	Estructura Básica de las Redes Neuronales Artificiales	14
2.1.2	Fundamentos de las Redes Neuronales Artificiales	15
2.1.3	Tipos y selección de la Redes Neuronales Artificiales	19
2.2	Algoritmos de Regresión Lineal	20
2.2.1	Tipos de regresión lineal	21
2.2.2	Aplicaciones de la regresión lineal	24
2.3	Regresión Logística	24
2.3.1	Diferencias entre Regresión Lineal y Regresión Logística	25
2.3.2	Tipos de Regresión Logística	26
2.4	Algoritmo K-Nearest Neighbors	26
2.4.1	Aplicación del Algoritmo	27
2.5	Árboles de Decisión	30
2.5.1	Ventajas del Árbol de decisiones	30
2.5.2	Tipos de nodos	31
3	Área de Estudio: Finanzas	33
3.1	Definición	33
3.2	Tipos de finanzas	33
3.2.1	Finanzas corporativas	33
3.2.2	Finanzas personales	34
3.2.3	Finanzas públicas	35
3.2.4	Finanzas Internacionales	36
3.3	Usos de los Datos a Gran Escala en las Finanzas Corporativas y Personales	38
3.3.1	Gestión de riesgos	39
3.3.2	Evaluación de la solvencia	39
3.3.3	Prevención del fraude	39
3.3.4	Mejora de la atención al cliente	39
3.3.5	Personalizar servicios financieros	39
3.3.6	Trading de alta frecuencia	40
3.3.7	Trading de productos primarios	40
3.3.8	Asesoramiento en inversiones	40
3.3.9	Fidelización de clientes	40
3.3.10	Casos de uso de los Datos a Gran Escala aplicados a las Empresas Financieras	41
4	Aplicación de la metodología CRISP-DM	42
4.1	Comprensión del dominio del problema	42
4.1.1	Determinación de los objetivos del proyecto	43
4.1.2	Valoración de la situación actual del objetivo del proyecto	43
4.1.3	Determinación de los objetivos de minería de datos	45
4.1.4	Propuesta del enfoque metodológico (plan de proyecto de minería de datos) en forma de tabla	45
4.2	Comprensión de los datos	46

4.2.1	Recopilación de los datos iniciales	46
4.2.2	Descripción de los datos	46
4.2.3	Exploración de los datos	48
4.2.4	Verificación de la calidad de los datos	52
4.3	Preparación de los datos	53
4.3.1	Selección de datos	53
4.3.2	Limpieza de datos	53
4.3.3	Integración de datos	54
4.3.4	Construcción de nuevos datos	56
4.3.5	Formato de datos	57
4.3.6	Nueva exploración de los datos	58
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	Generación del modelo: Red Neuronal Perceptron Backpropagation	66
4.4.5	Ejecución del modelo: Red Neuronal Perceptron Backpropagation .	69
4.4.6	Generación del modelo: Árbol de decisión C&R	74
4.4.7	Ejecución del modelo: Árbol de decisión C&R	77
4.4.8	Generación del modelo: Regresión logística	80
4.4.9	Ejecución del modelo: Algoritmo de Regresión Logística	82
4.4.10	Generación del modelo: Algoritmo KNN	83
4.4.11	Ejecución del modelo: Algoritmo KNN	88
4.4.12	Generación del modelo: Algoritmo de Regresión Lineal	88
4.4.13	Ejecución del modelo: Algoritmo de Regresión Lineal	92
4.5	Evaluación de los Modelos	93
5	Herramienta Intelligent Data Analysis Tool	103
	Conclusiones	114
	Bibliografía	116

Índice de figuras

1.1	Las 7V del Big Data.	3
1.2	Tipos de datos dentro del Big Data.	5
1.3	Diagrama UML de los tipos de Datos.	6
1.4	Ciclo de vida genérico del análisis de datos.	7
1.5	Diagrama UML del ciclo de vida genérico del análisis de datos.	9
1.6	Diagrama UML de casos de uso del big data.	12
2.1	Diagrama de una Neurona Artificial (PE).	14
2.2	Arquitectura de una Red Neuronal Simple. [1]	15
2.3	Modelo de Neurona Artificial. [2]	16
2.4	Red Neuronal de una Capa.	17
2.5	Red Neuronal de dos Capas.	18
2.6	Ej. Regresión Lineal Simple. [3]	21
2.7	Ej. Regresión Lineal Múltiple. [3]	22
2.8	Ej. Regresión Lineal Múltivariante. [3]	23
2.9	Ej. Regresión Lineal Múltiple Multivariante. [3]	24
2.10	Ej. Función Logística. [4]	25
2.11	Ej. Conjunto de datos clasificados en dos categorías. [5]	27
2.12	Ej. Clasificación de un nuevo vector de datos. [5]	28
2.13	Ej. Selección de los K=5 vecinos más cercanos al nuevo elemento. [5]	29
2.14	Ej. 3 elementos de la categoría 1 y 2 elementos de la categoría 2. [5]	29
2.15	Ej. 3 elementos de la categoría 1 y 2 elementos de la categoría 2. [5]	30
2.16	Estructura de un Árbol de Decisión.	32
3.1	Diagrama UML, finanzas corporativas.	34
3.2	Diagrama UML, finanzas personales.	35
3.3	Diagrama UML, Finanzas Públicas.	36
3.4	Diagrama UML, Finanzas Internacionales.	37
3.5	Diagrama UML, Finanza.	38
4.1	Vista Previa del dataset "South German Credit".	46
4.2	Histograma de la edad de los clientes.	48
4.3	Gráfica Monto x Duración del crédito.	49
4.4	Gráfica Edad de los clientes x Núm. de Créditos obtenidos.	50
4.5	Gráfica de la edad de los clientes x el monto que solicitan.	51

4.6	Histograma del propósito del crédito. Donde 0 = otros, 1 = Carro Nuevo, 2 = Carro Usado, 3 = Muebles o equipamiento, 4 = Radio o televisión, 5 = Usos domésticos, 6 = Reparaciones, 7 = Educación, 8 = vacaciones, 9 = Cursos, 10 = Negocios.	52
4.7	SPSS Modeler, aplicación de preparación automática de datos y aplicación de un filtro a los datos.	53
4.8	SPSS Modeler, Eliminación del registro "Teléfono".	54
4.9	SPSS Modeler, Anexar nuevos datos adquiridos.	55
4.10	SPSS Modeler, Resultado de Anexar datos.	56
4.11	SPSS Modeler, Derivación de un nuevo campo.	57
4.12	Gráfica Monto vs Duración del Crédito.	58
4.13	Histograma del Tipo de Vivienda según la edad de los clientes.	59
4.14	Gráfica de la edad vs el número de créditos obtenidos.	60
4.15	Gráfica entre el propósito del crédito vs el monto.	61
4.16	Histograma del sexo de los clientes.	62
4.17	Propósito del Crédito. dónde 0.0 = Vacaciones, 1.0 = Radios / Televisiones, 2.0 = Negocios, 3.0 = Aplicaciones Domésticas, 4.0 = Reparos, 5.0 Cursos de capacitación, 6.0 = Carro Nuevo, 7.0 = Carro Usado, 8.0 = Otros, 9.0 = Mobiliario / Maquinaria.	63
4.18	Histograma del cumplimiento del crédito según la edad.	64
4.19	Vista del Proyecto, <i>IBS SPSS Modeler</i>	66
4.20	<i>IBS SPSS Modeler</i> , Generación del modelo red neuronal.	67
4.21	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo red neuronal.	68
4.22	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo red neuronal 2.	69
4.23	<i>IBS SPSS Modeler</i> , Resumen del modelo red neuronal.	70
4.24	<i>IBS SPSS Modeler</i> , Importancia del predictor del modelo red neuronal.	71
4.25	<i>IBS SPSS Modeler</i> , Clasificaciones del registro Destino del modelo red neuronal.	72
4.26	<i>IBS SPSS Modeler</i> , Red neuronal Perceptron Backpropagation.	73
4.27	<i>IBS SPSS Modeler</i> , Generación del modelo árbol de decisión C&R.	74
4.28	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo árbol de decisión C&R.	75
4.29	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo árbol de decisión C&R 2.	76
4.30	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo árbol de decisión C&R 3.	77
4.31	<i>IBS SPSS Modeler</i> , Importancia del predictor del modelo árbol de decisión C&R.	78
4.32	<i>IBS SPSS Modeler</i> , Modelo árbol de decisión C&R.	78
4.33	<i>IBS SPSS Modeler</i> , Nodo 0 del árbol de decisión C&R.	79
4.34	<i>IBS SPSS Modeler</i> , Nodo 1 del árbol de decisión C&R.	79
4.35	<i>IBS SPSS Modeler</i> , Nodo 2 del árbol de decisión C&R.	80
4.36	<i>IBS SPSS Modeler</i> , Generación del modelo Regresión logística.	81
4.37	<i>IBS SPSS Modeler</i> , Opciones para la Generación del modelo Regresión Logística.	82

4.38	<i>IBS SPSS Modeler</i> , Análisis del modelo Regresión Logística.	83
4.39	<i>IBS SPSS Modeler</i> , Generación del modelo KNN.	84
4.40	<i>IBS SPSS Modeler</i> , Selección del objetivo para el modelo KNN.	85
4.41	<i>IBS SPSS Modeler</i> , Opciones para la generación del modelo KNN.	86
4.42	<i>IBS SPSS Modeler</i> , Opciones para la generación del modelo KNN 2.	87
4.43	<i>IBS SPSS Modeler</i> , Resultados del modelo KNN.	88
4.44	<i>IBS SPSS Modeler</i> , Selección del objetivo para el modelo Algoritmo de Regresión Lineal.	89
4.45	<i>IBS SPSS Modeler</i> , Opciones para la generación del modelo Algoritmo de Regresión Lineal.	90
4.46	<i>IBS SPSS Modeler</i> , Opciones para la generación del modelo Algoritmo de Regresión Lineal 2.	91
4.47	<i>IBS SPSS Modeler</i> , Opciones para la generación del modelo Algoritmo de Regresión Lineal 3.	92
4.48	<i>IBS SPSS Modeler</i> , Resultados del modelo Algoritmo de Regresión Lineal.	93
4.49	<i>IBS SPSS Modeler</i> , Evaluación de los resultados del modelo Red Neuronal Perceptron Backpropagation.	94
4.50	<i>IBS SPSS Modeler</i> , Evaluación de los resultados del modelo Árbol de Decisión C&R.	94
4.51	<i>IBS SPSS Modeler</i> , Evaluación de los resultados del modelo Algoritmo de Regresión Logística.	94
4.52	<i>IBS SPSS Modeler</i> , Evaluación de los resultados del modelo Algoritmo KNN.	95
4.53	<i>IBS SPSS Modeler</i> , Evaluación de los resultados del modelo Algoritmo de Regresión Lineal.	95
4.54	Datos para la evaluación.	96
4.55	Tipos de los campos de los datos de prueba.	97
4.56	Tipos de los campos de los datos de prueba.	98
4.57	Salida esperada en la evaluación de los modelos.	99
4.58	<i>IBM SPSS Modeler</i> , Evaluación del modelo Red Neuronal Perceptron Backpropagation con nuevos datos.	99
4.59	<i>IBM SPSS Modeler</i> , Salida generada por el modelo Red Neuronal Perceptron Backpropagation.	99
4.60	<i>IBM SPSS Modeler</i> , Evaluación del modelo Árbol de Decisión C&R con nuevos datos.	100
4.61	<i>IBM SPSS Modeler</i> , Salida generada por el modelo Árbol de Decisión C&R.	100
4.62	<i>IBM SPSS Modeler</i> , Evaluación del modelo Algoritmo de Regresión Logística.	100
4.63	<i>IBM SPSS Modeler</i> , Salida generada por el modelo Algoritmo de Regresión Logística.	101
4.64	<i>IBM SPSS Modeler</i> , Evaluación del modelo Algoritmo KNN con nuevos datos.	101
4.65	<i>IBM SPSS Modeler</i> , Salida generada por el modelo Algoritmo KNN.	101
4.66	<i>IBM SPSS Modeler</i> , Evaluación del modelo Algoritmo de Regresión Lineal con nuevos datos.	102
4.67	<i>IBM SPSS Modeler</i> , Salida generada por el modelo Algoritmo de Regresión Lineal.	102

5.1	Interfaz de introducción.	104
5.2	Interfaz principal.	105
5.3	Visualización de los datos en la herramienta.	106
5.4	Interfaz para graficar los datos.	106
5.5	Gráfico.	107
5.6	Opciones de limpieza de los datos.	108
5.7	Opciones de técnicas de modelado.	109
5.8	Selección de campos de entrada y salida del modelo.	110
5.9	Selección de los datos de prueba para el modelo.	111
5.10	Comenzar con la predicción.	111
5.11	Gráfico de la evolución en la precisión.	112
5.12	Predicción y tasa de efectividad del modelo.	113

Índice de tablas

1.1	Casos de uso del big data.	10
2.1	Funciones más utilizadas en las Redes Neuronales Artificiales	17
2.2	Clasificación de las ANN	19
3.1	Casos de uso del big data en las empresas.	41
4.1	Plan de proyecto de minería de datos.	45
4.2	Descripción y tipo de los datos.	47
4.3	Formato de los datos.	57

Capítulo 1

Big Data

1.1. Definición

El Big Data son conjuntos de datos de mayor tamaño y más complejos, procedentes particularmente de nuevas fuentes de datos como: sensores, correos electrónicos, páginas de internet, bases de datos, redes sociales, smartwatches, etc, estos datos se presentan en volúmenes que tienen una mayor velocidad de crecimiento.

Debido a su volumen y complejidad, se dificulta su captura, gestión, procesamiento o análisis por lo tanto, no se puede usar tecnologías, software o sistemas de cómputo convencionales, por lo que, para su análisis se usan métodos estadísticos, de aprendizaje automatizado, de aprendizaje profundo, entre otros.

1.2. Características de los Datos a Gran Escala

Las características más importantes de los Datos a Gran Escala (Big Data) son las conocidas como las cuatro V del Big Data, que se refieren al Volumen, Variedad, Velocidad y Veracidad, pero con el paso del tiempo, además de estas cuatro V, se les han añadido tres V más, que corresponden a Viabilidad, Visualización de los Datos y el Valor, esta última considerada como la principal.[6]

A continuación se realizará una breve descripción de cada una de estas V's de los Datos a Gran Escala [7]:

■ 1.2.1. Volumen

Es la característica mayormente asociada al Big Data, ya que se hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesar los datos para obtener un análisis de estos.

■ 1.2.2. Velocidad

Se refiere a la rapidez en la que los datos son creados, almacenados y procesados en tiempo real.

■ 1.2.3. Variedad

Se refiere a las formas, tipos y fuentes en la que se registran los datos. Pueden ser datos estructurados (como bases de datos) o no estructurados (como correos electrónicos, datos de sensores, audios, videos o imágenes, redes sociales, etc).

■ 1.2.4. Veracidad

Se refiere a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recopilada.

■ 1.2.5. Viabilidad

Se trata de la capacidad que tienen las compañías en generar un uso eficaz del gran volumen de datos que manejan.

■ 1.2.6. Visualización de los datos

Se refiere al modo en el que los datos son presentados una vez que son procesados, para encontrar patrones en el tema a investigar.

■ 1.2.7. Valor

El valor se obtiene de datos que se transforman en información; esta a su vez se transforma en conocimiento, y este último en acción o en decisión. El valor de los datos está en poder tomar la mejor decisión en base a estos datos.

La Figura 1.1 es una representación gráfica de las Características de los Datos a Gran Escala 1.2.



Figura 1.1: Las 7V del Big Data.

1.3. Tipos de datos dentro del Big Data y de dónde provienen

Existen muchas categorías de información dentro del Big Data, la IBM clasifica cinco tipos de datos [8]:

1. 1.3.1. Web y Social Media

Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, Instagram, Uber, LinkedIn, etc.

2. 1.3.2. Máquina a Máquina (M2M)

Se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza sensores o medidores que capturan eventos en particular (velocidad, temperatura, presión, variables químicas, etc) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3. 1.3.3. Gran transacción de datos

Incluye registros de facturación, en telecomunicaciones como los registros detallados de las llamadas telefónicas que realizamos.

4. 1.3.4. Biométrica

Información en la que se incluyen huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5. 1.3.5. Generada por humanos

Las personas generamos grandes y diversas cantidades de datos, por ejemplo, llamadas telefónicas, notas de voz, correos electrónicos, estudios médicos, encuestas de marketing, etc.

La Figura 1.2 es una representación gráfica de la clasificación de los cinco tipos de datos según la IBM.

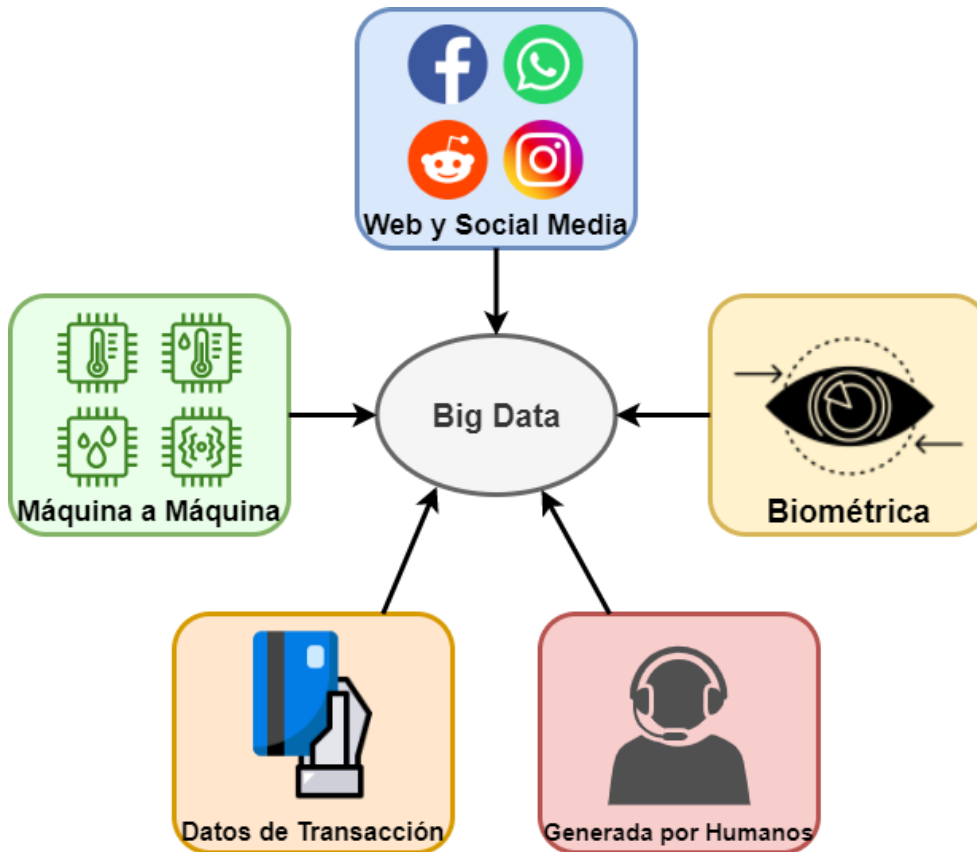


Figura 1.2: Tipos de datos dentro del Big Data.

1.4. Tipos de Datos

Los tipos de datos descritos en el apartado 1.3 pueden ser organizados de tres formas diferentes. Pueden ser de tipo estructurados, semi-estructurados o no estructurados [9]:

■ 1.4.1. Datos estructurados

Son los datos que se encuentran ordenados mediante un serie de filas y columnas definidas, como una tabla de Excel, hojas de cálculo, bases de datos relacionales, entre otros.

Este tipo de datos es más fácil de gestionar, y gracias a su organización, permiten una mayor predictibilidad que los otros tipos de datos.

■ 1.4.2. Datos semi-estructurados

Son los datos que tienen un cierto nivel de estructura, jerarquía y organización pero carecen de un esquema fijo, como los correos electrónicos, archivos comprimidos,

ejecutables binarios, paquetes TCP/IP, etc.

Si se procesan este tipo de datos, se puede conseguir almacenamiento en una base de datos relacional y también en filas y columnas. Sin embargo, no todos los que se colocan en un grupo tienen siempre las mismas propiedades. A veces difieren en tipo y tamaño. Además, contienen metadatos (etiquetas y elementos) que se utilizan para agruparlos y describir cómo se almacenan.

■ 1.4.3. Datos no estructurados

Son los datos que no están estructurados a través de modelos o esquemas de datos fijos y predefinidos, como documentos en archivos de texto, imágenes, archivos PDF, datos de redes sociales, archivos de audio, etc.

Este tipo de datos puede ser almacenado dentro de una base de datos relacional o NoSQL, y requieren análisis avanzados para encontrar información valiosa.

La Figura 1.3 es una representación en diagrama UML de los tipos de datos.

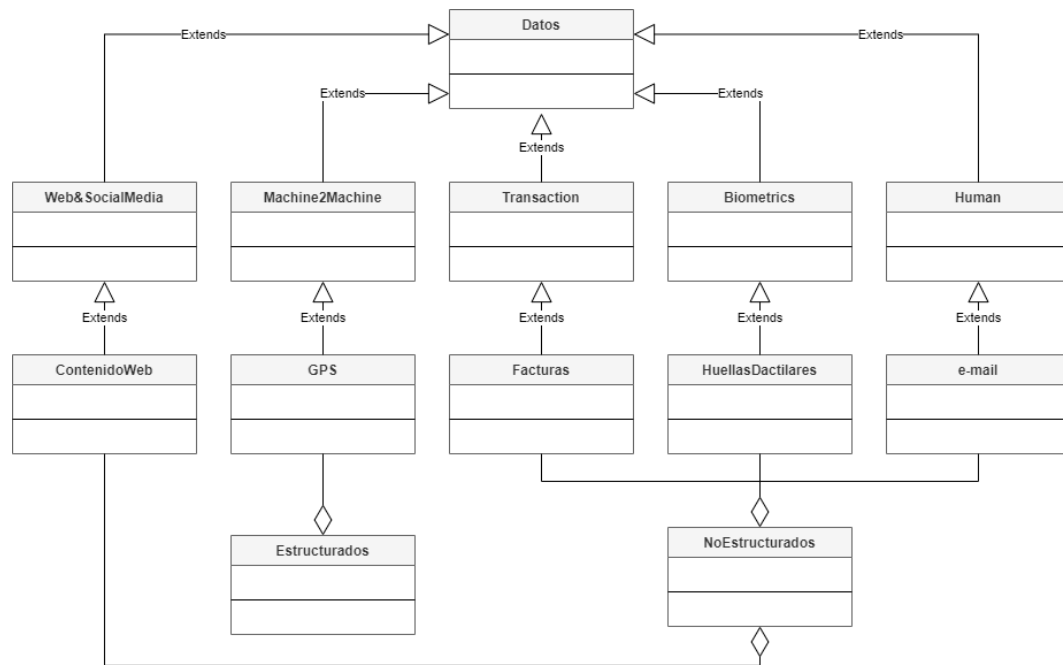


Figura 1.3: Diagrama UML de los tipos de Datos.

Dato Interesante: La IBM estima que la información digital disponible en el mundo, actualmente, es de 5 Zetabytes (cien trillones de bits), y esta se duplica cada dos años y medio. [8]

1.5. Ciclo de vida genérico del análisis de datos

Los datos masivos sin procesar tienen poco valor por sí mismos, para que puedan tener un valor significativo es necesario hacer un procesamiento de estos. La disciplina surgida de esta necesidad fue denominada como Ciencia de Datos.

La Ciencia de Datos combina un conjunto amplio de técnicas provenientes de múltiples disciplinas tales como las Ciencias de la Computación, Matemáticas, Estadística, Econometría e Investigación Operativa.

El ciclo de vida genérico del análisis de datos considera al menos etapas las cuales son representadas en la siguiente Figura 1.4:

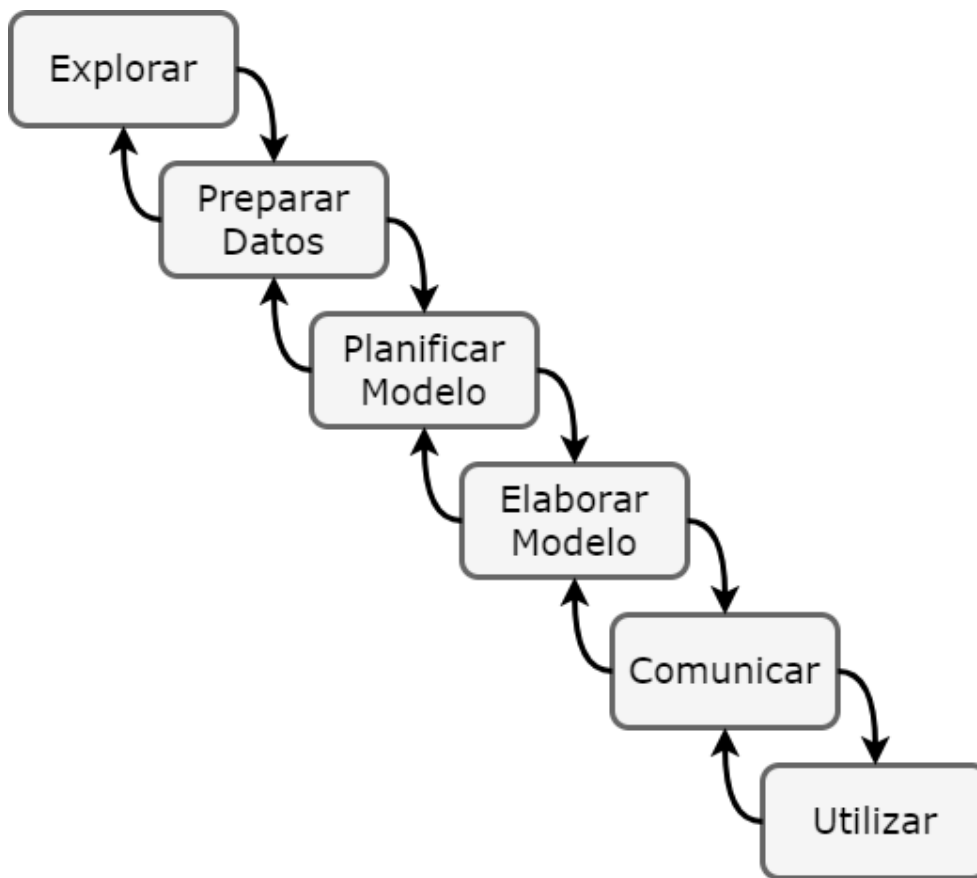


Figura 1.4: Ciclo de vida genérico del análisis de datos.

Como se puede apreciar en la figura 1.4, el ciclo de vida genérico del análisis de datos no es un proceso lineal, ya que muchas ocasiones es necesario reformular las preguntas de los objetivos del proyecto en función de la disponibilidad de los datos, o reinterpretar los resultados debido al surgimiento de nueva información. Por ende, este ciclo es más un proceso iterativo en el cual se puede retroceder a etapas previas.

1.5.1. Etapas y principales actividades del ciclo de vida genérico del análisis de datos

A continuación, se mostrará un listado de las etapas del ciclo de vida genérico y las principales actividades que se deben realizar en cada etapa de este ciclo. [10]

1. Etapa de Exploración

Las actividades principales a realizar en esta etapa son:

- Explorar los datos disponibles.
- Formular las preguntas que fijen el objetivo.
- Formular una hipótesis.

2. Etapa de Preparación de los Datos

Las actividades principales a realizar en esta etapa son:

- Determinar los datos necesarios.
- Recopilar los datos.
- Limpiar los datos.
- Analizar la consistencia de los datos.

3. Etapa de Planificación del Modelo

Las actividades principales a realizar en esta etapa son:

- Determinar qué variables explican y cuáles predecir.
- Seleccionar posibles modelos y algoritmos a utilizar.
- Definir métricas de desempeño.

4. Etapa de Elaboración del Modelo

Las actividades principales a realizar en esta etapa son:

- Implementar los modelos.
- Determinar el mejor modelo según ajuste y significancia.
- Validar el modelo.

5. Etapa de Comunicación

Las actividades principales a realizar en esta etapa son:

- Interpretar resultados.
- Generar visualizaciones adecuadas para comunicar los resultados.
- Hacer recomendaciones de mejora.

6. Etapa de Utilizar la información obtenida

Las actividades principales a realizar en esta etapa son:

- Tomar decisiones basadas en los resultados.
- Definir estándares de servicio.
- Metas para alcanzar estándares.
- Planificar y asignar recursos.

La Figura 1.5 muestra un diagrama de actividades UML el cual es ocupado para representar las etapas y principales actividades del ciclo de vida genérico del análisis de datos.

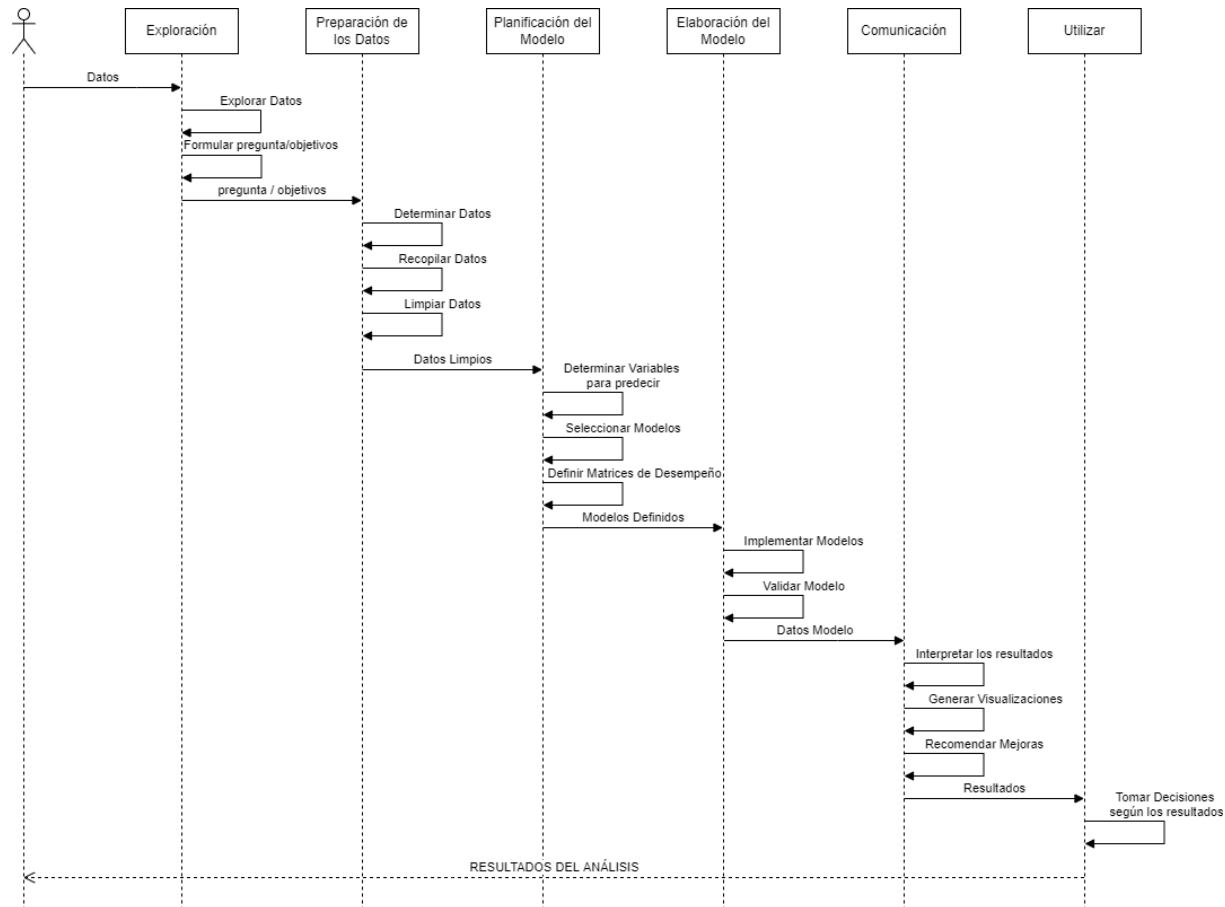


Figura 1.5: Diagrama UML del ciclo de vida genérico del análisis de datos.

1.6. Casos de uso de los datos a gran escala

Gracias al surgimiento de los datos a gran escala nos es posible dar respuesta a problemas que con la tecnología y la información de antes hubiera sido imposible, ya que, como su nombre lo indica, estos ofrecen una mayor cantidad de información, la cual es usada por distintas empresas para ofrecernos más y mejores servicios, experiencias personalizadas, entre otros beneficios.

A continuación, se mostrará una recopilación de casos de uso sobre el Big Data.

Tabla 1.1: Casos de uso del big data.

[11]

APLICACIÓN	DESCRIPCIÓN
Motor de recomendación	Los minoristas en línea utilizan herramientas de coincidencias para que los usuarios se recomienden entre sí o para recomendar productos y servicios basados en el análisis del perfil de usuario y los datos de comportamiento. LinkedIn utiliza este enfoque para alimentar su mensaje «la gente puede saber», mientras que Amazon lo usa para sugerir productos relacionados con las compras de los consumidores.
Análisis de percepciones	En este enfoque se utilizan herramientas de Big Data junto con otras herramientas de texto avanzadas que revisan el contenido de mensajes no estructurado de las redes sociales, incluyendo los Tweets y mensajes de Facebook, para determinar el “sentimiento de usuario” o percepción relacionado con determinadas empresas, marcas o productos.
Análisis y modelos de riesgo	Empresas financieras, bancos y otras entidades crediticias utilizan Big Data junto a Data Warehouse para analizar grandes volúmenes de datos transaccionales y determinar el riesgo y la exposición de los activos financieros. Trabajan con escenarios del tipo «qué pasaría si» basados en simulaciones de mercado. El objetivo es calificar el riesgo de clientes potenciales que solicitan préstamos.
Detección de Fraude	Las empresas financieras, las tarjetas de crédito, los minoristas y otros utilizan técnicas de Big Data para combinar el comportamiento del cliente, datos históricos y transaccionales para detectar la actividad fraudulenta.
Análisis campaña de marketing	Los departamentos de marketing en todas las industrias han utilizado durante mucho tiempo la tecnología para monitorear y determinar la eficacia de las campañas. El Big Data permite a los equipos de marketing para incorporar mayores volúmenes de datos cada vez más granulares, como las secuencias de clics para aumentar la exactitud del análisis.
Abandono de clientes	Las empresas utilizan tecnologías de Big Data para analizar los datos de comportamiento de los clientes e identificar patrones que indican qué clientes son más propensos a dejar por un proveedor o servicio y pasarse a un competidor.

Análisis de la experiencia de cliente	Las empresas de consumo usan tecnologías Big Data para integrar los datos de los canales de interacción con el cliente, previamente guardados en bases separadas, como los centros de atención telefónica, chat en línea, Twitter, etc., para obtener una visión completa de la experiencia del cliente. Esto les permite comprender el canal de interacción que un cliente tiene y su impacto en otro con el fin de optimizar todo el ciclo de vida de la experiencia del cliente.
Monitoreo de la red	Las tecnologías Big Data se utilizan para capturar, analizar y visualizar los datos recogidos de los servidores, dispositivos de almacenamiento y otros equipos de TI para permitir a los administradores supervisar la actividad de red y diagnosticar cuellos de botella y otros problemas. Este tipo de análisis también se puede aplicar a otro tipo de redes. Por ejemplo las de transporte a los efectos de mejorar la eficiencia de combustible.
Investigación y Desarrollo	Las empresas, tales como fabricantes de productos farmacéuticos, utilizan Big Data para leer, en enormes volúmenes de datos de texto y otros datos históricos para ayudarlas en el desarrollo de nuevos productos.

Como se pudo ver en la Tabla 1.1, existen diversas áreas en las cuales el Big Data puede ser aplicado. La Figura 1.6 es un Diagrama de casos de uso en UML que muestra los usos del big data en distintas áreas.

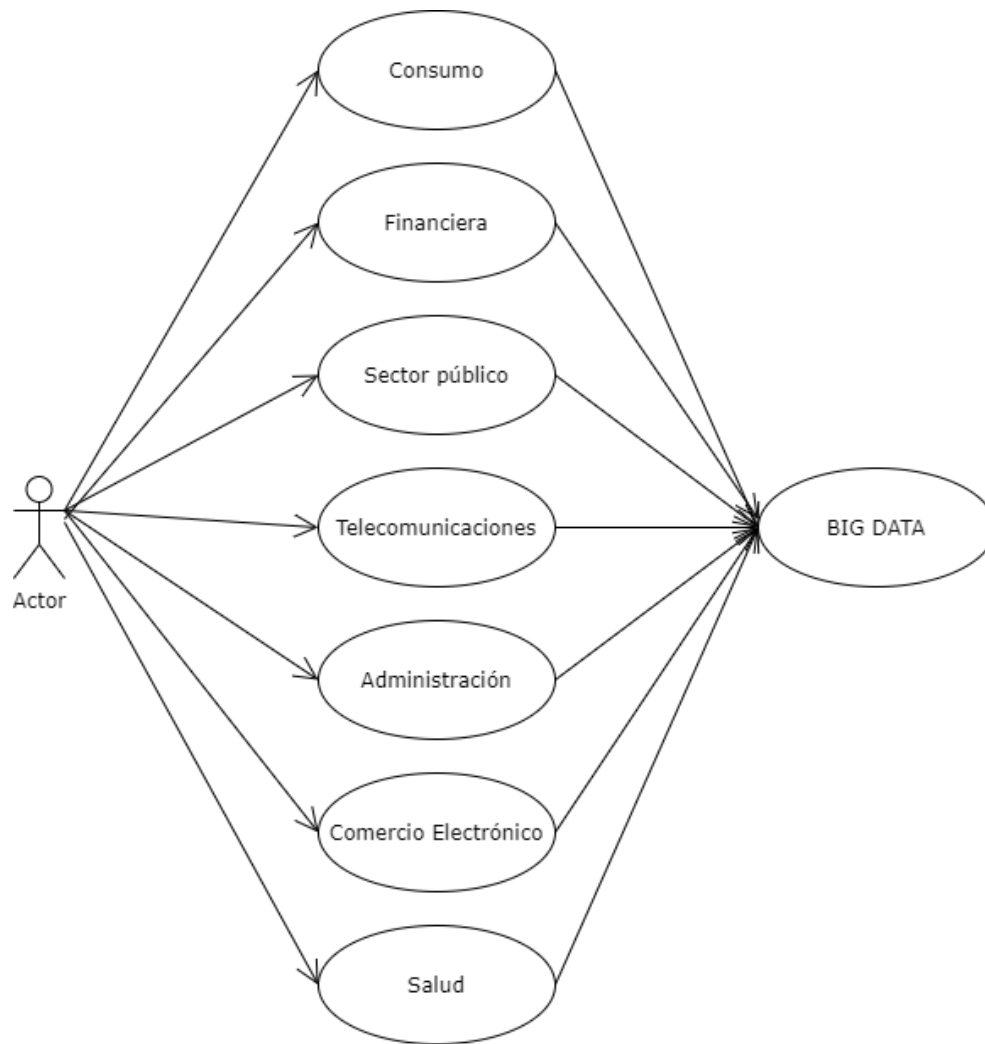


Figura 1.6: Diagrama UML de casos de uso del big data.

Capítulo 2

Técnicas de Modelado

En el Capítulo 1 se hizo mención a un Ciclo de vida genérico del análisis de datos (Sección 1.5), donde un par de sus etapas consisten en planificar y elaborar Modelos (Puntos 3 y 4) para el análisis de los datos. A continuación se presentarán las técnicas de modelado comúnmente utilizadas en la Ciencia de los Datos.

2.1. Redes Neuronales Artificiales.

Las Redes Neuronales Artificiales, ANN (Artificial Neuronal Networks) están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano.[12]

Las ANN al margen de “parecerse” al cerebro humano presentan una serie de características propias del cerebro. Por ejemplo las ANN aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos y abstraen las características principales de una serie de datos.

Algunos Conceptos aplicados a las ANN [12]:

- **Aprender:** Adquirir conocimientos de una cosa por medio del estudio, ejercicio o experiencia. Las ANN pueden cambiar su comportamiento en función del entorno. Se les muestra un conjunto de entradas y ellas mismas se ajustan para producir unas salidas consistentes.
- **Generalizar:** Extender o ampliar una cosa. Las ANN generalizan automáticamente debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.
- **Abstraer:** Aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas ANN son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos.

2.1.1. Estructura Básica de las Redes Neuronales Artificiales

En las Redes Neuronales Artificiales, ANN, la unidad análoga a la neurona biológica es el elemento procesador, PE (process element). Un elemento procesador tiene varias entradas y las combina, normalmente con una suma básica. La suma de las entradas es modificada por una función de transferencia y el valor de la salida de esta función de transferencia se pasa directamente a la salida del elemento procesador.

La salida del PE se puede conectar a las entradas de otras neuronas artificiales (PE) mediante conexiones ponderadas correspondientes a la eficacia de la sinapsis de las conexiones neuronales.[12]

La Figura 2.1 representa un elemento procesador de una red neuronal artificial.

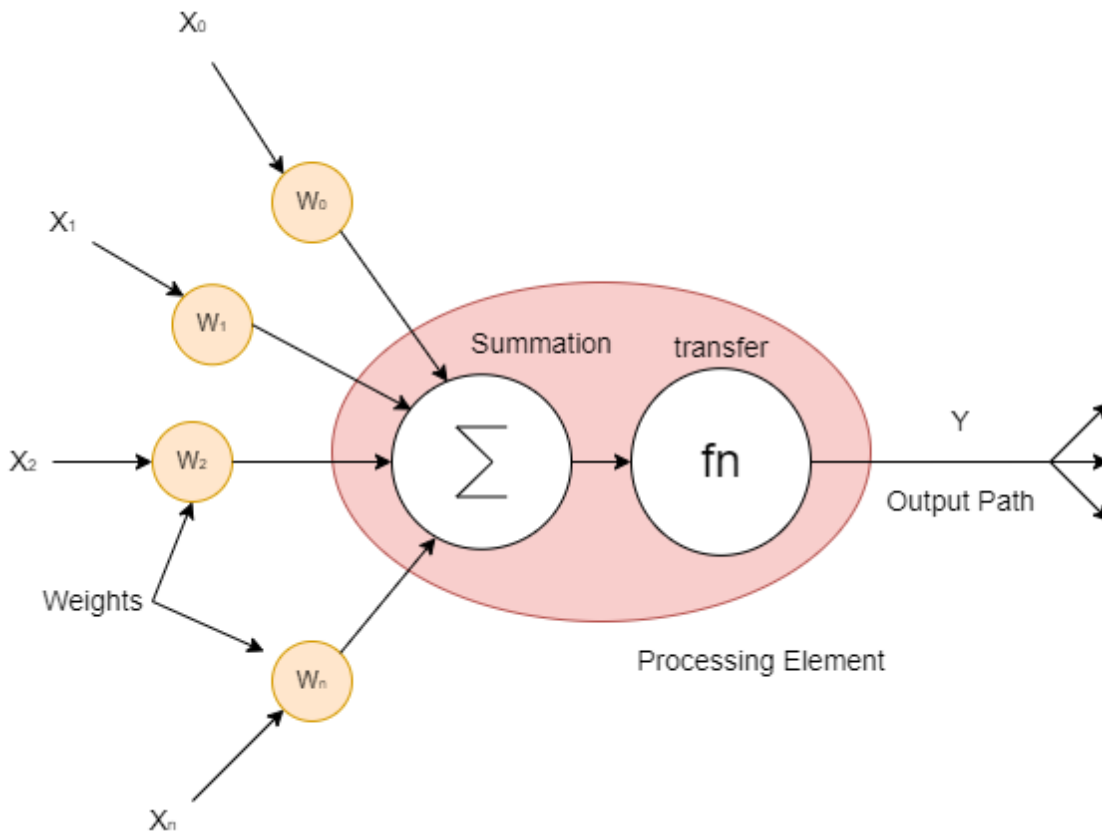


Figura 2.1: Diagrama de una Neurona Artificial (PE).

Una red Neuronal consiste en un conjunto de unidades elementales PE conectadas de una forma concreta. El interés de las ANN no reside solamente en el modelo del elemento PE sino en las formas en que se conectan estos elementos procesadores. Generalmente los elementos PE están organizados en grupos llamados niveles o capas. Una red típica consiste en una secuencia de capas con conexiones entre capas adyacentes consecutivas.

Existen dos capas con conexiones con el mundo exterior. Una capa de entrada, buffer de entrada, donde se presentan los datos a la red, y una capa buffer de salida que mantiene la respuesta de la red a una entrada. El resto de las capas reciben el nombre de capas ocultas. La Figura 2.2 muestra el aspecto de una Red Neuronal Artificial.[12]

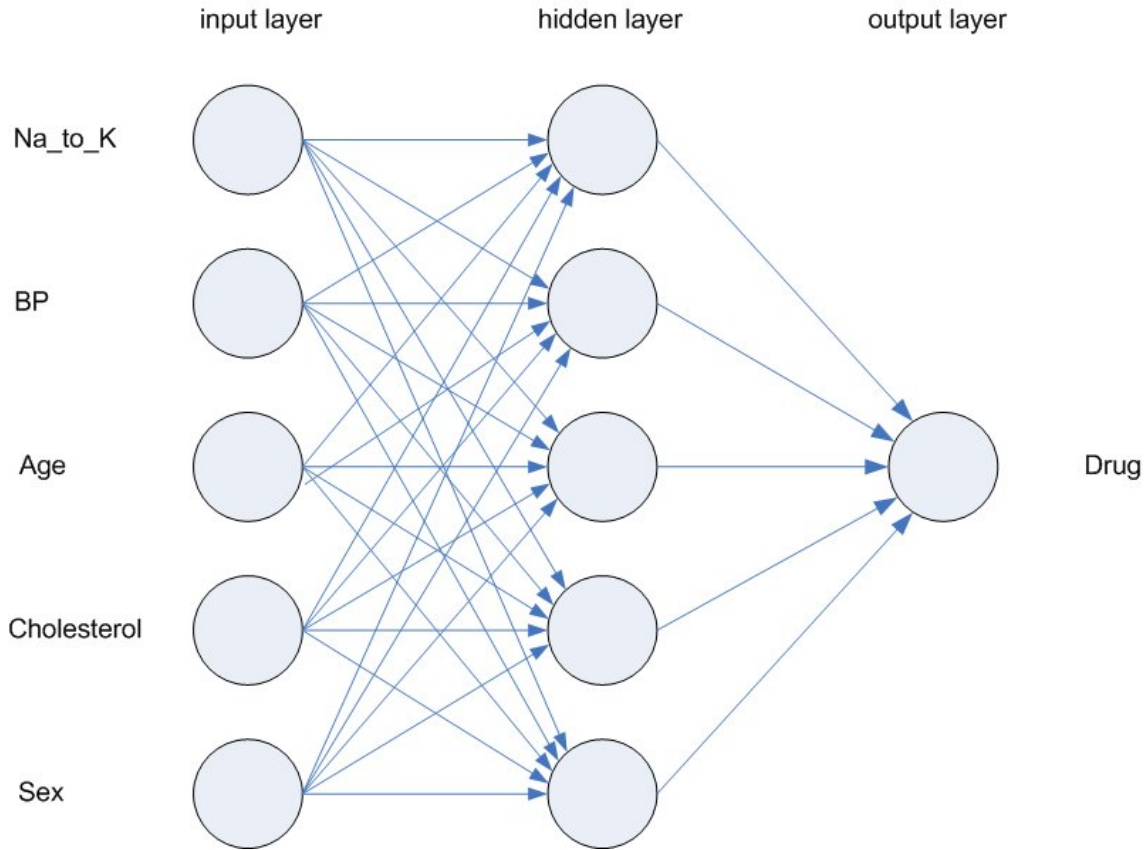


Figura 2.2: Arquitectura de una Red Neuronal Simple. [1]

2.1.2. Fundamentos de las Redes Neuronales Artificiales

La Neurona Artificial

La neurona artificial fue diseñada para “emular” las características del funcionamiento básico de la neurona biológica. En esencia, se aplica un conjunto de entradas a la neurona, cada una de las cuales representa una salida de otra neurona. Cada entrada se multiplica por su “peso” o ponderación correspondiente análogo al grado de conexión de la sinapsis. Todas las entradas ponderadas se suman y se determina el nivel de excitación o activación de la neurona. Una representación vectorial del funcionamiento básico de una neurona artificial se indica según la siguiente expresión de la Ecuación 2.1 [12]:

$$NET = X * W$$

Siendo NET la salida, X el vector de entrada y W el vector de pesos.

Normalmente la señal de salida NET suele ser procesada por una función de activación F para producir la señal de salida de la neurona OUT. La función F puede ser una función lineal, o una función umbral o una función no lineal que simula con mayor exactitud las características de transferencia no lineales de las neuronas biológicas.[12]

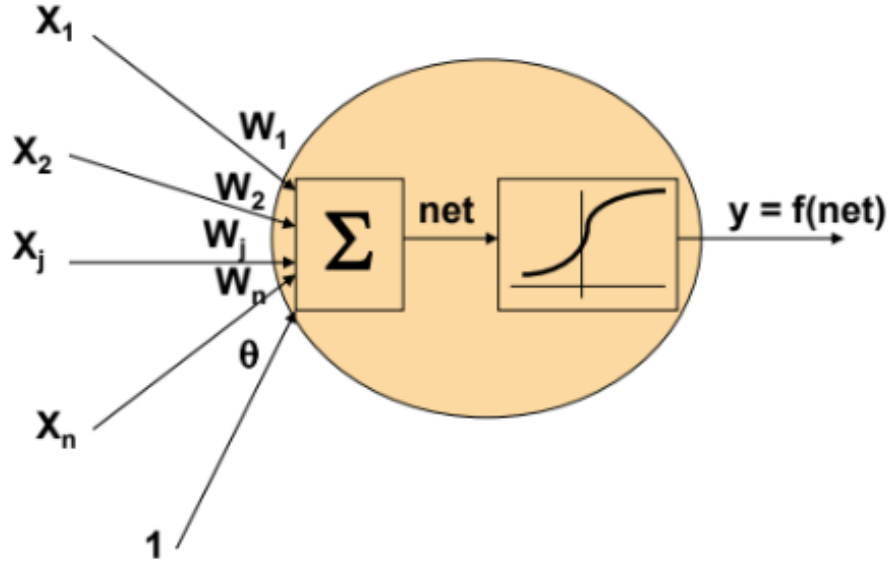


Figura 2.3: Modelo de Neurona Artificial. [2]

Donde:

- $\bar{W} = (W_1, \dots, W_n)$ es un vector de pesos, equivalente a las conexiones sinápticas en una neurona real.
- θ es el umbral de acción o activación.
- $\bar{X} = (X_1, \dots, X_n)$ es la entrada y el escalador.
- net es la suma ponderada entre el vector de entrada $\bar{X} = (X_1, \dots, X_n)$ y el vector de pesos $\bar{W} = (W_1, \dots, W_n)$ más un sesgo θ

$$net = \sum_{i=1}^n \bar{W}_i * \bar{X}_i + \theta$$

- $f()$ es la función de activación.
- y , es la salida generada a partir de la función de activación.

Las funciones $f()$ más utilizadas son la función Sigmoide y Tangente hiperbólica, estas son expresadas en la Tabla 2.1.

Tabla 2.1: Funciones más utilizadas en las Redes Neuronales Artificiales

APLICACIÓN	DESCRIPCIÓN
Sigmoide	$OUT = \frac{1}{(1+e^{-NET})}$
Tangente Hiperbólica	$OUT = \tanh NET$

Redes Neuronales Artificiales de una capa y multicapa

La capacidad de cálculo y potencia de la computación neuronal proviene de las múltiples conexiones de las neuronas artificiales que constituyen las redes ANN.

La red más simple es un grupo de neuronas ordenadas en una capa como se muestra en la siguiente figura (Figura 2.4). Los nodos circulares sólo son distribuidores y no se consideran constituyentes de una capa.

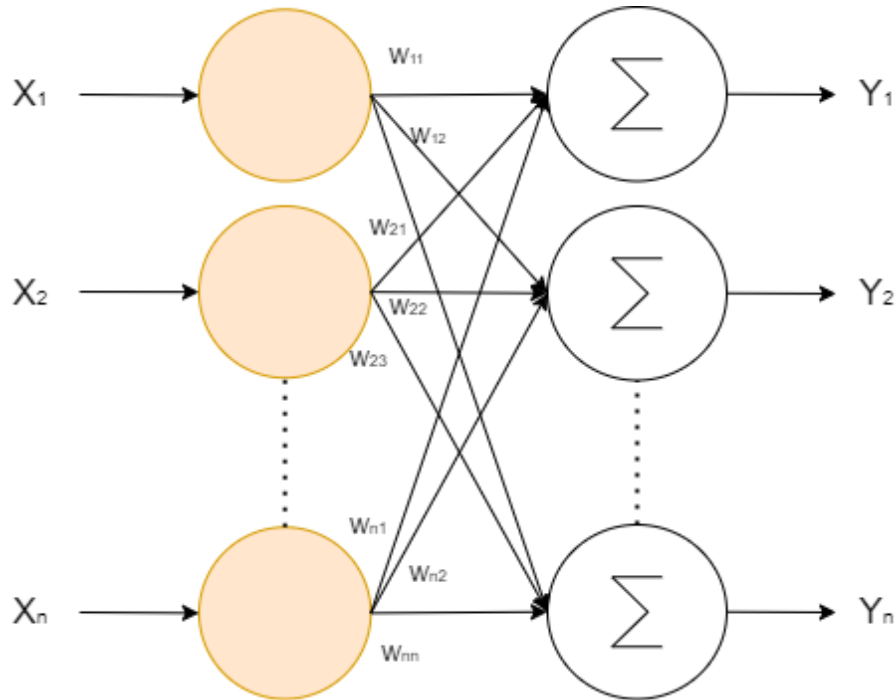


Figura 2.4: Red Neuronal de una Capa.

Cada una de las entradas está conectada a través de su peso correspondiente a cada neurona artificial. En la práctica existen conexiones eliminadas e incluso conexiones entre las salidas y entradas de las neuronas de una capa. No obstante la figura muestra una conectividad total por razones de generalización.

Normalmente las redes más complejas y más grandes ofrecen mejores prestaciones en el cálculo computacional que las redes simples. Las configuraciones de las redes construidas presentan aspectos muy diferentes pero tienen un aspecto común, el ordenamiento de las

neuronas en capas o niveles imitando la estructura de capas que presenta el cerebro en algunas partes.

Las redes multicapa se forman con un grupo de capas simples en cascada. La salida de una capa es la entrada de la siguiente capa. Se ha demostrado que las redes multicapa presentan cualidades y aspectos por encima de las redes de una capa simple. La Figura 2.5 muestra una red de dos capas.[12]

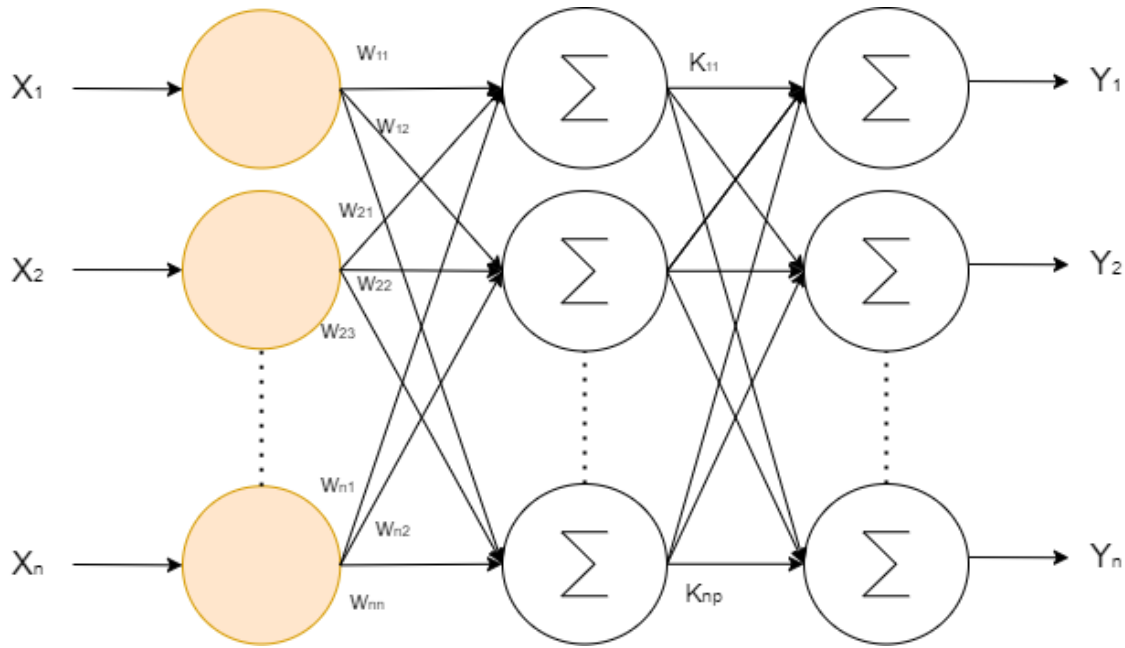


Figura 2.5: Red Neuronal de dos Capas.

Conviene destacar que la mejora de las redes multicapa estriba en la función de activación no lineal entre capas, pudiéndose llegar al caso de diseñar una red de una capa simple equivalente a una red multicapa si no se utiliza la función no lineal de activación entre capas.

Entrenamiento de las Redes Neuronales Artificiales

Una de las principales características de las ANN es su capacidad de aprendizaje. El objetivo del entrenamiento de una ANN es conseguir que una aplicación determinada, para un conjunto de entradas produzca el conjunto de salidas deseadas o mínimamente consistentes. El proceso de entrenamiento consiste en la aplicación secuencial de diferentes conjuntos o vectores de entrada para que se ajusten los pesos de las interconexiones según un procedimiento predeterminado. Durante la sesión de entrenamiento los pesos convergen gradualmente hacia los valores que hacen que cada entrada produzca el vector de salida deseado.[12]

Los algoritmos de entrenamiento o los procedimientos de ajuste de los valores de las co-

nexiones de las ANN se pueden clasificar en dos grupos: Supervisado y No Supervisado.[12]

- **Entrenamiento Supervisado:** estos algoritmos requieren el emparejamiento de cada vector de entrada con su correspondiente vector de salida. El entrenamiento consiste en presentar un vector de entrada a la red, calcular la salida de la red, compararla con la salida deseada, 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.

Las parejas de vectores del conjunto de entrenamiento se aplican secuencialmente y de forma cíclica. Se calcula el error y el ajuste de los pesos por cada pareja hasta que el error para el conjunto de entrenamiento entero sea un valor pequeño y aceptable.

- **Entrenamiento No Supervisado:** Los sistemas de aprendizaje no supervisado no requieren de un vector de salidas deseadas y por tanto no se realizan comparaciones entre las salidas reales y salidas esperadas. El conjunto de vectores de entrenamiento consiste únicamente en vectores de entrada. El algoritmo de entrenamiento modifica los pesos de la red de forma que produzca vectores de salida consistentes. El proceso de entrenamiento extrae las propiedades estadísticas del conjunto de vectores de entrenamiento y agrupa en clases los vectores similares.

2.1.3. Tipos y selección de la Redes Neuronales Artificiales

La selección de una red se realiza en función de las características del problema a resolver. La mayoría de estos se pueden clasificar en aplicaciones de Predicción, Clasificación, Asociación, Conceptualización, Filtrado y Optimización. Los tres primeros tipos de aplicaciones requieren un entrenamiento supervisado.

Tabla 2.2: Clasificación de las ANN

[12]

NOMBRE DE LA RED	CARACTERÍSTICAS	TIPO
Adaline y Madaline	Técnicas de Adaptación para el Reconocimiento de Patrones.	Predicción
Adaptive Resonance Theory Networks (ART)	Reconocimiento de Patrones y Modelo del Sistema Neuronal. Concepto de Resonancia Adaptativa.	Conceptualización
Back-Propagation	Solución a las limitaciones de su red predecesora el Perceptron.	Clasificación
Bi-Directional Associative Memory (BAM) Networks	Inspirada en la red ART.	Asociación
The Boltzmann Machine	Similar a la red Hopfield.	Asociación
Brain-State-in a Box	Red Asociativa Lineal.	Asociación

Cascade-Correlation-Networks	Adición de nuevas capas ocultas en cascada.	Asociación
Counter-Propagation	Clasificación Adaptativa de Patrones.	Clasificación
Delta-Bar-Delta (DBD) Networks	Métodos Heurísticos para Acelerar la Convergencia.	Clasificación
Digital Neural Network Architecture (DNNA) Networks	Implementación Hardware de la función Sigmoid.	Predicción
Directed Random Search (DRS) Networks	Técnica de valores Random en el mecanismo de Ajuste de Pesos.	Clasificación
Functional-link Networks (FLN)	Versión mejorada de la red Backpropagation.	Clasificación
Hamming Networks	Clasificador de vectores binarios utilizando la Distancia Hamming.	Asociación
Hopfield Networks	Concepto de la red en términos de energía.	Optimización
Learning Vector Quantization (LVQ) Networks	Red Clasificadora.	Clasificación
Perceptron Networks	Primer modelo de sistema Neuronal Artificial.	Predicción
Probabilistic Neural Network (PNN)	Clasificación de Patrones utilizando métodos estadísticos.	Asociación
Recirculation Networks	Alternativa a la red Backpropagation.	Filtrado
Self-Organizing Maps (SOM)	Aprendizaje sin supervisión.	Conceptualización
Spatio-Temporal-Pattern Recognition (SPR)	Red clasificatoria Invariante en el espacio y tiempo.	Asociación

2.2. Algoritmos de Regresión Lineal

La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuestas continúa 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.[3]

Las técnicas de regresión lineal permiten crear un modelo lineal. Este modelo describe la relación entre una variable dependiente y (también conocida como la respuesta) como una función de una varias variables independientes X_i (denominadas predictores). La ecuación

general correspondiente a un modelo de regresión lineal es:

$$Y = \beta_0 + \sum \beta_i X_i + \varepsilon_i$$

Donde β representa las estimaciones de parámetros lineales que se deben calcular y representa los términos de error.[3]

2.2.1. Tipos de regresión lineal

Existen diferentes tipos de regresión lineal según el predictor y las variables de respuestas. A continuación se presentan algunos tipos de regresión lineal.[3]

Regresión lineal simple: Modelo que utiliza un único predictor. La ecuación general es:

$$Y = \beta_0 + \beta_i X + \varepsilon_i$$

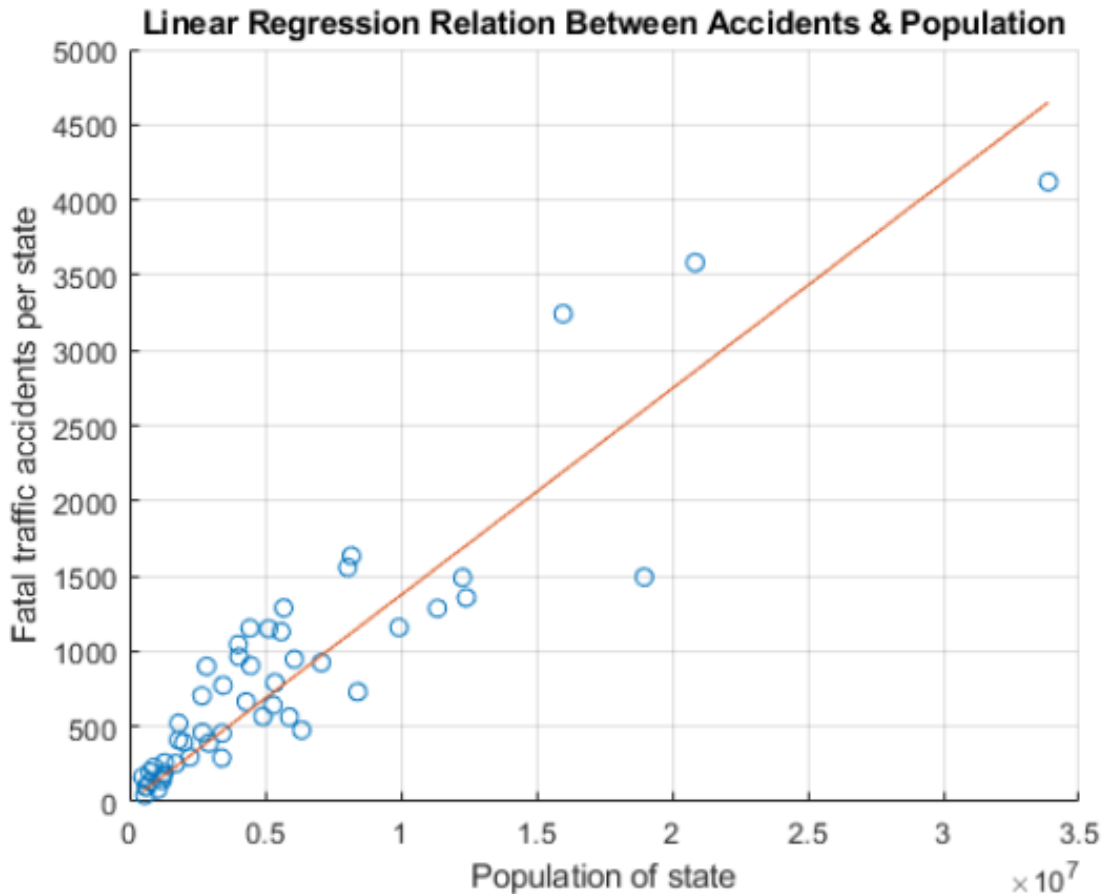


Figura 2.6: Ej. Regresión Lineal Simple. [3]

Regresión lineal múltiple: Modelo que utiliza múltiples predictores. Esta regresión

tiene múltiples X_1 para predecir la respuesta, Y . Este es un ejemplo de la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

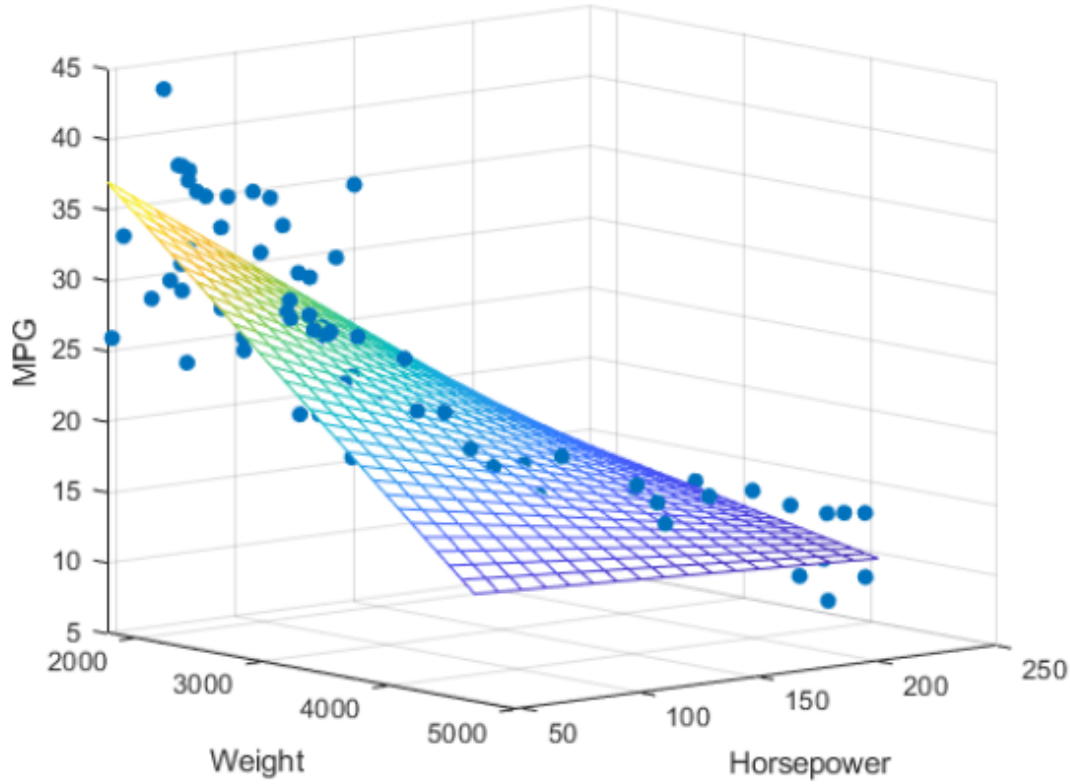


Figura 2.7: Ej. Regresión Lineal Múltiple. [3]

Regresión lineal multivariante: Modelo para varias variables de respuesta. Esta regresión tiene múltiples Y_i que derivan de los mismos datos Y . Se expresan con fórmulas diferentes. Este es un ejemplo del sistema con 2 ecuaciones:

$$Y_1 = \beta_{01} + \beta_{11} X_1 + \varepsilon$$

$$Y_2 = \beta_{02} + \beta_{12} X_1 + \varepsilon_2$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \beta_{01} & \beta_{11} \\ \beta_{02} & \beta_{12} \end{pmatrix} \begin{pmatrix} 1 \\ X_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

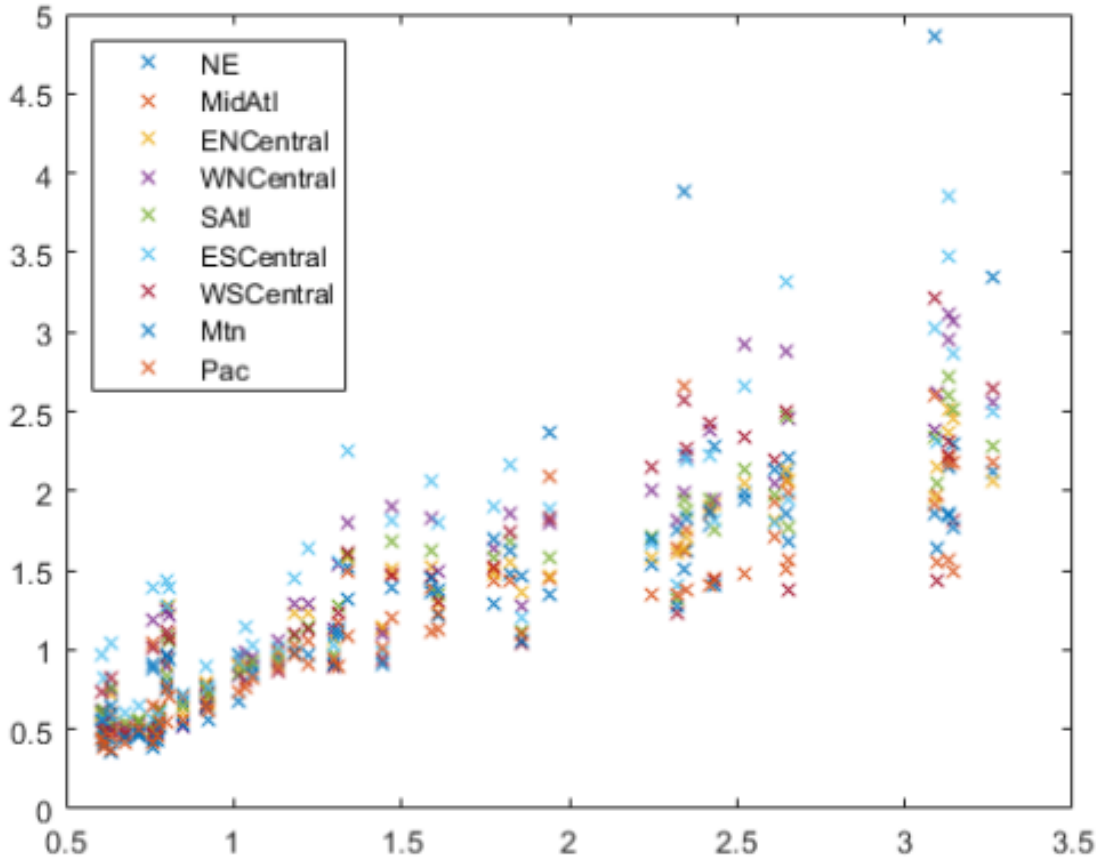


Figura 2.8: Ej. Regresión Lineal Múltivariante. [3]

Regresión lineal múltiple multivariante: Modelo que utiliza varios predictores para múltiples variables de respuesta. Esta regresión tiene múltiples X_1 para predecir varias respuestas Y . Esta es una generalización de las ecuaciones:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ 1 & X_{31} & X_{32} & \cdots & X_{3q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

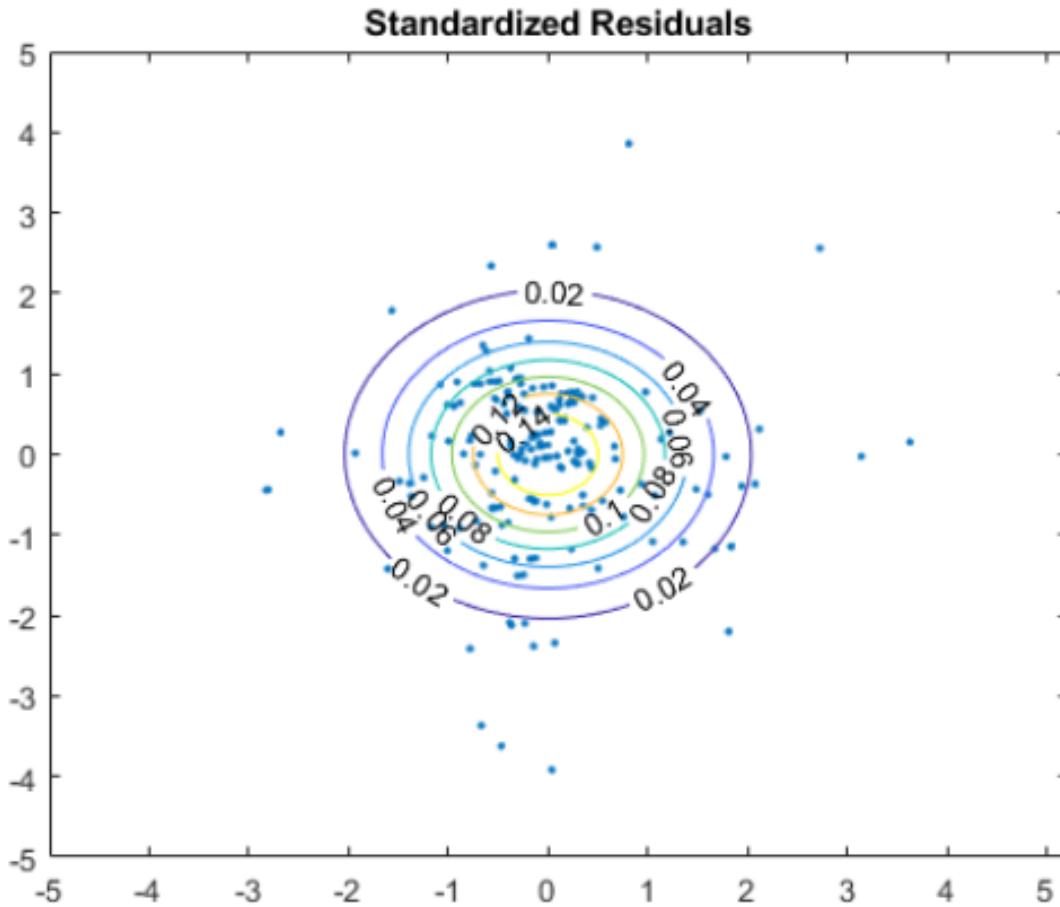


Figura 2.9: Ej. Regresión Lineal Múltiple Multivariante. [3]

2.2.2. Aplicaciones de la regresión lineal

La regresión lineal cuenta con ciertas características ideales para las siguientes aplicaciones [3]:

- **Predicción o pronóstico:** Se puede utilizar un modelo de regresión para crear un modelo de pronóstico para un conjunto de datos específico. A partir de la moda, se puede usar la regresión para predecir valores de respuesta donde solo se conocen los predictores.
- **Fuerza de la regresión:** Se puede utilizar un modelo de regresión para determinar si existe una relación entre una variable y un predictor, y cuán estrecha es esta relación.

2.3. Regresión Logística

La Regresión Logística es un método estadístico para predecir clases binarias. Tiene ciertas similitudes en su planteamiento con la regresión lineal, pero está orientada a resol-

ver problemas de clasificación y no de predicción.

Este es uno de los algoritmos de Machine Learning más simples. Es fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria.

La Regresión Logística describe y estima la relación entre una variable binaria dependiente y las variables independientes. Lleva el nombre de la función utilizada en el núcleo del método (función logística), esta función es también llamada función Sigmoide. La cual es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1.[4]

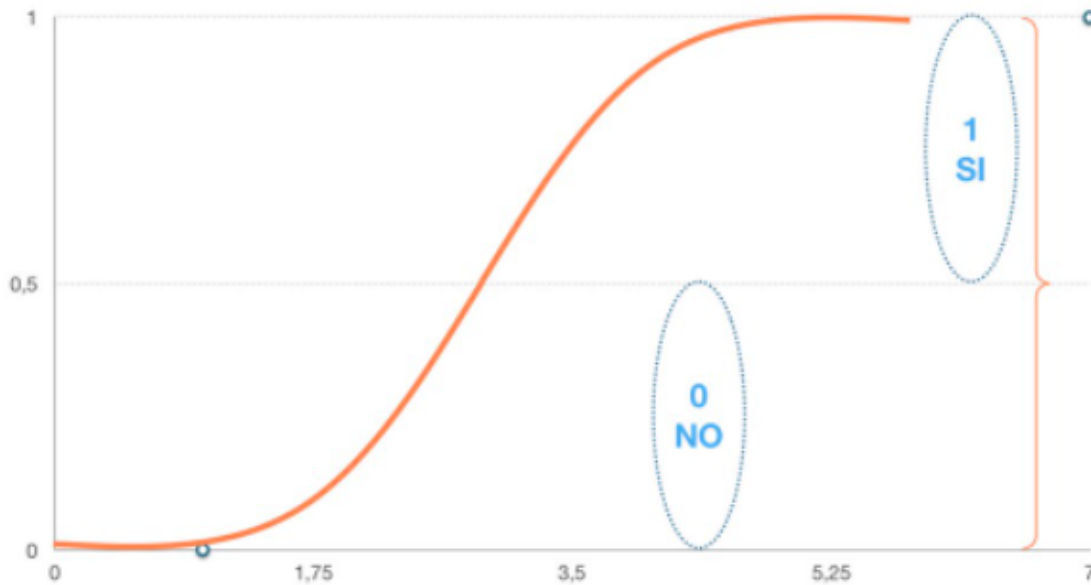


Figura 2.10: Ej. Función Logística. [4]

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO. Por su parte si el resultado es 0.75, podemos decir en términos de probabilidad como, hay un 75 % de probabilidades de que el paciente sufra cáncer.[4]

La ecuación que produce la función Sigmoide es la siguiente:

$$\text{Sigmoide} = \frac{1}{1 + e^{-y}}$$

2.3.1. Diferencias entre Regresión Lineal y Regresión Logística

La Regresión Lineal proporciona una salida continua, pero la Regresión Logística proporciona una salida discreta. Un ejemplo de una salida continua es conocer el porcentaje

de probabilidad de lluvia o el precio de una acción. Un ejemplo de una salida discreta, por su parte, es conocer si va a llover o no, o si el precio de una acción subirá o no.[4]

2.3.2. Tipos de Regresión Logística

La regresión logística también tiene distintos tipos según la variable objetivo, algunos tipos son los siguientes[4]:

Regresión Logística Binaria: La variable objetivo tiene solo dos resultados posibles, Llueve o NO Llueve, Sube o Baja.

Regresión Logística Multinomial: La variable objetivo tiene tres o más categorías nominales, como predecir el tipo de vino.

Regresión Logística Ordinal: La variable objetivo tiene tres o más categorías ordinales, como clasificar un restaurante o un producto del 1 al 5.

2.4. Algoritmo K-Nearest Neighbors

KNN es un método de clasificación supervisada que sirve para estimar la función de densidad:

$$f\left(\frac{x}{C_j}\right)$$

Donde X es la variable independiente y C_j la clase j , por lo que la función determina la probabilidad a posteriori de que la variable X pertenezca a la clase j .

En el reconocimiento de patrones, el algoritmo KNN es utilizado como método de clasificación de objetos con un entrenamiento a través de ejemplos cercanos en el espacio de diversos elementos. Cada elemento está descrito en términos de P atributos considerando Q clases para la clasificación.[5]

El espacio de los valores de la variable independiente es particionada en regiones por localizaciones y etiquetas de los elementos de entrenamiento. De esta forma un punto en el espacio es asignado a la clase C , si ésta es la clase más frecuente entre los k elementos más cercanos.[5]

Para determinar la cercanía de los elementos se utiliza comúnmente la distancia euclidiana:

$$d(X_i, X_j) = \sqrt{\sum_{r=1}^p (X_{ri} - X_{rj})^2}$$

La **fase de entrenamiento** consiste en almacenar los vectores característicos y las etiquetas de las clases de dichos elementos de entrenamiento.

En la **fase de clasificación** se calcula la distancia entre los vectores almacenados y el nuevo vector y se selecciona los k elementos más cercanos.

El **nuevo vector** es clasificado con la clase que más se repite en los vectores seleccionados.

2.4.1. Aplicación del Algoritmo

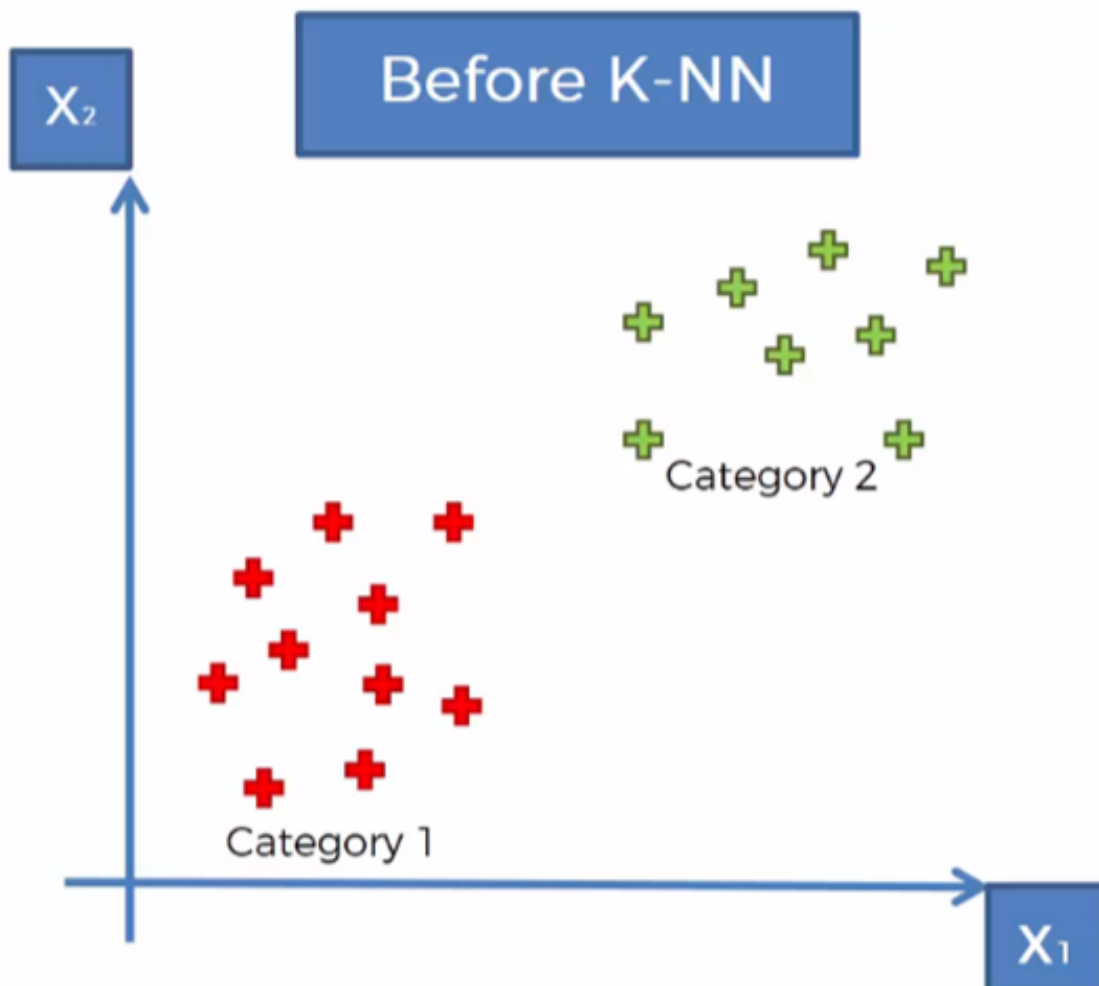


Figura 2.11: Ej. Conjunto de datos clasificados en dos categorías. [5]

Considerando un conjunto de datos clasificados en dos categorías, como se muestra en la Figura 2.11, se requiere clasificar un nuevo vector de datos que se encuentra en la región mostrada en la siguiente gráfica (Figura 2.12).

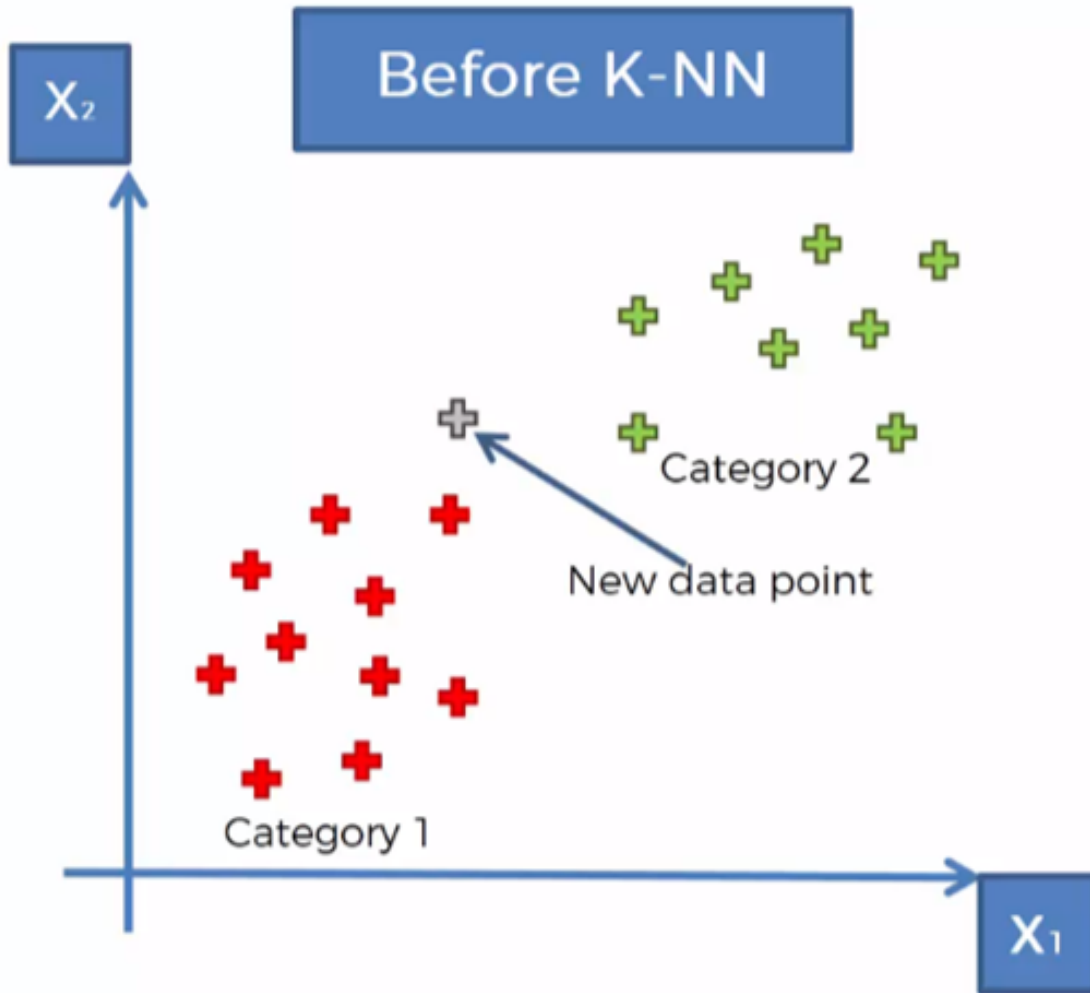


Figura 2.12: Ej. Clasificación de un nuevo vector de datos. [5]

El algoritmo KNN sigue los siguientes pasos para determinar a qué categoría pertenece el nuevo dato que se desea clasificar [5]:

- **Paso 1:** Selecciona el número de K vecinos.
- **Paso 2:** Toma los K vecinos más cercanos al nuevo elemento de acuerdo con la distancia euclidiana.
- **Paso 3:** Entre los K vecinos, contar el número de elementos que pertenece a cada categoría.
- **Paso 4:** Asignar el nuevo elemento a la categoría donde se contaron más vecinos.

Tomando para el ejemplo que $K = 5$, marcamos los 5 vecinos más cercanos al nuevo elemento.



Figura 2.13: Ej. Selección de los $K=5$ vecinos más cercanos al nuevo elemento. [5]

En la Figura 2.13 contamos que existen 3 elementos de la categoría 1 y dos elementos de la categoría 2 de entre los 5 vecinos más cercanos.



Figura 2.14: Ej. 3 elementos de la categoría 1 y 2 elementos de la categoría 2. [5]

Por lo tanto, la categoría con más elementos contados es la categoría 1, por lo que el nuevo elemento se asigna a la categoría 1.



Figura 2.15: Ej. 3 elementos de la categoría 1 y 2 elementos de la categoría 2. [5]

Como se observa en la Figura 2.15, el nuevo elemento se asignó a la categoría 1, dado que para $K = 5$ existen más vecinos de dicha categoría.

2.5. Árboles de Decisión

Cuando se analizan decisiones en condiciones de riesgo, el diagrama del árbol es un instrumento gráfico que obliga a la persona que toma las decisiones a “examinar todos los resultados posibles, incluidos los desfavorables. También la obliga a tomar decisiones de una manera lógica y consecutiva”. [13]

Los árboles de decisión son especialmente útiles cuando debe tomarse una sucesión de decisiones.

El análisis de un árbol de decisiones se basa en la teoría de probabilidades.

2.5.1. Ventajas del Árbol de decisiones

Los árboles de decisiones no ofrecen grandes ventajas como [13]:

- Claramente plantean el problema porque todas las opciones sean analizadas.
- Permiten analizar totalmente las posibles consecuencias de tomar una decisión.
- Proveen un esquema para cuantificar el costo de un resultado y la probabilidad de que suceda.
- Nos ayuda a tomar las mejores decisiones sobre la base de la información existente y de las mejores suposiciones.

- Provee una estructura sumamente efectiva dentro de la cual se puede estimar cuales son las opciones e investigar las posibles consecuencias de seleccionar cada una de ellas.
- También ayuda a construir una imagen balanceada de los riesgos y recompensas asociados con cada posible curso de acción.

2.5.2. Tipos de nodos

Un árbol de decisión se compone de distintos tipos de nodos, según su representación gráfica se cuenta con dos tipos de nodos los cuales son [13]:

- **Nodo de Decisión o de acción**, que se representará por medio de un cuadrado o rectángulo.

Estos cuadros indican que debe tomarse una decisión.

Simbolizan puntos de decisión, donde el tomador de decisiones debe elegir entre varias acciones posibles.

De estos nodos de decisión, sale una rama para cada acción posible.

- **Nodo de suceso o de probabilidad**, que se representará por medio de un círculo.

Estos empalmes circulares, de los que sale ramas representan un estado de la naturaleza posible, al que se asigna la probabilidad correspondiente.

Los círculos representan eventos aleatorios, donde ocurre algún estado de la naturaleza.

Estos eventos aleatorios no están bajo el control del tomador de decisiones.

De estos nodos aleatorios sale una rama para cada resultado posible.

Todos los árboles de decisión son parecidos a su estructura y tienen los mismos componentes. Para ser más específicos se requieren los siguientes componentes [13]:

- **Alternativas** de decisión en cada punto de decisión.
- **Eventos** que pueden ocurrir como resultados de cada alternativa de decisión.
- **Probabilidades** de que ocurran los eventos posibles como resultados de las decisiones.

- **Resultados** de las posibles interacciones entre las alternativas de decisión y los eventos.

La Figura 2.16 es la representación de un árbol de decisión y sus componentes.

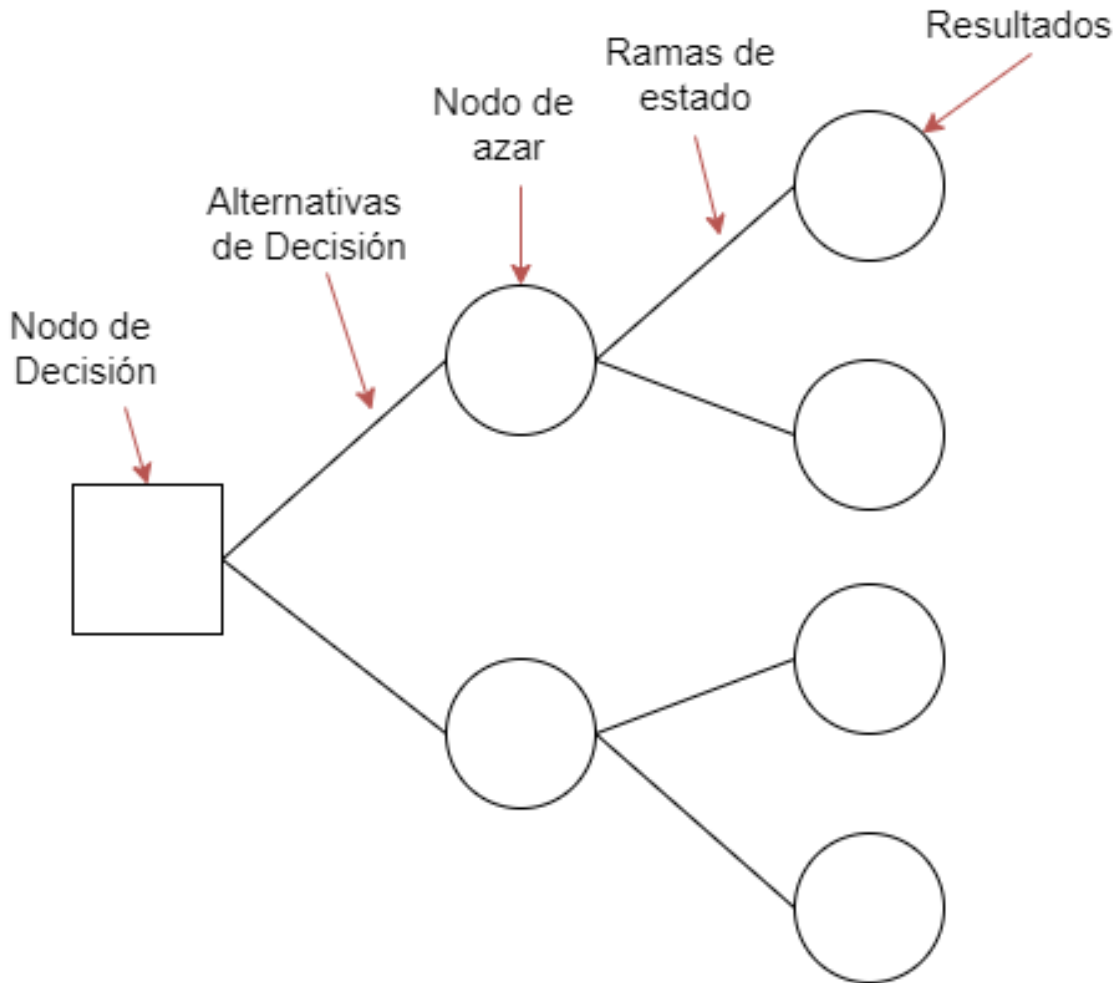


Figura 2.16: Estructura de un Árbol de Decisión.

Para iniciar el análisis del árbol de decisiones, tenemos dos reglas que dirigen este proceso [13]:

1. Si estamos analizando un nodo aleatorio (círculo), calculamos el valor esperado en ese nodo multiplicando la probabilidad en cada rama que sale por la ganancia al final de esta rama y luego sumando los productos de todas las ramas que salen del nodo.
2. Si estamos analizando un nodo de decisión (cuadrado), el valor esperado de ese nodo será el máximo de los valores esperados de todas las ramas que salen del nodo.

De esta forma, elegimos la acción con el mayor valor esperado y podamos las ramas que corresponden a las acciones menos rentables.

Capítulo 3

Área de Estudio: Finanzas

3.1. Definición

Definición técnica, “Las finanzas corresponden a un área de la economía que estudia la obtención y administración del dinero y el capital, es decir, los recursos financieros. Estudia tanto la obtención de esos recursos (financiación), así como la inversión y el ahorro de los mismos”. [14]

Las finanzas estudian cómo los agentes económicos (empresas, familias o Estado) deben tomar decisiones de inversión, ahorro y gasto en condiciones de incertidumbre. Al momento de elegir, los agentes pueden optar por diversos tipos de recursos financieros tales como: dinero, bonos, acciones o derivados, incluyendo la compra de bienes de capital como maquinarias, edificios y otras infraestructuras.

Las finanzas ayudan a controlar los ingresos y gastos, tanto al Gobierno, a las empresas, como a cada uno de nosotros. Tener un buen control de las finanzas nos permite gestionar mejor nuestros recursos, conociendo al detalle todos los ingresos y gastos, para tener un mayor control sobre ellos mismos. [14]

3.2. Tipos de finanzas

Las finanzas pueden dividirse en cuatro grandes grupos:

3.2.1. Finanzas corporativas

Las finanzas corporativas sirven para tomar decisiones sobre la retribución a los accionistas, las fuentes de financiamiento para la empresa y sus proyectos, el nivel de endeudamiento, la optimización del flujo de efectivo, la viabilidad de un proyecto de inversión, el modelo financiero, las fusiones y adquisiciones, e incluso influyen en rubros como la responsabilidad social corporativa y las política de incentivos a los colaboradores.

Las decisiones que se toman a partir de las finanzas corporativas pueden dividirse en [15]:

- **Decisiones de inversión:** Deben partir de un estudio detallado de las necesidades de la empresa para determinar los activos reales en los que se debe invertir para lograr un aumento en las ganancias. La inversión puede darse en stock, maquinaria o equipo, etc.
- **Decisiones de financiamiento:** Se trata de estudiar de dónde saldrán los fondos para las inversiones que se quiere hacer. Con estas decisiones se determina si es conveniente utilizar recursos propios o conviene solicitar créditos.
- **Decisiones sobre dividendos:** Trata de decidir cómo se hará la retribución a los accionistas, siempre buscando que se mantenga el equilibrio financiero de la empresa.
- **Decisiones directivas:** El análisis del estado financiero de la empresa ayuda a tomar decisiones estratégicas y de operación, por ejemplo: aumentos de sueldo, apertura de nuevas líneas de producción, etc.

La Figura 3.1 modela las finanzas corporativas mediante un diagrama UML.

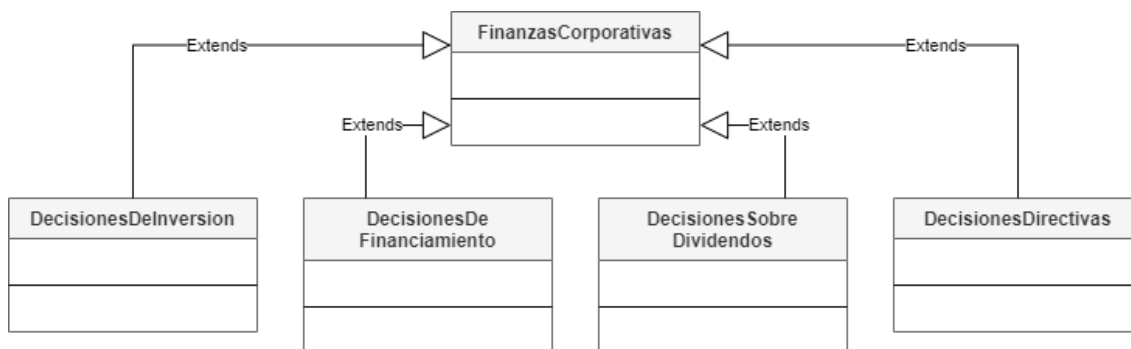


Figura 3.1: Diagrama UML, finanzas corporativas.

3.2.2. Finanzas personales

Las finanzas personales se ocupan de cómo los individuos o familias administran sus recursos a lo largo de su vida. En su análisis se incluyen no solo los ingresos y gastos recibidos o pagados durante la vida, sino también las herramientas o productos financieros con los que cuentan los individuos o familias para optimizar el manejo de sus recursos.

Su objetivo principal es ayudar a las personas y familias a que tomen decisiones informadas que permitan optimizar el manejo de sus recursos, esto contempla poder alcanzar una serie de subobjetivos como [16]:

- **Protección:** Contar con una protección adecuada ante riesgos o imprevistos.

- **Inversión:** Lograr acumular o conseguir suficientes recursos para poder invertir en activos que afectan positivamente la calidad de vida, pero que son costosos, por ejemplo, adquirir un coche o una casa, iniciar un negocio propio, o financiar una carrera universitaria.
- **Jubilación:** Mantener los recursos suficientes para poder vivir bien en la etapa de vida en donde se deje de trabajar.
- **Liquidez:** Contar con los recursos para financiar nuestras actividades cotidianas.

La Figura 3.2 modela las finanzas personales mediante un diagrama UML.

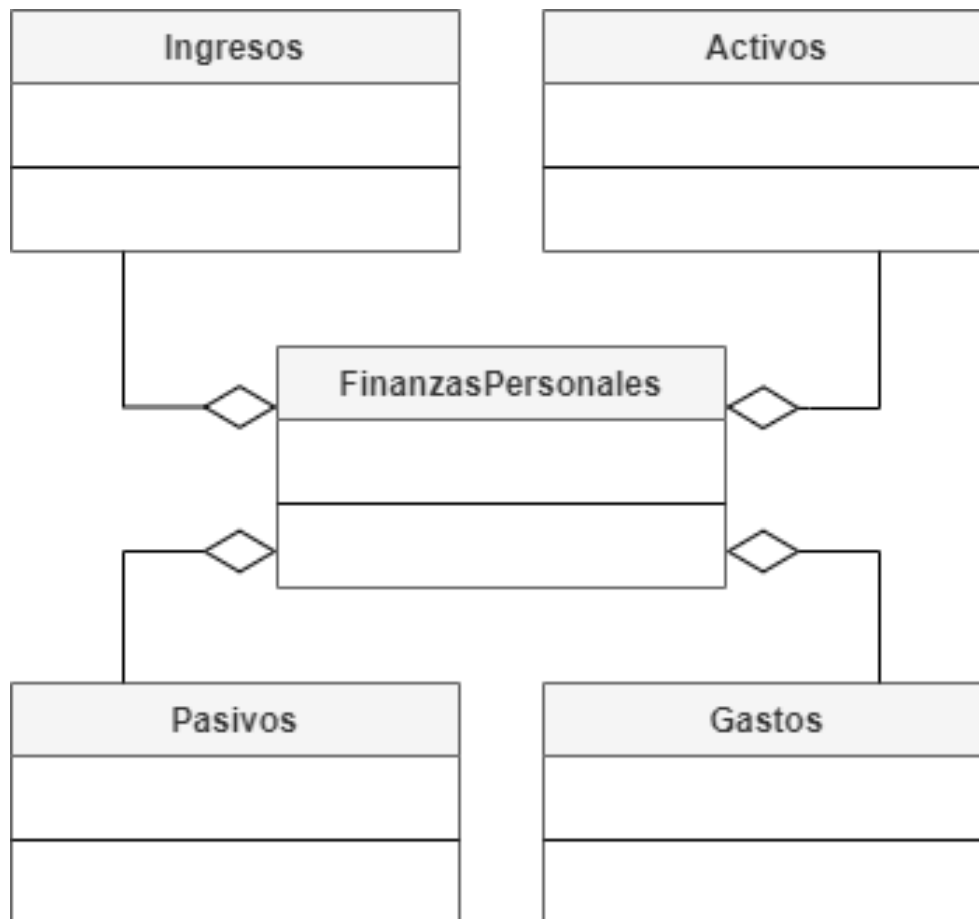


Figura 3.2: Diagrama UML, finanzas personales.

3.2.3. Finanzas públicas

Las finanzas públicas son la disciplina que se enfoca en la obtención de ingresos, realización de gastos y gestión de la deuda pública de un Estado. Se centra en dos frentes que son competencia del Gobierno: la recaudación de impuestos y el gasto público.

Parte de las tareas de las finanzas públicas es definir las herramientas de financiamiento del Estado. Entre sus objetivos principales debe estar tener un presupuesto público sostenible en el tiempo. En otras palabras, el plan consiste en no generar una deuda pública que en el largo plazo obligue a elevar impuestos o a recortar beneficios a los ciudadanos. [17]

- **Ingreso Público:** Es la cantidad total de recursos que reciben los organismos e instituciones que son manejadas por el Estado.

Este ingreso proviene de Tributos (impuestos, tasas, contribuciones especiales), Cotizaciones, Contractuales (ingresos derivados de la celebración de un contrato), Deuda pública, y voluntarios. [18]

- **Gasto Público:** Es la cuantía monetaria total que desembolsa el sector público para desarrollar sus actividades. Entre sus objetivos se encuentran distribuir la riqueza, mejorar el acceso a la salud, asegurar la justicia, mejorar el empleo, fomentar el crecimiento económico, garantizar una vida digna, fuerzas armadas, etc.

Algunos tipos de gastos públicos son: Gasto corriente, Gasto de capital, Gasto de transferencia, Gasto de inversión. [19]

La Figura 3.3 modela las finanzas públicas mediante un diagrama UML.

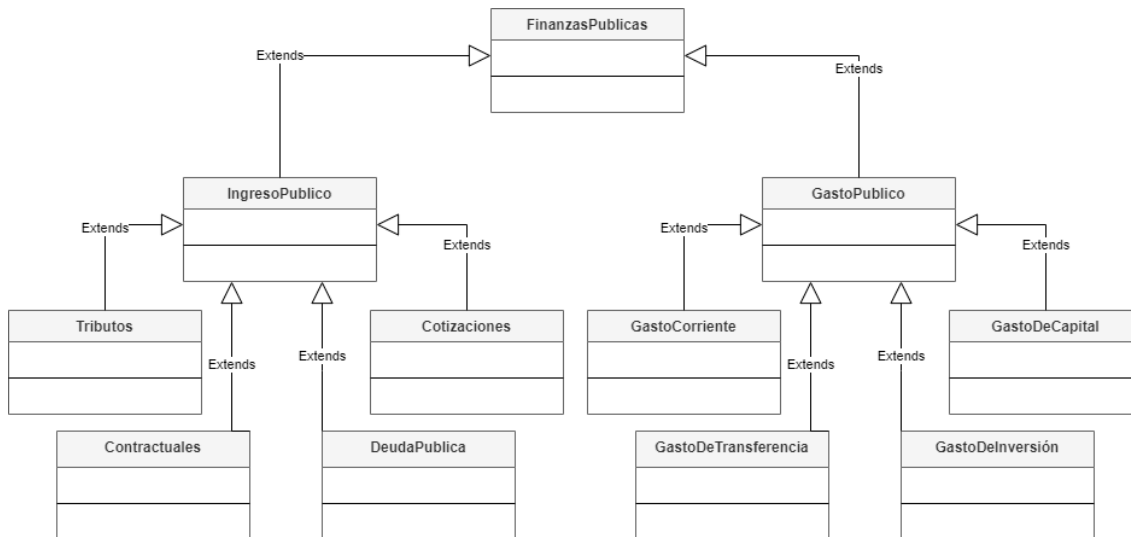


Figura 3.3: Diagrama UML, Finanzas Públicas.

3.2.4. Finanzas Internacionales

Es la rama que estudia y analiza las interrelaciones monetarias y macroeconómicas entre dos o más países, como los sistemas financieros, la balanza de pagos, las inversiones directas o los tipos de cambio. Esta disciplina se divide en dos ramas [20] :

- **La economía internacional:** Estudia las relaciones generales entre países, que incluye aspectos financieros y comerciales, como la mayor parte del mundo son economías abiertas están relacionadas con las demás, por lo tanto estudia cómo intercambian bienes o servicios y la forma en que se financian estas actividades.
- **Las finanzas corporativas:** Esta rama se centra en las relaciones internacionales entre empresas. En este caso, el objetivo es saber cómo maximizar el valor para propietarios y accionistas de una empresa que opere en mercados internacionales. Está íntimamente relacionado, entre otros, con las inversiones directas de las empresas en los países.

La Figura 3.4 modela las finanzas internacionales mediante un diagrama UML.

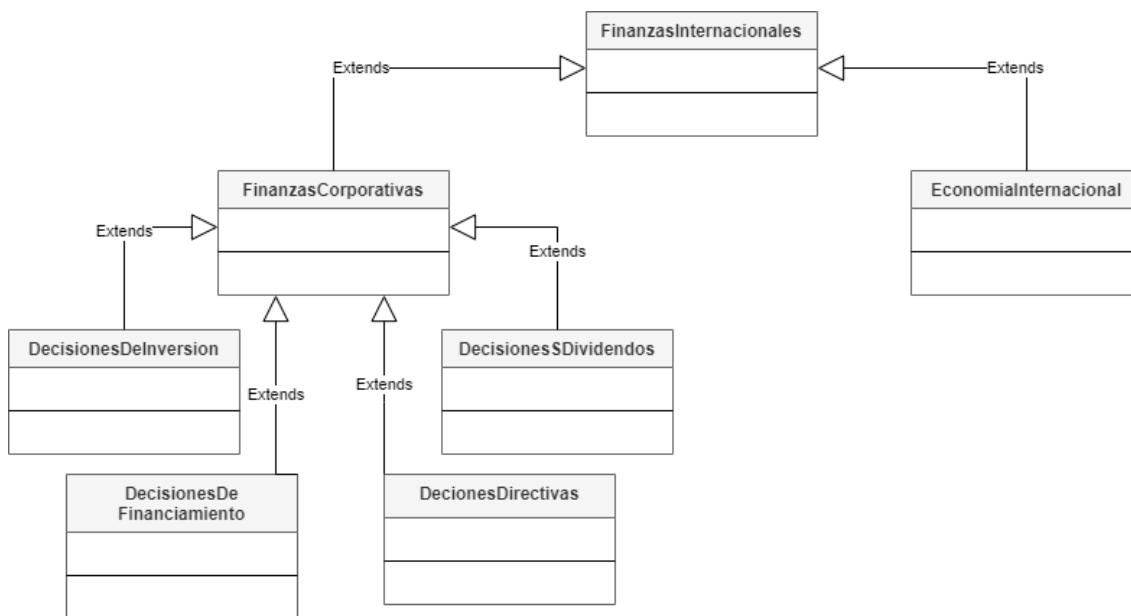


Figura 3.4: Diagrama UML, Finanzas Internacionales.

Uniendo las Figuras 3.1, 3.2, 3.3, 3.4 sobre los tipos de finanzas, obtenemos una vista global sobre todo el área de las finanzas, representado por un diagrama UML.

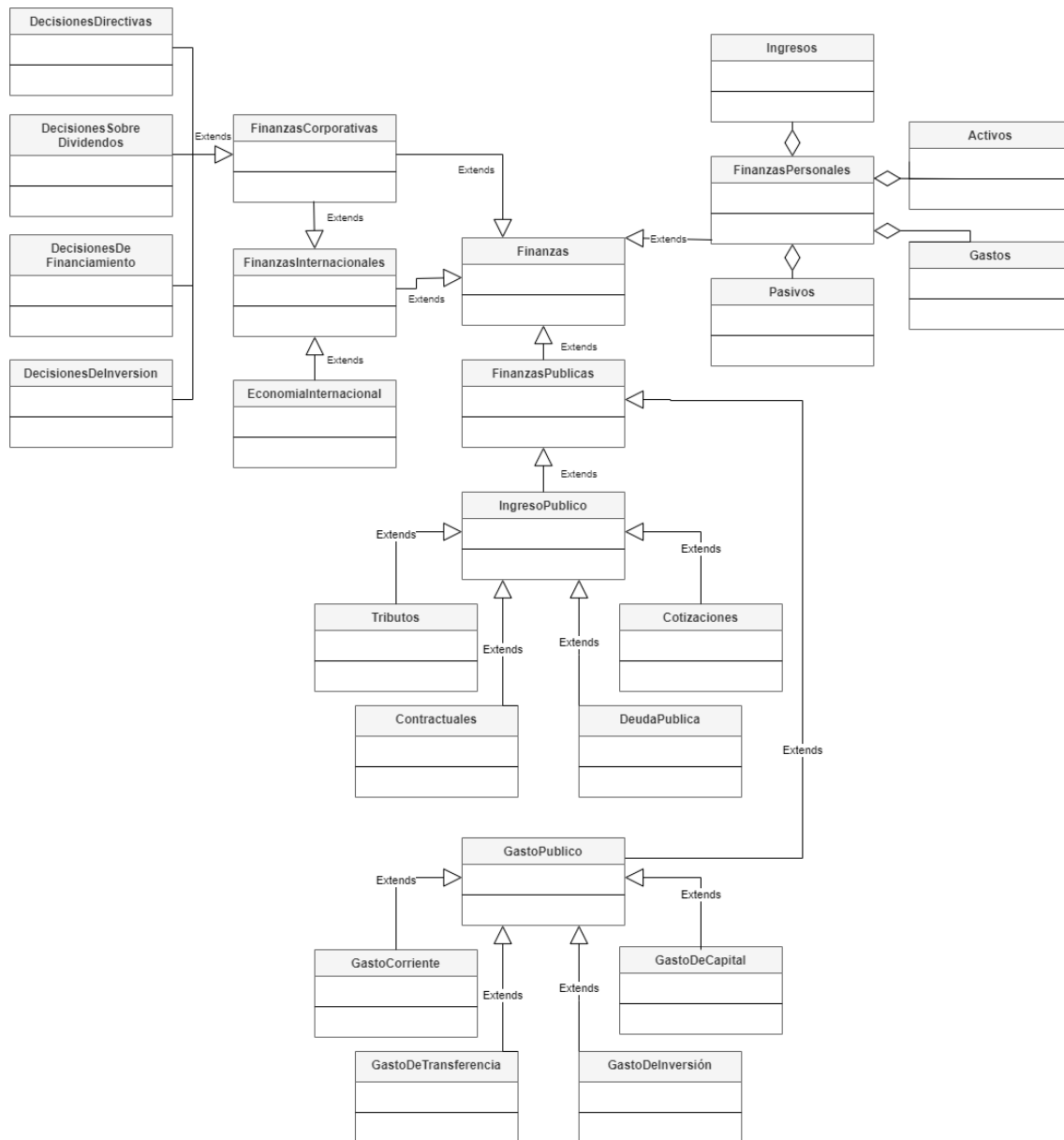


Figura 3.5: Diagrama UML, Finanzas.

3.3. Usos de los Datos a Gran Escala en las Finanzas Corporativas y Personales

El Big Data en el área de las finanzas cuenta con muchas aplicaciones, su uso nos puede proporcionar información valiosa sobre algún cliente para evaluar su condición financiera o poder para mejorar los servicios que se le ofrece, también nos sirve para detectar casos de fraude, así como en la elaboración de planes de inversión haciendo trading de alta frecuencia, entre muchas otras aplicaciones como las que se describen a continuación. [21]

3.3.1. Gestión de riesgos

Una de las principales prioridades de las empresas es minimizar los riesgos en la toma de decisiones financieras. El Big Data en las finanzas permite conocer los mercados y estar preparado ante los continuos cambios. Gracias a esta tecnología se pueden analizar riesgos como la incertidumbre de los mercados, influencia de la competencia, subidas o bajadas de tipos de interés, influencia de la adaptación a nuevas normativas legales, o incluso imprevistos de forma de catástrofes naturales. En definitiva, el procesamiento de datos en tiempo real permite a las compañías conocer su estado y posición respecto a los posibles riesgos, y actuar en consecuencia.

3.3.2. Evaluación de la solvencia

Las entidades o instituciones que prestan servicios financieros usan el Big Data para conocer la solvencia de los clientes y determinar el riesgo a la hora de otorgar préstamos o créditos. Por ejemplo, la aplicación del Big Data en las finanzas permite saber quiénes son sus clientes, cuál es su situación económica, qué otros acuerdos han cerrado, dónde realizan sus compras, etc.

3.3.3. Prevención del fraude

Las empresas, sobre todo las grandes compañías, son susceptibles de ser víctimas de fraudes que podrían acarrear millones en pérdidas. Gracias a las herramientas de Big Data en las finanzas se pueden detectar comportamientos o patrones fraudulentos o actividades sospechosas. A través del análisis predictivo se pueden conocer patrones de compras, datos de geolocalización, historiales de transacciones y mucha otra información que contribuye a prevenir los fraudes y estafas financieras.

3.3.4. Mejora de la atención al cliente

En realidad se trata de un uso que se puede aplicar en cualquier sector. Sin embargo, el Big Data en las finanzas cobra especial importancia a la hora de ofrecer los mejores servicios de soporte a los clientes. Con esta tecnología se pueden hacer predicciones del estado, problemas o necesidades de los clientes y ofrecerles las soluciones adecuadas. Los bancos, entidades financieras o compañías de seguros emplean el Big Data para analizar la situación de sus clientes en tiempo real con sus propias herramientas online. Esto se traduce en un menor tiempo de respuesta, una reducción de los costes operativos y una mayor satisfacción de los clientes.

3.3.5. Personalizar servicios financieros

Gracias al Big Data en las finanzas, los bancos, entidades financieras y entidades de todo tipo pueden conocer mejor a sus clientes, segmentarlos por perfiles y ofrecerles productos y servicios adaptados a sus necesidades. El análisis de información como las preferencias de compra, la relación con las compañías, la frecuencia de visitas al sitio web, o la media de

gasto, ayuda a elaborar perfiles más certeros. Con ello se obtienen ventajas competitivas a través de la personalización de campañas publicitarias o la optimización de las estrategias de venta cruzada o cross-selling.

3.3.6. Trading de alta frecuencia

El High-Frequency trading se basa en el análisis de datos propios y externos para optimizar los procesos de compra y venta de acciones. Las herramientas Big Data permiten procesar una enorme volumen de operaciones de forma automatizada, permitiendo a los traders detectar oportunidades, minimizar errores y tomar las decisiones adecuadas en el momento correcto. La precisión del Big Data a la hora de analizar la información financiera en tiempo real es tal que algunos expertos llaman a esta tecnología High Intelligence trading.

3.3.7. Trading de productos primarios

Se conoce como “commodities” a aquellos bienes básicos que constituyen componentes fundamentales de otros productos más complejos. Por ejemplo, el uso del Big Data en las finanzas ha supuesto un importante avance a la hora de acercar las relaciones entre el mundo financiero y los sectores de la agricultura y la ganadería. Mediante la instalación de sensores y otros dispositivos en grandes o cultivos, se pueden conocer datos en tiempo real sobre el estado de las cosechas, fechas de recolección, enfermedades del ganado, volumen de producción, etc. Todo esto permite a los traders, inversores o empresas agrícolas y ganaderas tomar decisiones basadas en datos reales y no en el azar.

3.3.8. Asesoramiento en inversiones

El Big Data en las finanzas ha cambiado la forma en que se prestan servicios de asesoramiento financiero a los clientes. Hasta hace unos años, los asesores trabajaban basándose en la información patrimonial del cliente. Sin embargo, el Big Data permite obtener y procesar datos de muchas otras fuentes, permitiendo a los profesionales financieros contar con muchos más datos para ofrecer un mejor asesoramiento sobre inversiones. Además, esta nueva forma de operar ha dado lugar a nuevos modelos basados en el asesoramiento online automatizado a través de algoritmos Big Data.

3.3.9. Fidelización de clientes

La retención de clientes es uno de los factores más complejos para las empresas que prestan servicios financieros. Debido a la competitividad del mercado, la volatilidad de los clientes es enorme. Basta con que una entidad rival baje los tipos de interés o haga una mejor oferta sobre algún producto o servicio para que el cliente se vaya con ellos. Sin embargo, con el Big Data se puede trabajar con indicadores que detecten cuando un cliente está perdiendo interés en un producto o servicio financiero de la compañía, por ejemplo, si llevan demasiado tiempo sin entrar en el sitio web o si han interactuado con la competencia en las redes sociales, y actuar antes de que sea demasiado tarde.

3.3.10. Casos de uso de los Datos a Gran Escala aplicados a las Empresas Financieras

Tabla 3.1: Casos de uso del big data en las empresas.

[21]

EMPRESA	DESCRIPCIÓN
BANCO SANTANDER	El Banco Santander puso en marcha en el año 2016 diversas iniciativas basadas en el Big Data con el objetivo de mejorar la atención a sus clientes. Entre ellas está la creación de Santander Analytics, un departamento integrado por expertos en Big Data cuyo objetivo es el control de riesgos y la prevención del fraude. Gracias a la implantación de estas nuevas medidas y herramientas, la entidad bancaria espera ahorrar hasta 2.500 millones de euros en costes operativos.
BANCO BBVA	Otro de los bancos que ha apostado por el Big Data es BBVA, que en el año 2014 puso en marcha el proyecto BBVA Data & Analytics. Se trata de una nueva división integrada por 50 expertos en Big Data y con oficinas en España y México. Entre los proyectos más exitosos desarrollados por este departamento está ‘Commerce360’, una herramienta de business intelligence basada en Big Data y enfocada a pequeñas y medianas empresas.
MORGAN STANLEY	Esta multinacional financiera se dio cuenta de que el grid computing y las bases de datos tradicionales no eran suficientes para hacer frente a la enorme cantidad de datos que manejaban. Por ello, en el año 2010 empezaron a utilizar la herramienta de Big Data Hadoop. Primero la aplicaron solo en 15 servidores destinados a la gestión en el mercado de las commodities. Debido a la eficacia de la herramienta y a los resultados obtenidos hoy en día ya se apoyan en Hadoop a la hora de tomar decisiones para sus proyectos más importantes.

Capítulo 4

Aplicación de la metodología CRISP-DM

CRISP-DM (Cross Industry Standar Process for Data Mining) es una metodología especializada para minería de datos. [22]

- Como **metodología**, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como **modelo de proceso**, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

CRISP-DM está compuesta por seis fases principales, las cuales con excepción de las últimas dos se seguirán en este proyecto:

1. Comprensión del dominio del problema (negocio).
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Despliegue

Para la aplicación de esta metodología se utilizará la herramienta **IBM SPSS Modeler**.

4.1. Comprensión del dominio del problema

Primera Fase de la Metodologia CRISP-DM.

4.1.1. Determinación de los objetivos del proyecto

- Analizar el comportamiento financiero de las personas que solicitan un crédito a alguna institución bancaria.
- Del punto anterior se podrá conseguir un plan para evitar otorgar créditos a personas que no puedan cumplir con el pago.
- Este plan consistirá en tener indicadores de riesgo según el cliente que solicite el crédito.

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 problema que se intenta abordar es un problema de finanzas corporativas y finanzas personales.

La institución bancaria que proporcionó los datos a analizar, requiere saber si es viable otorgar créditos a las personas según sus datos personales y financieros, con el fin de tomar medidas de precaución según un indicador de riesgo.

Algunos aspectos importantes de las finanzas corporativas y personales que se deben de tener en consideración son los siguientes:

Crédito: Un crédito es una operación bancaria de financiación donde una persona llamada “acreedor” (normalmente una entidad financiera), presta una cierta cifra monetaria a otro, llamado “deudor”, quien a partir de ese momento, garantiza al acreedor que retornará esta cantidad solicitada en el tiempo previamente estipulado más una cantidad adicional, llamada “intereses”.

Entidad financiera: Una entidad financiera es cualquier entidad o agrupación que tiene como objetivo ofrecer servicios de carácter financiero y que van desde la simple intermediación y asesoramiento al mercado de los seguros o créditos bancarios.

Estado de cuenta: El estado de cuenta es un documento de validez oficial que expide una institución bancaria o financiera, en donde se visualiza el saldo de la cuenta o crédito y se registran los movimientos que se hayan realizado por un periodo determinado. Éste se expide de manera mensual, en formato físico o digital.

Estos son útiles para la institución bancaria ya que permite:

- Tener un comprobante oficial de una cuenta bancaria.
- Contar con un control interno de movimientos bancarios y pagos realizados.

- Consultar la tasa de interés de los préstamos solicitados, periodos de pago y pagos mínimos para no generar intereses.
- Mejorar la toma de decisiones para pagos mensuales y fomentar el cuidado de las finanzas personales.

Tasa de interés: Una tasa de interés es el costo de pedir dinero prestado o la recompensa por ahorrarlo. Se calcula como un porcentaje del monto que fue entregado por un prestamista en un financiamiento bancario o por una persona que lo guarda en una cuenta de ahorro.

Las características de la tasa de interés son:

- Se miden en porcentaje.
- Se agregan al saldo total que falta por pagar.
- Están predeterminadas.
- Se cobran con base en lo establecido por Banxico.

¿Existen datos disponibles para efectuar el análisis?

Sí, existen datos disponibles. Los datos con los que se cuenta son extraídos del repositorio UCI Machine Learning así como de otros banco de datos y son de tipo multivariante.

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

Sí, se cuentan con los recursos necesarios para desarrollar el proyecto.

- Recursos Humanos:
 - Especialista en minería de datos.
 - Especialista en aprendizaje automatizado.
 - Especialista en finanzas.
- Recursos Tecnológicos:
 - 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 clasificación: Clasificar los datos financieros y personales de las personas que solicitan un crédito, y clasificar el tipo de acción según su identificador de riesgo en el cumplimiento del pago.

Problema de predicción: A partir de los datos disponibles de los clientes, elaborar modelos de predicción que dé como resultado un identificador de riesgo, el cual se usará para determinar la posibilidad de otorgar el crédito que solicitan, y a su vez determinar el riesgo de incumplimiento de pago. De este punto se seleccionará el modelo predictivo con mayor precisión (menor tasa de error).

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

Tabla 4.1: Plan de proyecto de minería de datos.

FASE	TIEMPO A DEDICAR	RECURSOS HUMANOS Y TECNOLÓGICOS	RIESGOS ATRIBUIBLES
1. Comprensión del dominio del problema	5 semanas	Experto en el dominio del problema, experto en minería de datos	No se han identificado riesgos
2. Comprensión de los datos	5 semanas	Experto en el dominio del problema, experto en minería de datos. Tablas, gráficos y resúmenes estadísticos que faciliten la comprensión de los datos. Paquete IBM SPSS Modeler. R Studio. Python.	No se han identificado riesgos
3. Preparación de los datos	5 semanas	Experto en el dominio del problema, experto en minería de datos. Tablas, gráficos y resúmenes estadísticos que faciliten la comprensión de los datos. Paquete IBM SPSS Modeler. R Studio. Python.	No se han identificado riesgos
4. Modelado	6 semanas	Experto en técnicas de machine learning. Herramientas para la implementación de modelos de machine learning. Paquete IBM SPSS Modeler. R Studio. Python.	No se han identificado riesgos

5. Evaluación	6 semanas	Experto en el dominio del problema, experto en minería de datos. Paquete IBM SPSS Modeler. R Studio. Python.	No se han identificado riesgos
6. Presentación	3 semanas	Experto en el dominio del problema, experto en minería de datos. Paquete IBM SPSS Modeler. R Studio. Python.	No se han identificado riesgos

4.2. Comprensión de los datos

Segunda Fase de la Metodología CRISP-DM.

4.2.1. Recopilación de los datos iniciales

Los datos a usar en este proyecto fueron recopilados del repositorio “UCI Machine Learning Repository” dentro del área de Negocios.

El dataset tiene como nombre “South German Credit”, el cual cuenta con 21 atributos y 1,000 instancias.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	status	duration	credit_hist	purpose	amount	savings	employment	installment	personal	stz	other_debt	present_res	property	age	other_install	housing	number_cre	job	people_liab	telephone	foreignn_wc	credit_risk
2	1	18	4	2	1049	1	2	4	2	1	4	2	21	3	1	1	3	2	1	2	1	
3	1	9	4	0	2799	1	3	2	3	1	2	1	36	3	1	2	3	1	1	2	1	
4	2	12	2	9	841	2	4	2	2	1	4	1	23	3	1	1	2	2	1	2	1	
5	1	12	4	0	2122	1	3	3	3	1	2	1	39	3	1	2	2	1	1	1	1	
6	1	12	4	0	2171	1	3	4	3	1	4	2	38	1	2	2	2	2	1	1	1	
7	1	10	4	0	2241	1	2	1	3	1	3	1	48	3	1	2	2	1	1	1	1	
8	1	8	4	0	3398	1	4	1	3	1	4	1	39	3	2	2	2	2	1	1	1	
9	1	6	4	0	1361	1	2	2	3	1	4	1	40	3	2	1	2	1	1	1	1	
10	4	18	4	3	1098	1	1	4	2	1	4	3	65	3	2	2	1	2	1	2	1	
11	2	24	2	3	3758	3	1	1	2	1	4	4	23	3	1	1	1	2	1	2	1	
12	1	11	4	0	3905	1	3	2	3	1	2	1	36	3	1	2	3	1	1	2	1	
13	1	30	4	1	6187	2	4	1	4	1	4	3	24	3	1	2	3	2	1	2	1	
14	1	6	4	3	1957	1	4	1	2	1	4	3	31	3	2	1	3	2	1	2	1	
15	2	48	3	10	7582	2	1	2	3	1	4	4	31	3	2	1	4	2	2	2	1	
16	1	18	2	3	1936	5	4	2	4	1	4	3	23	3	1	2	2	2	1	2	1	
17	1	6	2	3	2647	3	3	2	3	1	3	1	44	3	1	1	3	1	1	2	1	
18	1	11	4	0	3939	1	3	1	3	1	2	1	40	3	2	2	2	1	1	2	1	
19	2	18	2	3	3213	3	2	1	4	1	3	1	25	3	1	1	3	2	1	2	1	
20	2	36	4	3	2337	1	5	4	3	1	4	1	36	3	2	1	3	2	1	2	1	
21	4	11	4	0	7228	1	3	1	3	1	4	2	39	3	2	2	2	2	1	2	1	
22	1	6	4	0	3676	1	3	1	3	1	3	1	37	3	1	3	3	1	1	2	1	
23	2	12	4	0	3124	1	2	1	3	1	3	1	49	1	2	2	2	1	1	2	1	
24	2	36	2	5	2384	1	2	4	3	1	1	4	33	3	1	1	2	2	1	2	0	
25	2	12	4	4	1424	1	4	4	3	1	3	2	26	3	2	1	3	2	1	2	1	
26	1	6	4	0	4716	5	2	1	3	1	3	1	44	3	2	2	2	1	1	2	1	
27	2	11	3	3	4771	1	4	2	3	1	4	2	51	3	2	1	3	2	1	2	1	
28	1	12	2	2	652	1	5	4	2	1	4	2	24	3	1	1	3	2	1	2	1	
29	2	9	4	3	1154	1	5	2	3	1	4	1	37	3	2	3	2	2	1	2	1	
30	4	15	2	0	3556	5	3	3	3	1	2	4	29	3	2	1	3	2	1	2	1	
31	3	42	4	1	4796	1	5	4	3	1	4	4	56	3	3	1	3	2	1	2	1	
32	3	30	4	3	3017	1	5	4	3	1	4	2	47	3	2	1	3	2	1	2	1	
33	4	36	4	0	3535	1	4	4	3	1	4	3	37	3	2	2	3	2	2	2	1	
34	4	36	4	0	6614	1	5	4	3	1	4	3	34	3	2	2	4	2	2	2	1	
35	4	24	2	3	1376	3	4	4	2	1	1	3	28	3	2	1	3	2	1	2	1	
36	1	15	2	0	1721	1	2	2	3	1	3	1	36	3	2	1	3	2	1	2	1	
37	1	6	4	0	860	1	5	1	2	1	4	4	39	3	2	2	3	2	2	2	1	
38	4	12	4	0	1495	1	5	4	3	1	1	1	38	3	2	2	2	1	1	2	1	

Figura 4.1: Vista Previa del dataset "South German Credit".

4.2.2. Descripción de los datos

La tabla está compuesta por 21 atributos de entrada y 1000 instancias. Los atributos, descripción y tipos son los siguientes:

Tabla 4.2: Descripción y tipo de los datos.

ATRIBUTO	DESCRIPCIÓN	TIPO
status	Estado de la cuenta del deudor en el banco	Categorico
duration	Duración del crédito en meses	Cuantitativo
credit_history	Historial de cumplimiento de contratos de crédito anteriores	Categorico
purpose	Propósito por el cual se necesita el crédito	Categorico
amount	Monto del crédito en DM	Cuantitativo
savings	Ahorros del deudor	Categorico
employment_duration	Duración del empleo del deudor con el empleador actual	Ordinal
installment_rate	Tasa de pago, porcentaje del ingreso disponible del deudor	Ordinal
personal_status_sex	Información que combina el sexo y el estado civil del deudor	Categorico
other_debtors	¿Hay otro deudor o garante del crédito?	Categorico
present_residence	Tiempo (en años) que el deudor tiene viviendo en su residencia actual	Ordinal
property	Propiedad más valiosa del deudor	Ordinal
age	Edad en años	Cuantitativo
other_installment_plans	Planes de pago a plazos a otras instituciones que no son el banco que otorga el crédito	Categorico
housing	Tipo de vivienda en la que vive el deudor	Categorico
number_credits	Número de créditos incluyendo el actual que tiene (o tuvo) el deudor con el banco prestamista	Ordinal
job	Calidad del trabajo del deudor	Ordinal

people_liable	Número de personas que dependen económicamente del deudor	Binario
telephone	¿Existe un teléfono fijo registrado a nombre del deudor?	Binario
foreign_worker	¿El deudor es un trabajador extranjero?	Binario
credit_risk	¿Se ha cumplido (bien) o no (mal) el contrato de crédito?	Binario

4.2.3. Exploración de los datos

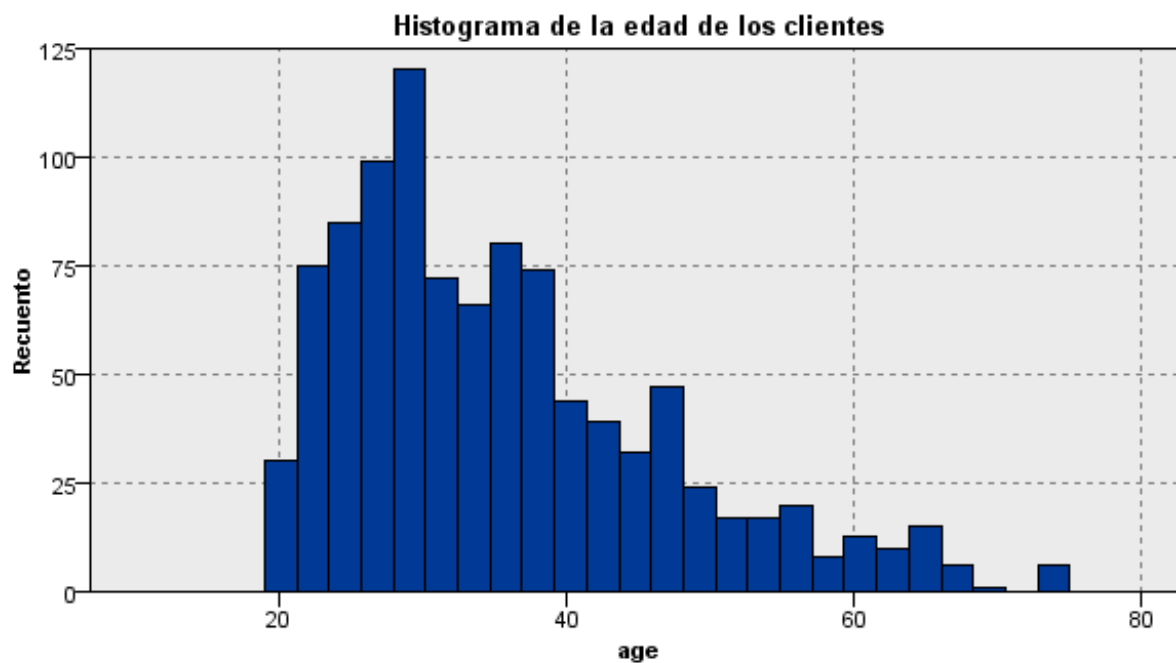


Figura 4.2: Histograma de la edad de los clientes.

Como se puede observar en la Gráfica 4.2, la mayor cantidad de clientes que solicitan un préstamo a la institución financiera tienen entre 23 a 39 años.

Al contrario, las personas entre 60 y 80 años no son tan recurrentes.

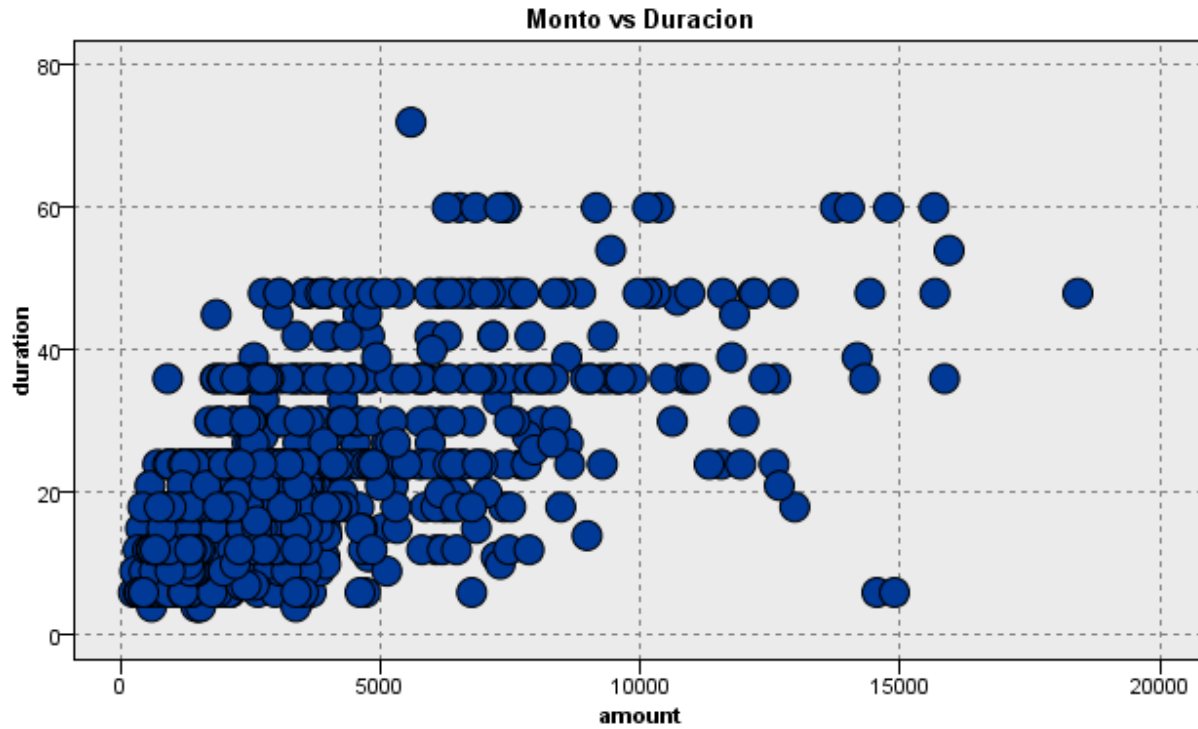


Figura 4.3: Gráfica Monto x Duración del crédito.

En la Gráfica 4.3. Se puede observar que la gente solicita más préstamos de 1 a 10000 DM y los pagan en un plazo de 1 a 45 meses.

Por otra parte, la gráfica nos muestra que la gente evita pedir préstamos de más de 15000 DM.

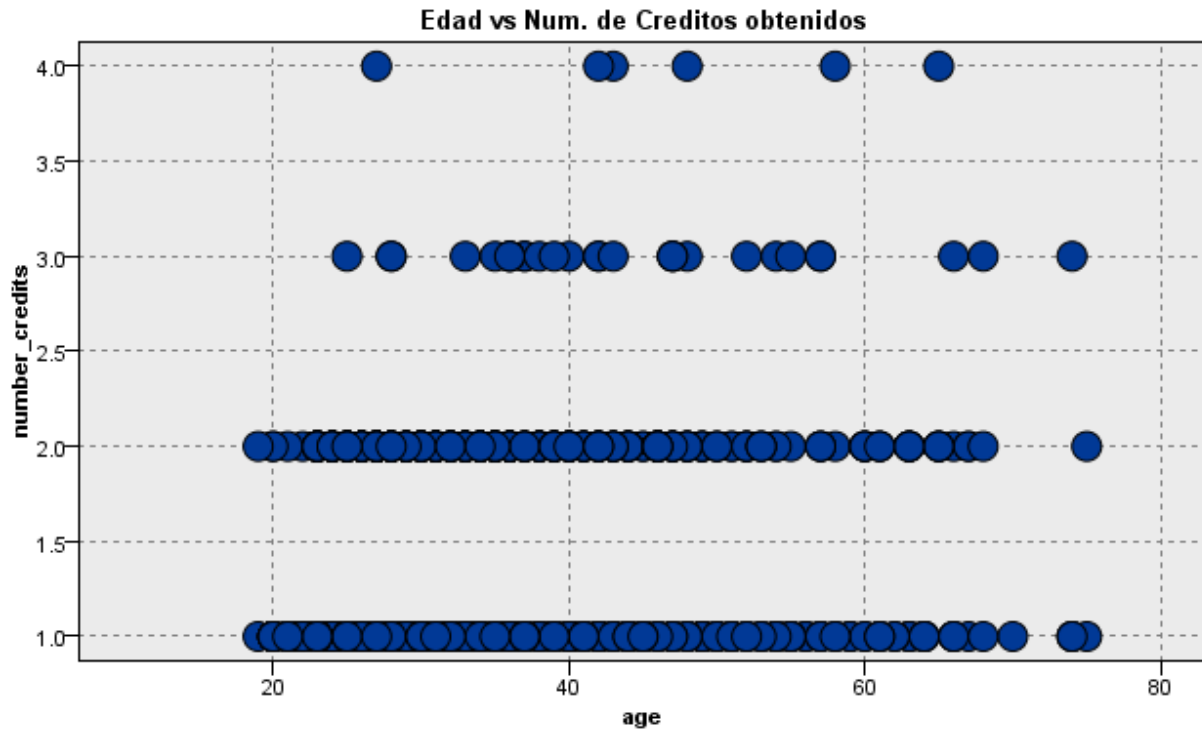


Figura 4.4: Gráfica Edad de los clientes x Núm. de Créditos obtenidos.

En la Gráfica 4.4. Se observa que para mucha gente es su primer o segunda solicitud de un crédito. Además, las personas entre 22 a 59 años son las que más créditos han solicitado.

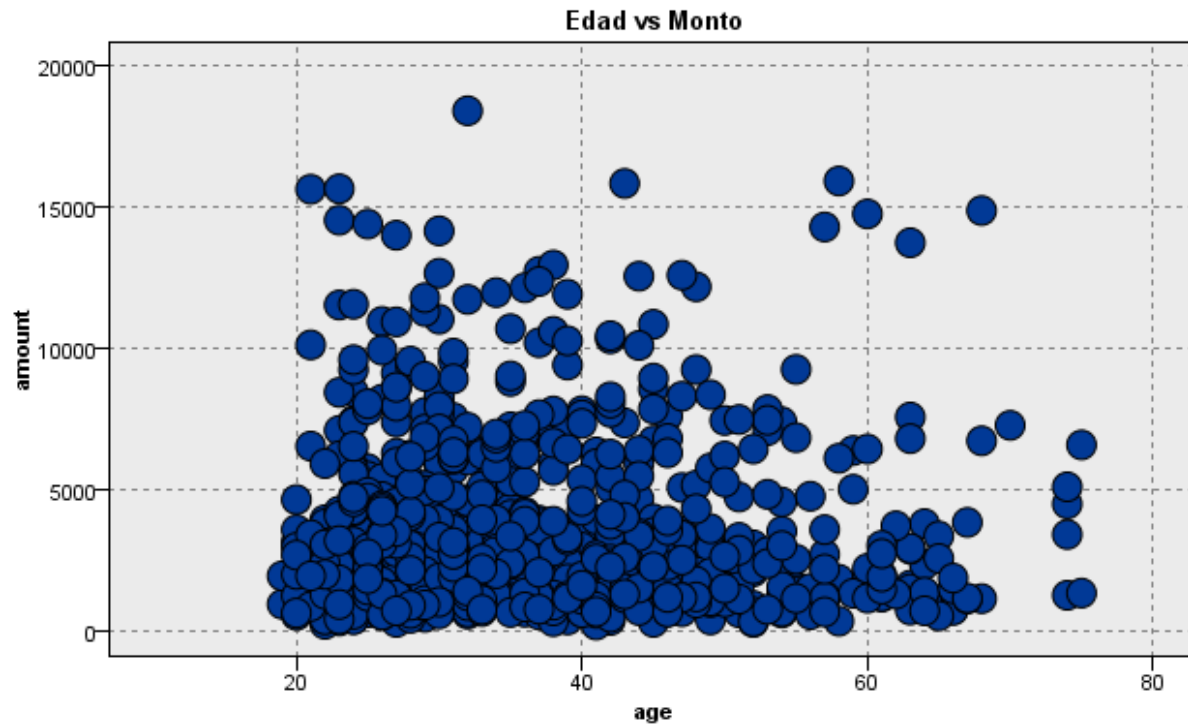


Figura 4.5: Gráfica de la edad de los clientes x el monto que solicitan.

En la Gráfica 4.5. se puede observar que la mayor cantidad de gente que tiene entre 20 a 50 años de edad solicitan préstamos de 0 a 10000 DM. Además las personas que tienen entre 20 a 40 años suelen pedir más dinero prestado.

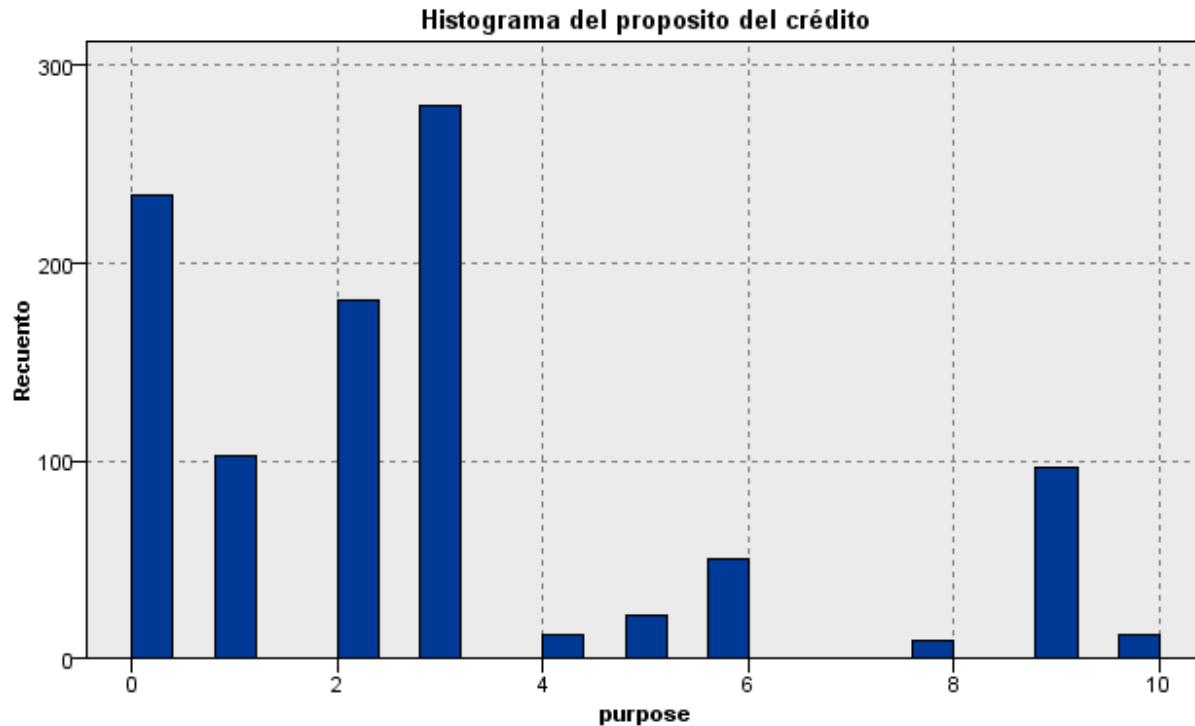


Figura 4.6: Histograma del propósito del crédito. Donde 0 = otros, 1 = Carro Nuevo, 2 = Carro Usado, 3 = Muebles o equipamiento, 4 = Radio o televisión, 5 = Usos domésticos, 6 = Reparaciones, 7 = Educación, 8 = vacaciones, 9 = Cursos, 10 = Negocios.

En la Gráfica 4.6. se puede observar que la mayor cantidad de las solicitudes de algún crédito son utilizados para adquirir Muebles o equipamiento, después para otras razones, y en tercer lugar para adquirir un carro Usado.

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

El Volumen, Variedad y Veracidad de los datos existentes están garantizados, pero aun así es necesario adquirir más datos, de forma que se mejore el volumen y la certeza de los datos.

4.3. Preparación de los datos

Tercera Fase de la Metodología CRISP-DM.

4.3.1. Selección de datos

Inicialmente se cuenta con 21 atributos de entrada y 1000 instancias dentro del dataset, pero el atributo **telephone** nos resulta irrelevante, por lo que no será tomado en cuenta.

Más adelante se considerará la derivación de un nuevo atributo, lo que podría ocasionar la exclusión de los atributos existentes, al dejar de ser relevantes.

Como se indicó en la fase “Comprensión de los datos” 4.2, el número de registro podría resultar insuficiente para garantizar precisión en los objetivos del proyecto, por lo cual se considerará la integración de nuevos datos.

4.3.2. Limpieza de datos

Según lo determinado en la fase de “Comprensión de los datos” 4.2, no se habían encontrado problemas con los datos, pero en la Selección de Datos se determinó que el atributo **telephone** es irrelevante, entonces se eliminará del dataset. Además se hará una preparación automática de los datos por medio de la herramienta IBM SPSS Modeler, con el fin de solucionar errores que no se pudieron encontrar.



Figura 4.7: SPSS Modeler, aplicación de preparación automática de datos y aplicación de un filtro a los datos.

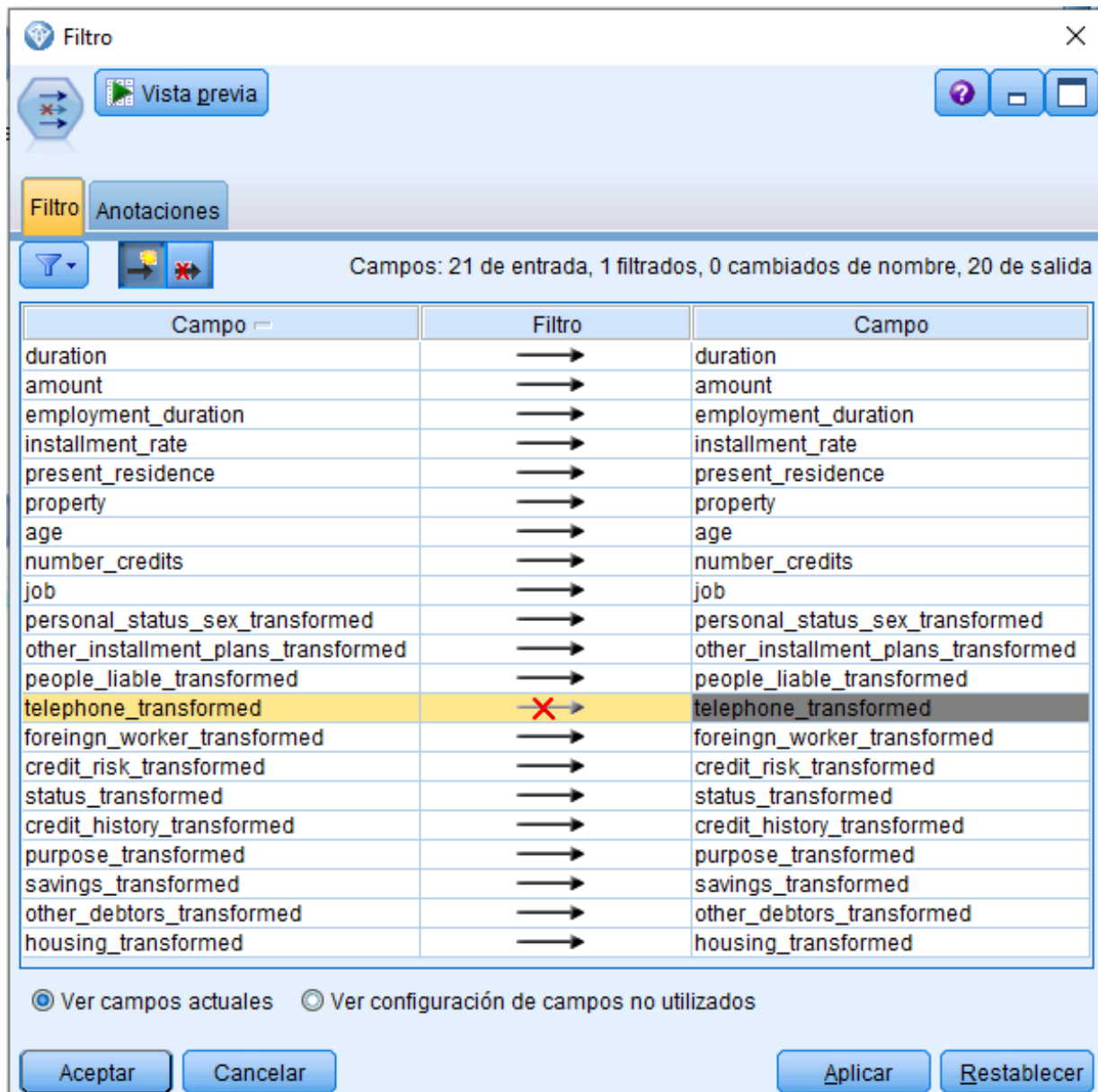


Figura 4.8: SPSS Modeler, Eliminación del registro “Teléfono”.

4.3.3. Integración de datos

Para poder garantizar el volumen y la certeza de los datos se ha considerado hacer la integración de nuevos datos los cuales son recopilados del repositorio “UCI Machine Learning Repository” dentro del área de Negocios. El nuevo dataset tiene como nombre “South German Credit (UPDATE)”, el cual cuenta con 21 atributos y 1,000 instancias.

Este dataset es la continuación del dataset presentado en la fase “Compresión de los datos”. Por lo tanto cuenta con los mismos atributos lo que nos facilita la integración de estos nuevos datos.

Antes de esta integración, se aplicó la selección y limpieza de los datos, la cual fue planteada al inicio de esta fase:

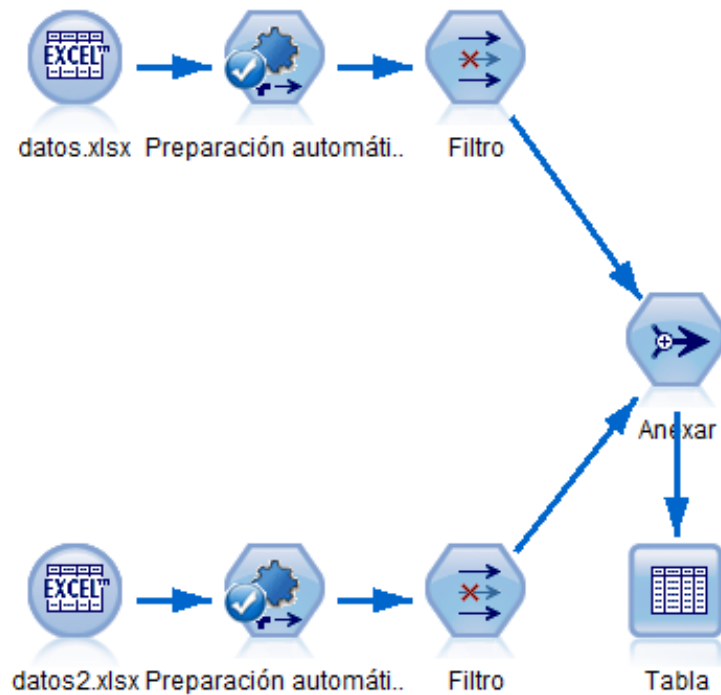


Figura 4.9: SPSS Modeler, Anexar nuevos datos adquiridos.

El resultado de esta integración es una tabla que cuenta con 20 atributos y un total de 2,000 instancias, de esta manera garantizamos la certeza y volumen de los datos, e incrementamos la precisión en los objetivos del proyecto.

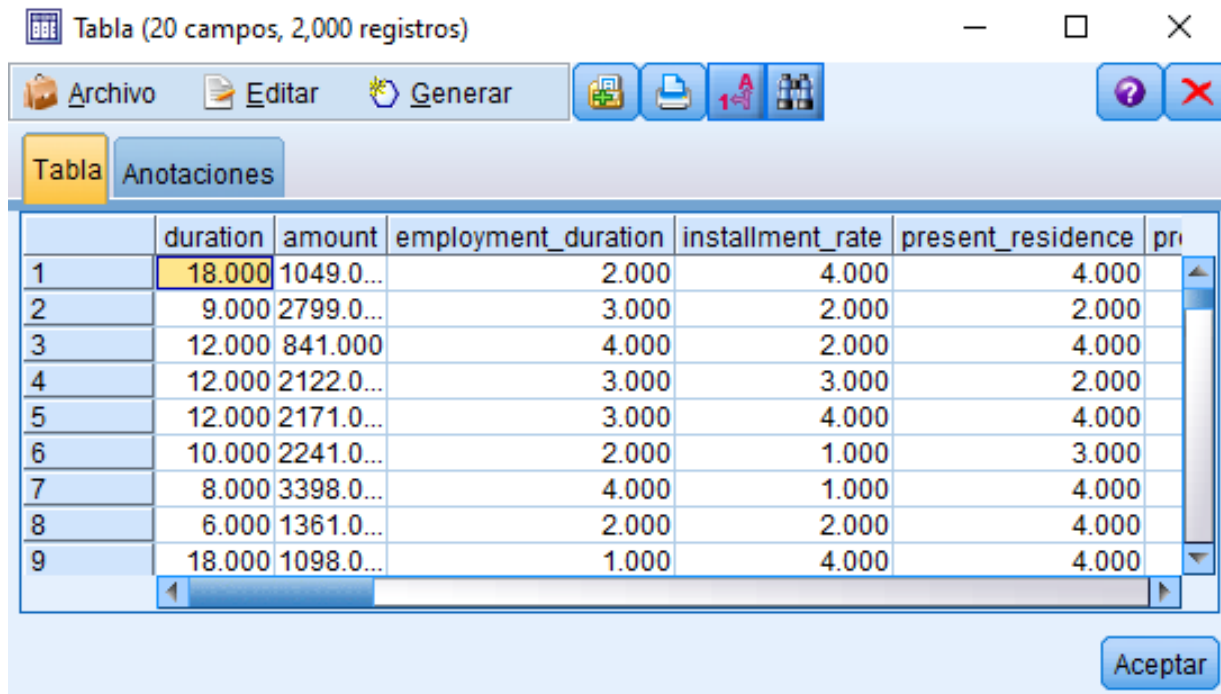


Tabla (20 campos, 2,000 registros)

Archivo Editar Generar

Tabla Anotaciones

	duration	amount	employment_duration	installment_rate	present_residence	pri
1	18.000	1049.0...	2.000	4.000	4.000	
2	9.000	2799.0...	3.000	2.000	2.000	
3	12.000	841.000	4.000	2.000	4.000	
4	12.000	2122.0...	3.000	3.000	2.000	
5	12.000	2171.0...	3.000	4.000	4.000	
6	10.000	2241.0...	2.000	1.000	3.000	
7	8.000	3398.0...	4.000	1.000	4.000	
8	6.000	1361.0...	2.000	2.000	4.000	
9	18.000	1098.0...	1.000	4.000	4.000	

Aceptar

Figura 4.10: SPSS Modeler, Resultado de Anexar datos.

4.3.4. Construcción de nuevos datos

Siguiendo el objetivo del proyecto de minería de datos, es necesario derivar el atributo **“Riesgo_Derivado”**, a partir de los atributos **credit_history**, **job** y **employment_duration**. De esta manera, el nuevo atributo será el atributo cuyo valor se desea predecir.

Riesgo_Derivado

Vista previa

Derivar como: Condicional

Configuración Anotaciones

Modo: ☒ Único ☐ Múltiple

Derivar campo:

Riesgo_Derivado

Derivar como: Condicional

Tipo de campo: Ordinal

Si:

1 ((credit_history_transformed >= 2) and ((job >= 2) or (employment_duration >= 1)))

Entonces:

1 0

En caso contrario:

1 1

Aceptar Cancelar Aplicar Restablecer

Figura 4.11: SPSS Modeler, Derivación de un nuevo campo.

En la Imagen 4.11, el valor 0 significa que no hay un riesgo en la otorgación de un crédito, por el contrario, el valor 1 significa que hay un posible riesgo en la otorgación del crédito.

4.3.5. Formato de datos

Tabla 4.3: Formato de los datos.

ATRIBUTO	TIPO	ROL
status	Nominal	Entrada
duration	Continuo	Entrada
credit_history	Nominal	Entrada

purpose	Nominal	Entrada
amount	Continuo	Entrada
savings	Nominal	Entrada
employment_duration	Nominal	Entrada
installment_rate	Nominal	Entrada
personal_status_sex	Nominal	Entrada
other_debtors	Nominal	Entrada
present_residence	Nominal	Entrada
property	Nominal	Entrada
age	Continuo	Entrada
other_installment_plans	Nominal	Entrada
housing	Nominal	Entrada
number_credits	Nominal	Entrada
job	Nominal	Entrada
people_liable	Ordinal	Entrada
foreign_worker	Ordinal	Entrada
credit_risk	Ordinal	Entrada
Riesgo_Derivado	Ordinal	Salida

4.3.6. Nueva exploración de los datos

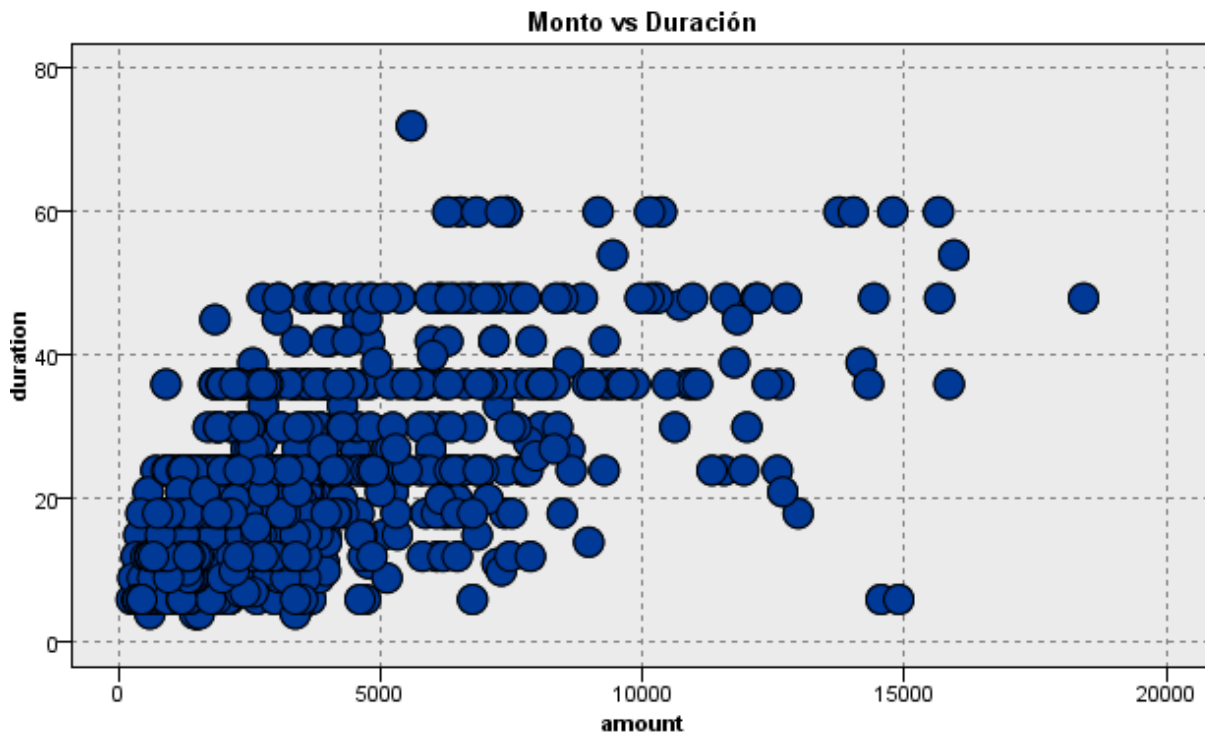


Figura 4.12: Gráfica Monto vs Duración del Crédito.

En la Gráfica 4.12, se puede observar que la gente solicita más préstamos de 1 a 10000 DM y los pagan en un plazo de 1 a 45 meses.

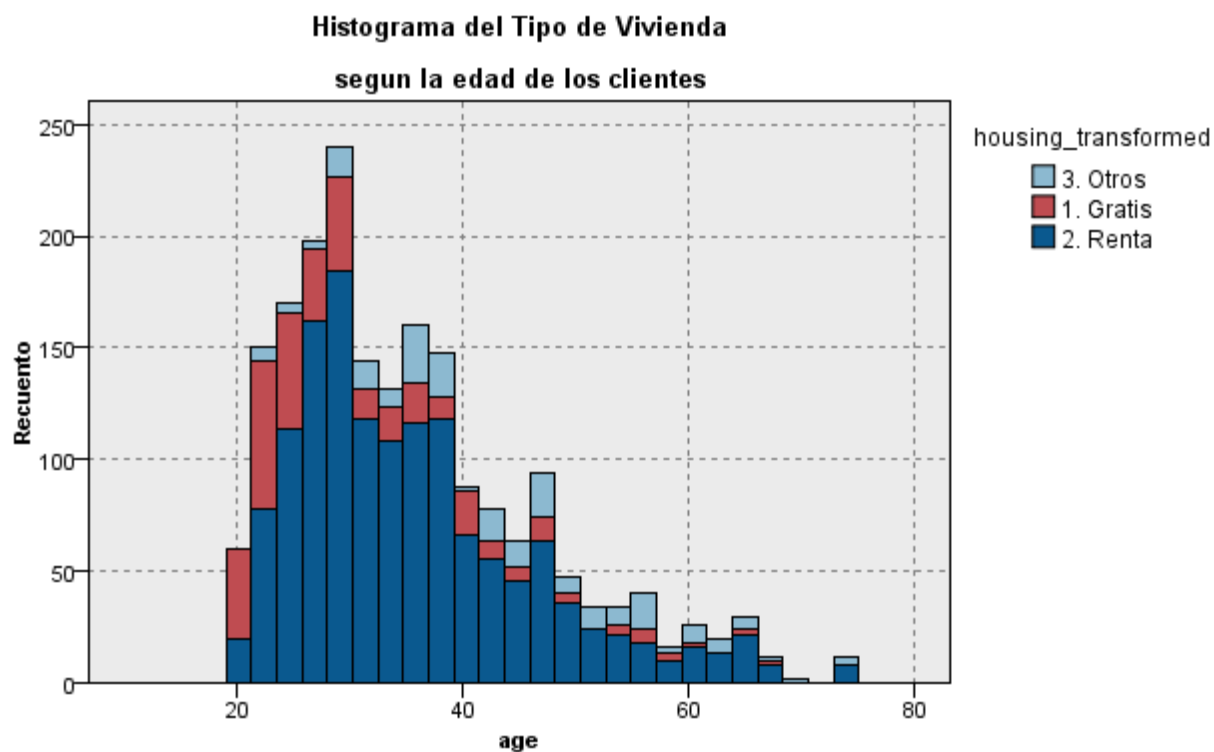


Figura 4.13: Histograma del Tipo de Vivienda según la edad de los clientes.

En esta Gráfica 4.13 se puede observar que la mayor parte de la gente que solicita un préstamo renta su vivienda.

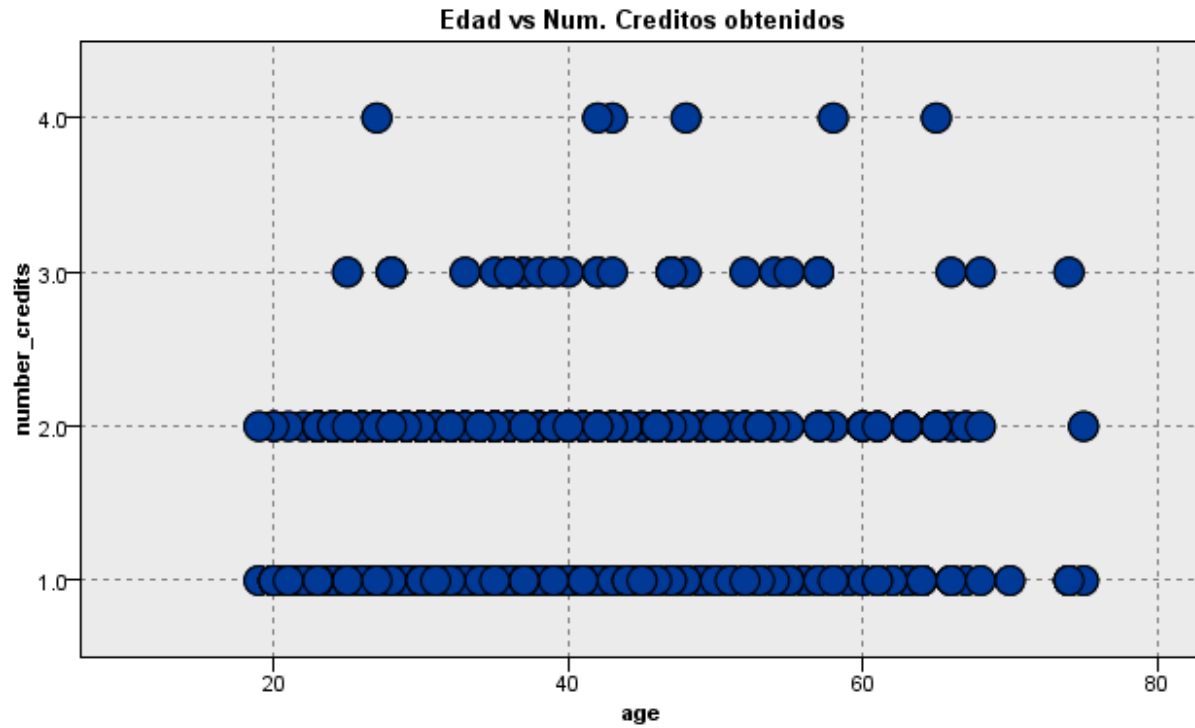


Figura 4.14: Gráfica de la edad vs el número de créditos obtenidos.

En la Gráfica 4.14, se observa que para mucha gente es su primer o segunda solicitud de un crédito. Además, las personas entre 22 a 59 años son las que más créditos han solicitado.

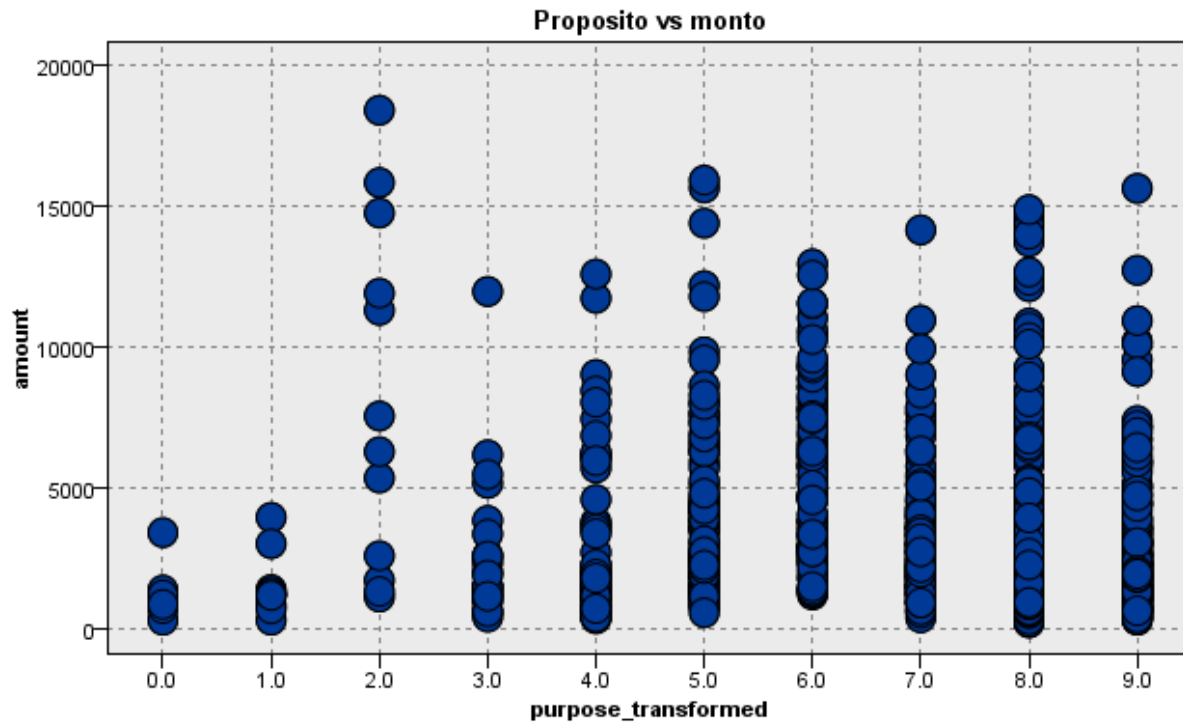


Figura 4.15: Gráfica entre el propósito del crédito vs el monto.

En la Gráfica 4.15, se observa que las menores cantidades de prestamos las ocupan para irse de vacaciones y para comprar una Televisión o Radio. A su vez los prestamos de más de 15,000 DM se ocupan principalmente para Negocios, Cursos de capacitación y la compra de mobiliario o maquinaria.

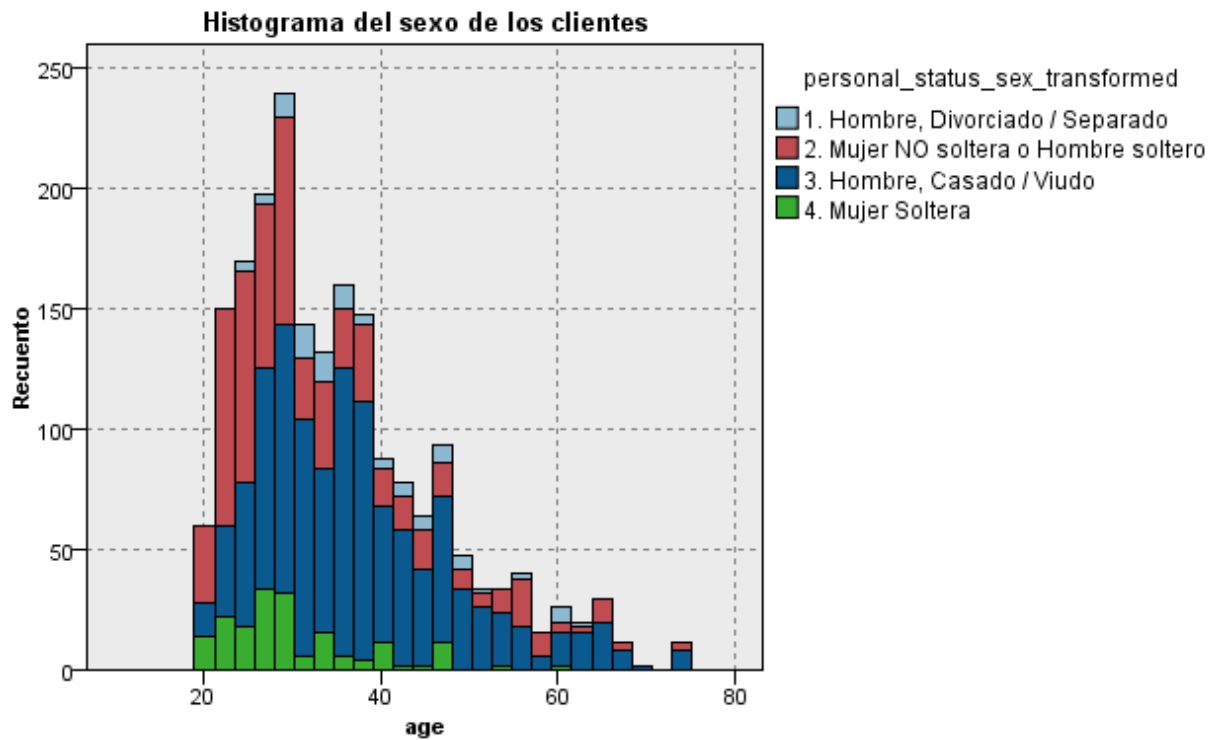


Figura 4.16: Histograma del sexo de los clientes.

De la Gráfica 4.16, se puede observar que la mayor cantidad de gente que solicita un préstamo son hombres y mujeres casados o que no son solteros.

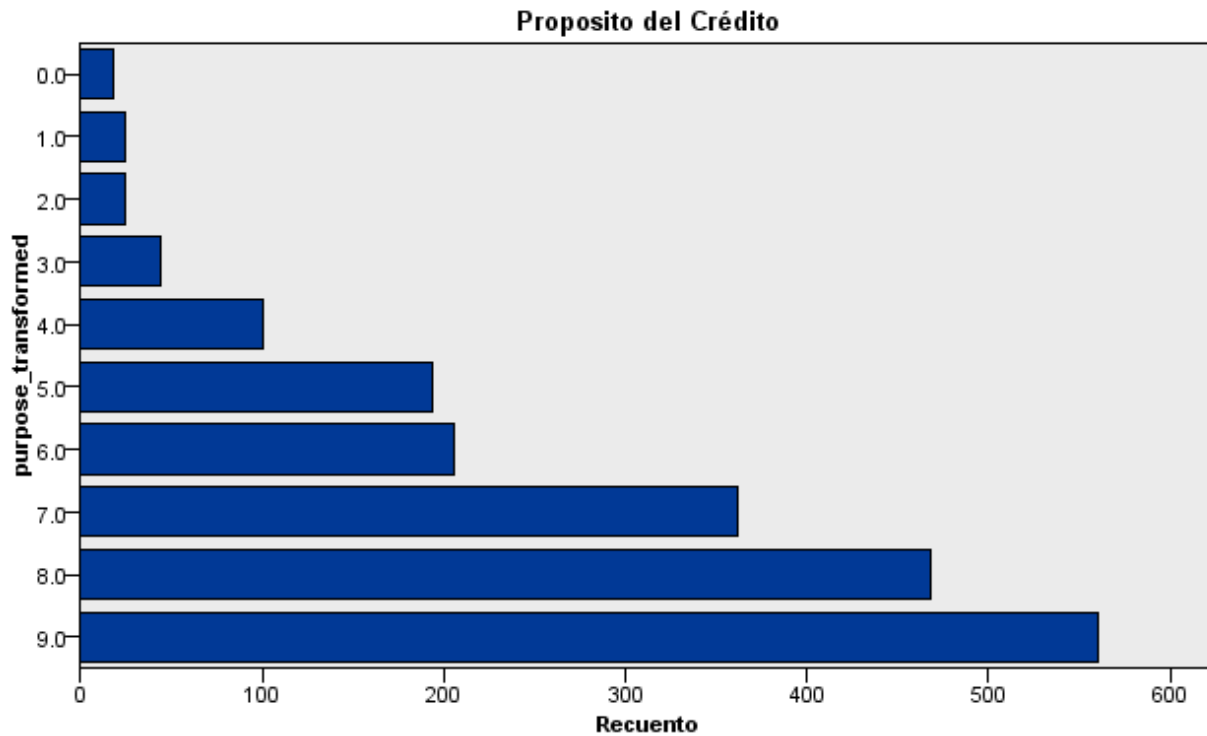


Figura 4.17: Propósito del Crédito. dónde 0.0 = Vacaciones, 1.0 = Radios / Televisiones, 2.0 = Negocios, 3.0 = Aplicaciones Domésticas, 4.0 = Reparos, 5.0 Cursos de capacitación, 6.0 = Carro Nuevo, 7.0 = Carro Usado, 8.0 = Otros, 9.0 = Mobiliario / Maquinaria.

De la Gráfica 4.17, se puede observar que la mayor cantidad de créditos son ocupados principalmente para comprar mobiliario o maquinaria, para otros propósitos de los listados y para comprar carros usados.

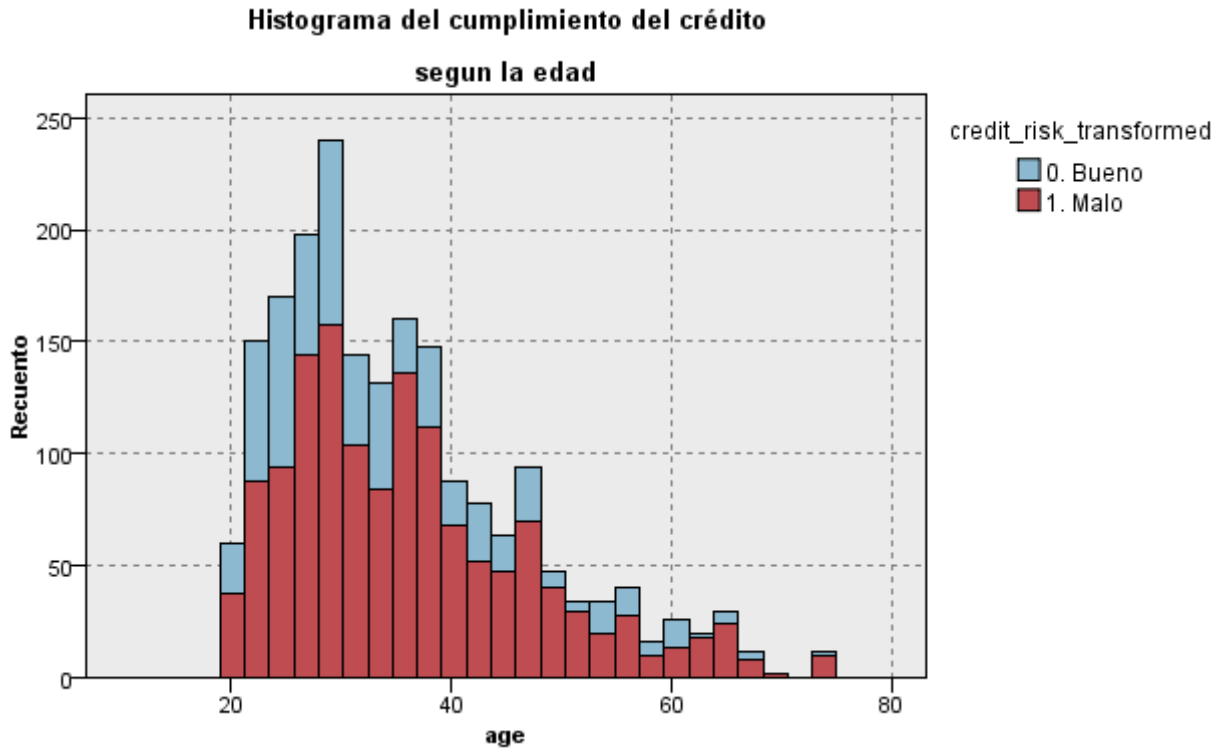


Figura 4.18: Histograma del cumplimiento del crédito según la edad.

De la Gráfica 4.18, podemos observar que más del 50 % de la gente según su edad no han cumplido el pago de su crédito.

4.4. Modelado

Cuarta Fase de la Metodología CRISP-DM.

4.4.1. Selección de técnicas de modelado

Se escogieron 5 modelos que responden al objetivo de minería de datos que nos ocupa: predicción de un indicador de riesgo que sirva para evitar otorgar un crédito a las personas que no sean aptas. Los modelos seleccionados fueron los siguientes:

- Árbol de decisión C&R.
- Algoritmo KNN.
- Algoritmo de regresión lineal.
- Algoritmo de regresión logística.
- Red Neuronal, Perceptron Backpropagation.

4.4.2. Métodos de Comprobación

Considerando que los modelos propuestos son técnicas de aprendizaje supervisado, entonces el método de comprobación seleccionado viene dado por el criterio de bondad del modelo, siendo en este caso la tasa de error del modelo.

Para comprobar el criterio de bondad del modelo, los 5 modelos propuestos incluyen entre sus prestaciones la partición de los datos en dos conjuntos, uno para el entrenamiento del modelo y el otro para la prueba o verificación del mismo.

4.4.3. Generación de los Modelos

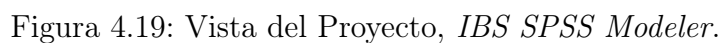
Los siguientes modelos fueron generados utilizando el paquete *IBM SPSS Modeler*:

- Red Neuronal Perceptron Backpropagation.
- Árbol de decisión C&R.
- Algoritmo de Regresión Logística.
- Algoritmo KNN.
- Algoritmo de Regresión Lineal.

Inicialmente fueron utilizados los parámetros por defecto propuestos por cada modelo.

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

La Figura 4.19 representa una vista general del avance del proyecto de minería de datos.



A continuación se mostrarán los pasos a seguir para generar el modelo de Red Neuronal Perceptron Backpropagation, en la herramienta *IBM SPSS Modeler*. Como ya se planteó, se utilizarán los parámetros por defecto propuestos por cada modelo.

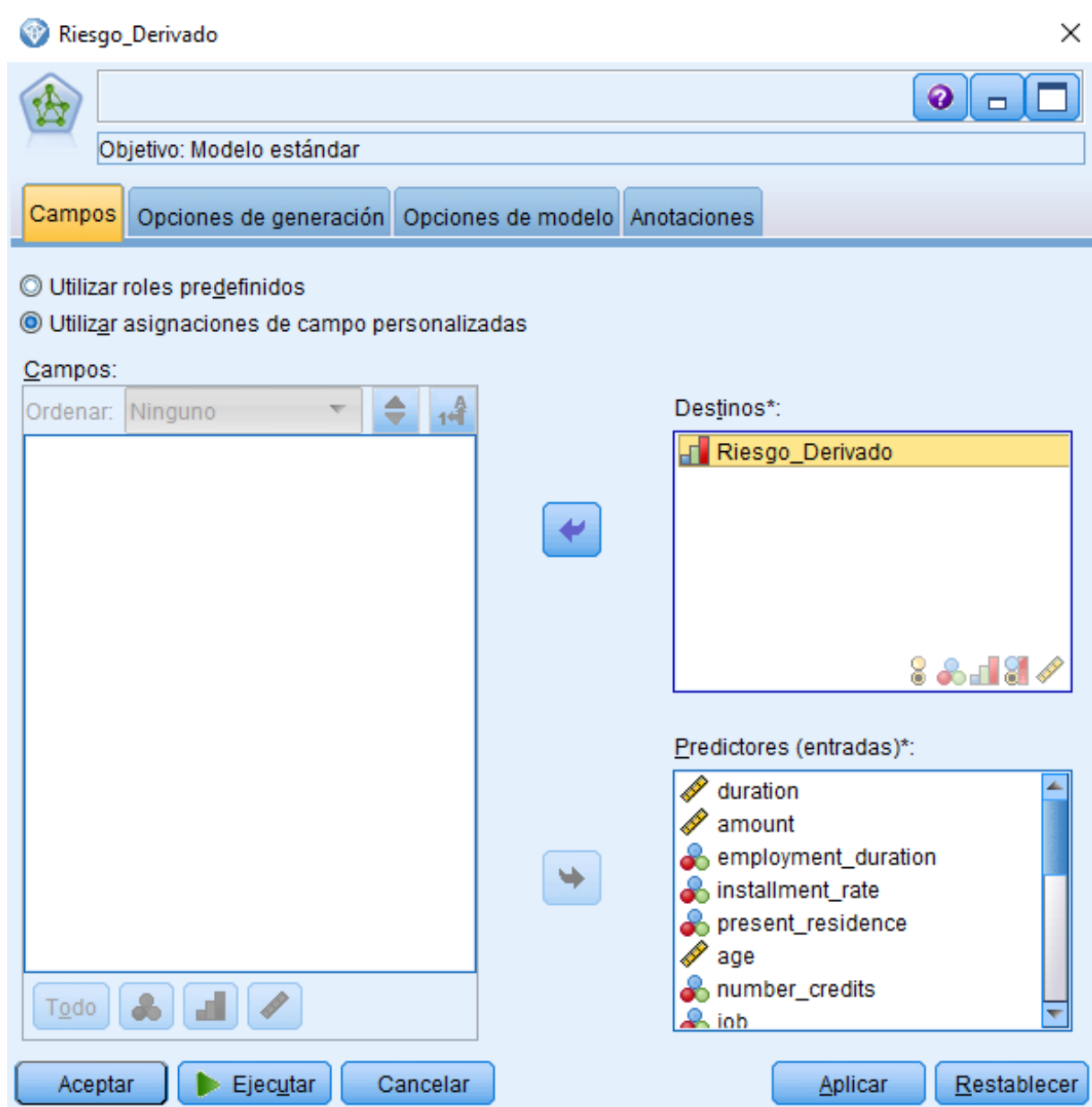


Figura 4.20: IBS SPSS Modeler, Generación del modelo red neuronal.

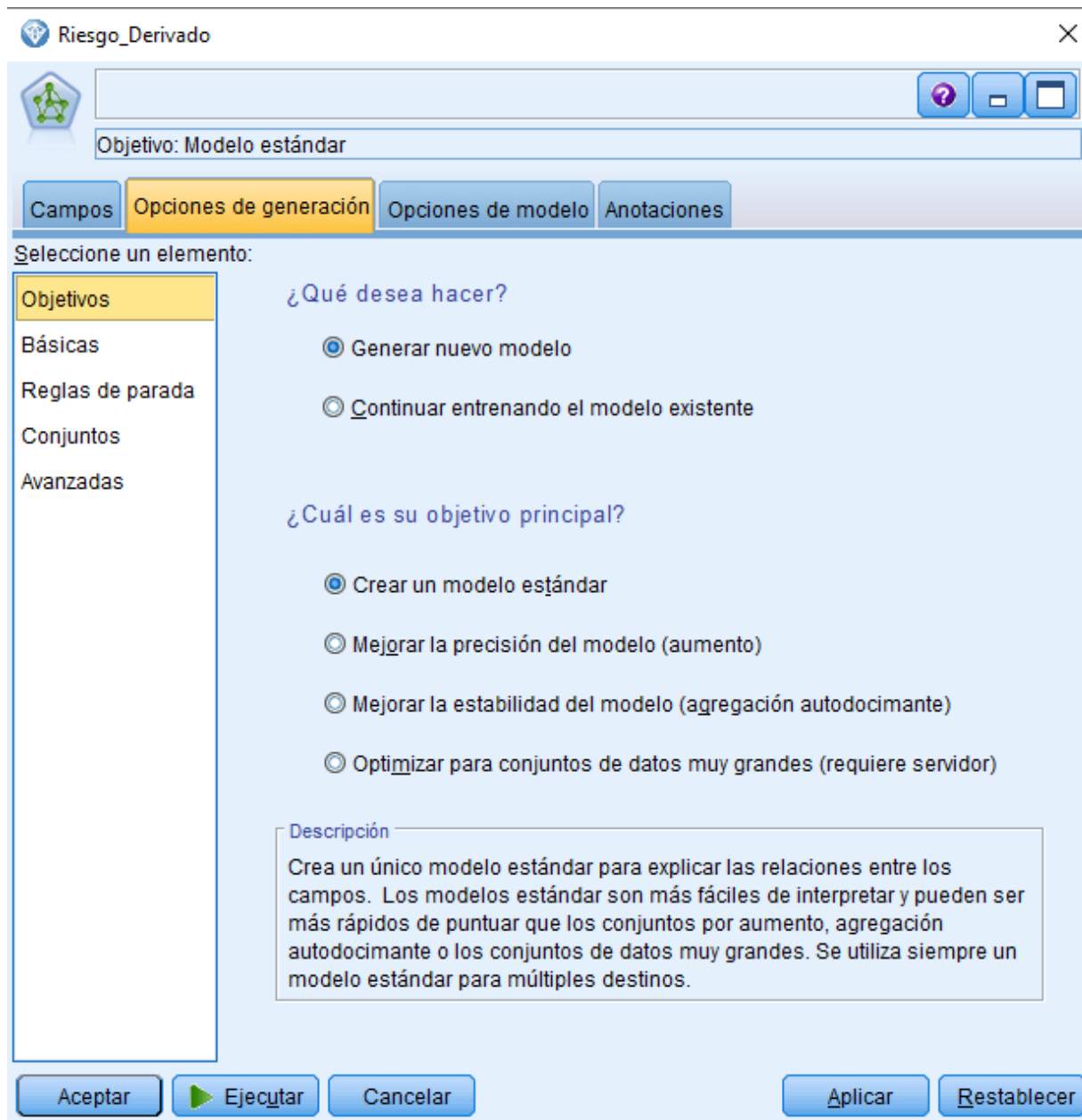


Figura 4.21: IBS SPSS Modeler, Opciones para la Generación del modelo red neuronal.

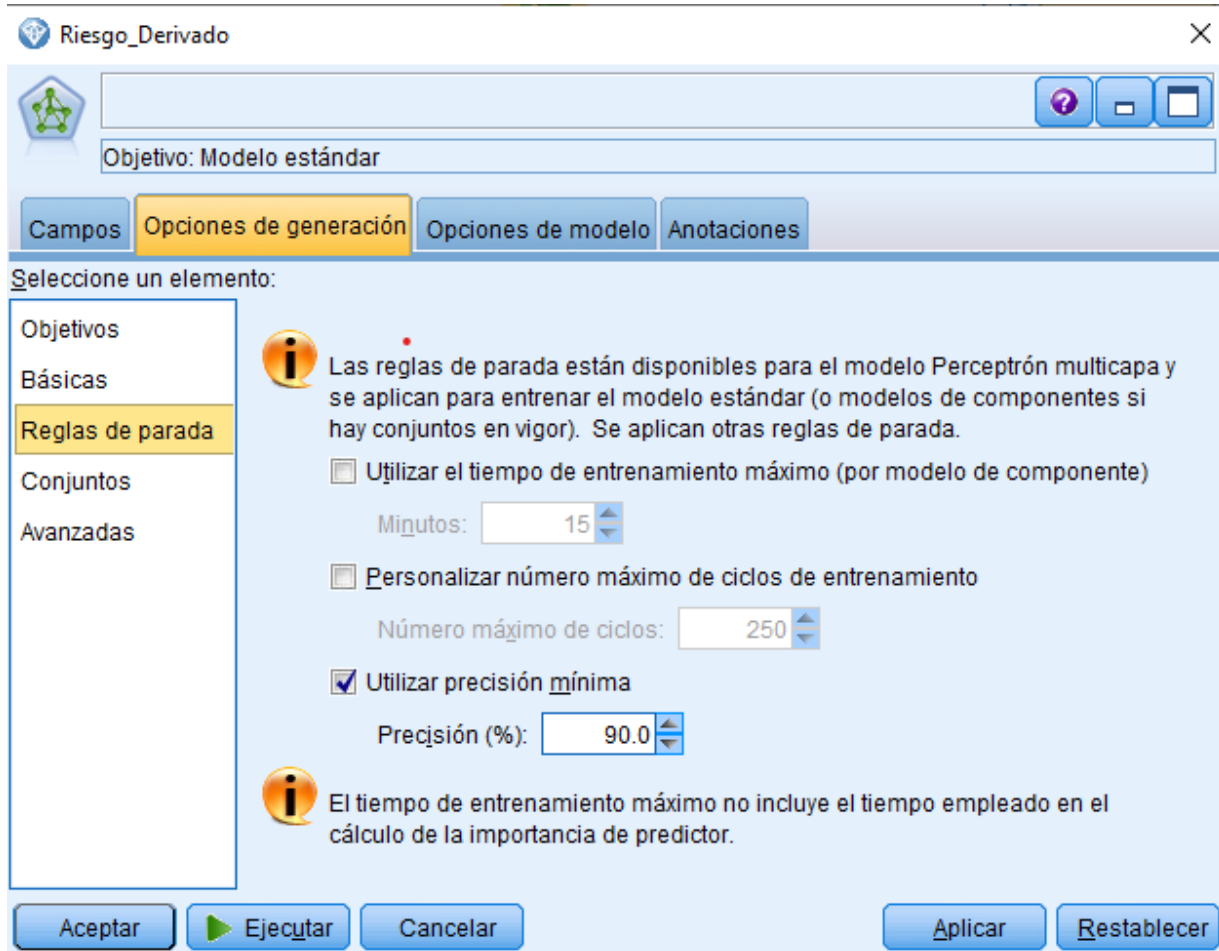


Figura 4.22: IBS SPSS Modeler, Opciones para la Generación del modelo red neuronal 2.

4.4.5. Ejecución del modelo: Red Neuronal Perceptron Backpropagation

En esta sección se mostrarán todos datos generados por el modelo. Para la Red Neuronal Perceptron Backpropagation se mostrará el Resumen del modelo, la importancia del predictor, la clasificación para el registro de salida y la Red generada.

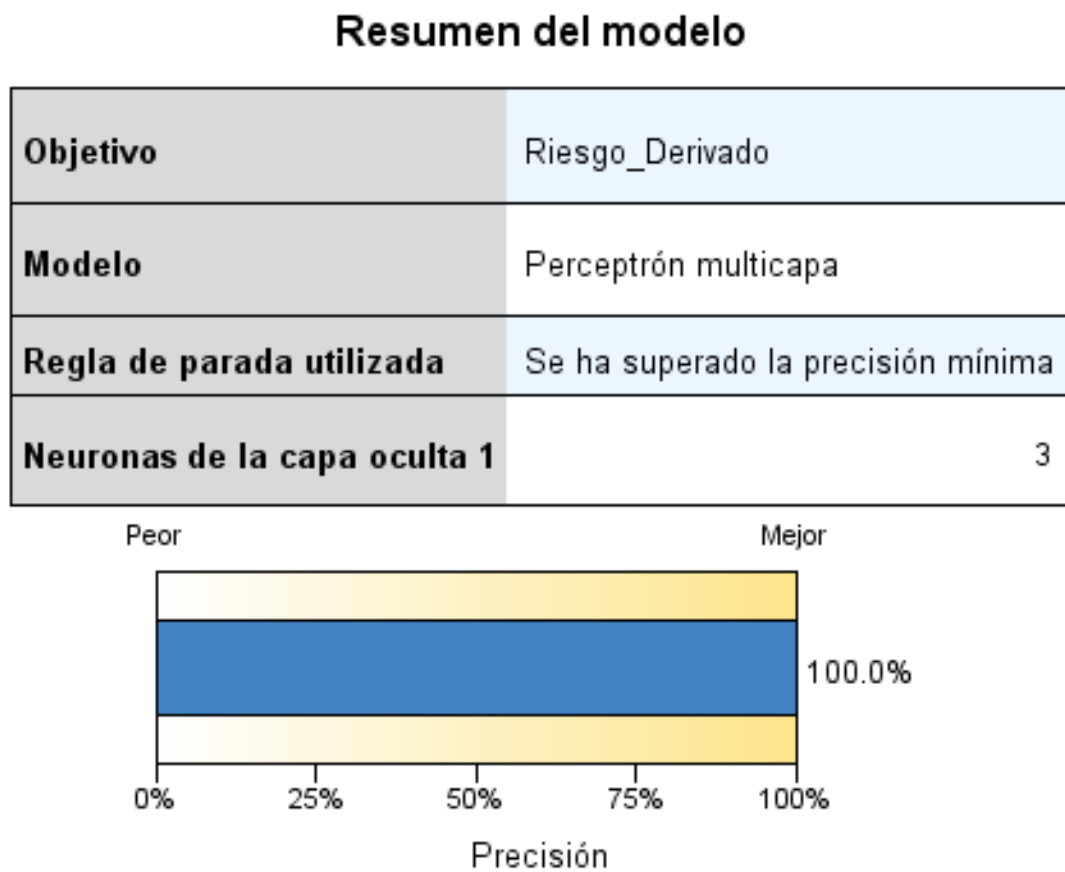


Figura 4.23: *IBS SPSS Modeler*, Resumen del modelo red neuronal.

Como se puede apreciar en la Figura 4.23 el modelo de red neuronal Perceptron Back-propagation no arrojó un resultado del 100 % el cual es el ideal al momento de pronosticar datos nuevos ya que no se han encontrado datos erróneos o que generen medidas falsas, al modelo.

También se aprecia que fueron generadas una sola capa oculta en la Red Neuronal con 3 neuronas en la parte de abajo se mostrará el diagrama de la Red.

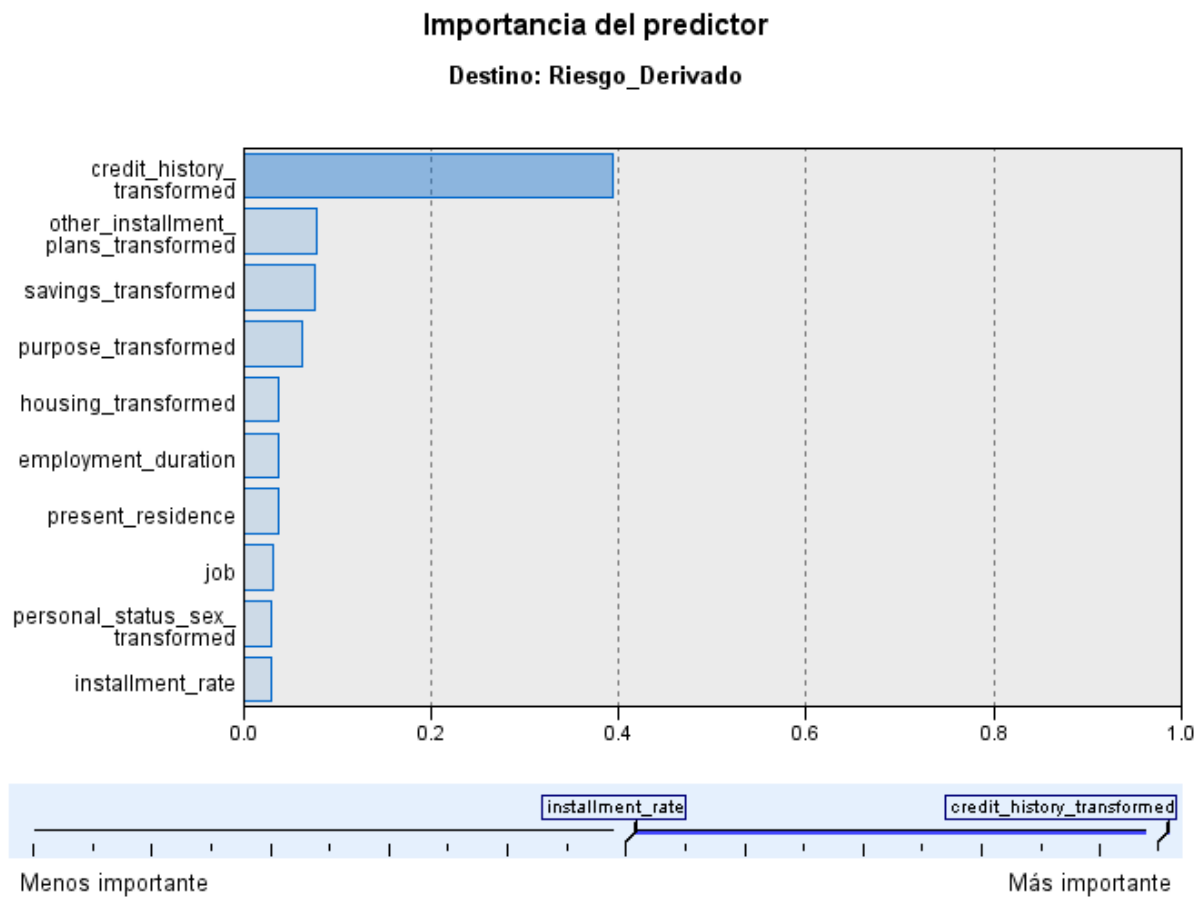


Figura 4.24: *IBS SPSS Modeler*, Importancia del predictor del modelo red neuronal.

De acuerdo con esta imagen se puede observar que el predictor que más peso tiene es el de “credit_history_transformed” el cual nos indica el historial de cumplimiento de contratos de crédito anteriores, en los siguientes dos lugares se encuentran los predictores “other_installment_plans_transformed” y “savings_transformed” que nos indican si el cliente tiene algún otro crédito o deuda con alguna otra institución bancaria y si el cliente cuenta con ahorros respectivamente.

Por lo que se puede observar se deduce que si el cliente solicitante del préstamo tiene mal historial crediticio con el cumplimiento de sus créditos es muy probable que no se le pueda otorgar el crédito que solicita.

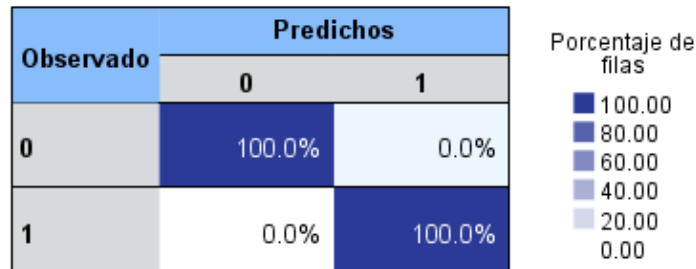
Clasificación para Riesgo_Derivado**Porcentaje correcto global = 100.0%**

Figura 4.25: *IBS SPSS Modeler*, Clasificaciones del registro Destino del modelo red neuronal.

Como se comentó en la parte de arriba, el modelo logró predecir al 100 % cada valor de la salida recordemos que el valor 0, significa que no hay un riesgo en la otorgación de un crédito, por el contrario, el valor 1 significa que hay un posible riesgo en la otorgación del crédito.

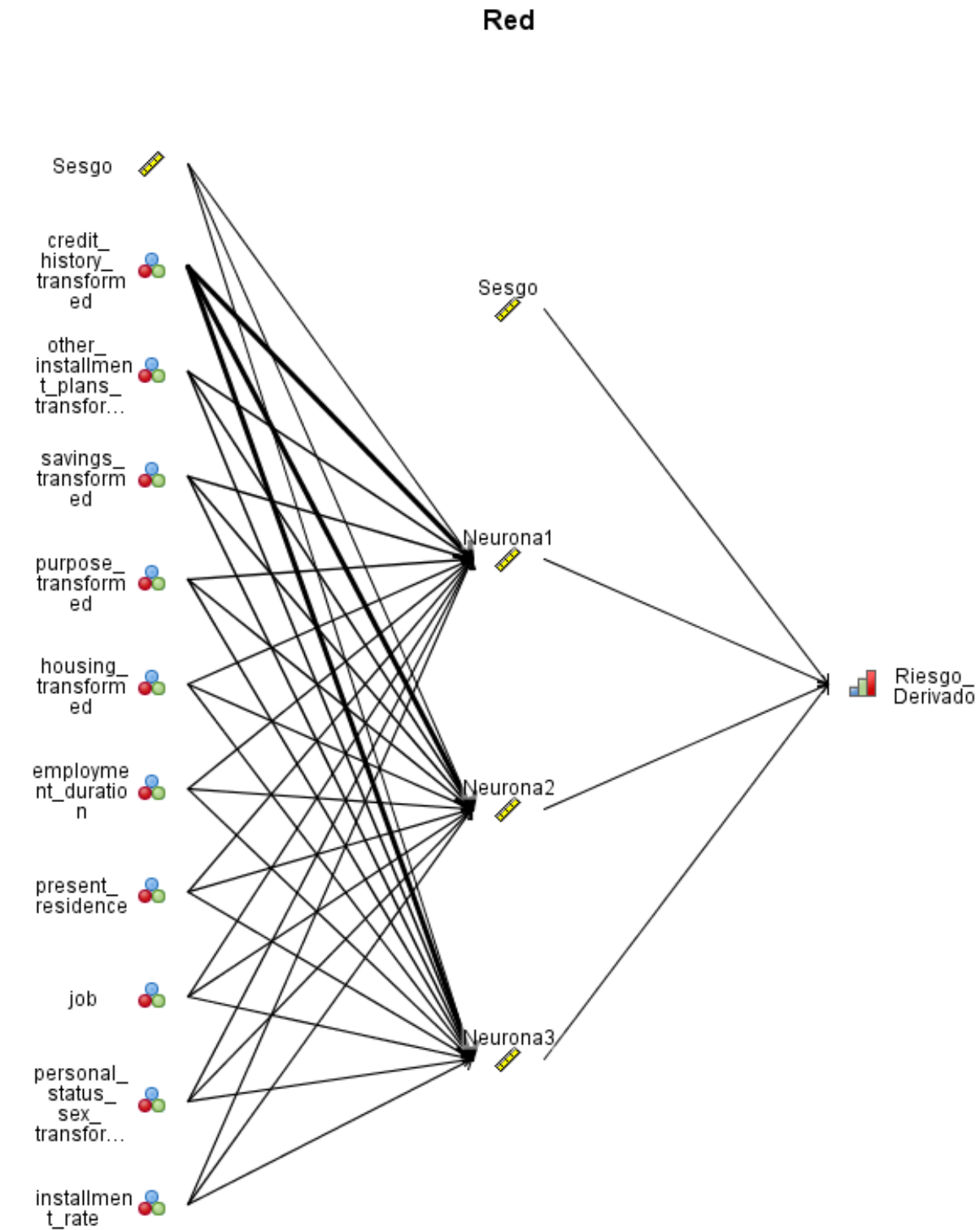


Figura 4.26: *IBS SPSS Modeler*, Red neuronal Perceptron Backpropagation.

La Figura 4.24 representa diagrama de la red Neuronal, que cuenta con una capa oculta la cual tiene 3 neuronas, donde sus principales predictores ya han sido descritos en la parte

de arriba, los cuales son “credit_history_transformed”, “other_installment_plans_transformed” y “savings_transformed”.

4.4.6. Generación del modelo: Árbol de decisión C&R

A Continuación se mostrarán los pasos a seguir para generar el modelo de Árbol de decisión C&R en la herramienta *IBM SPSS Modeler*. Como ya se planteó, se utilizarán los parámetros por defecto propuestos por cada modelo.

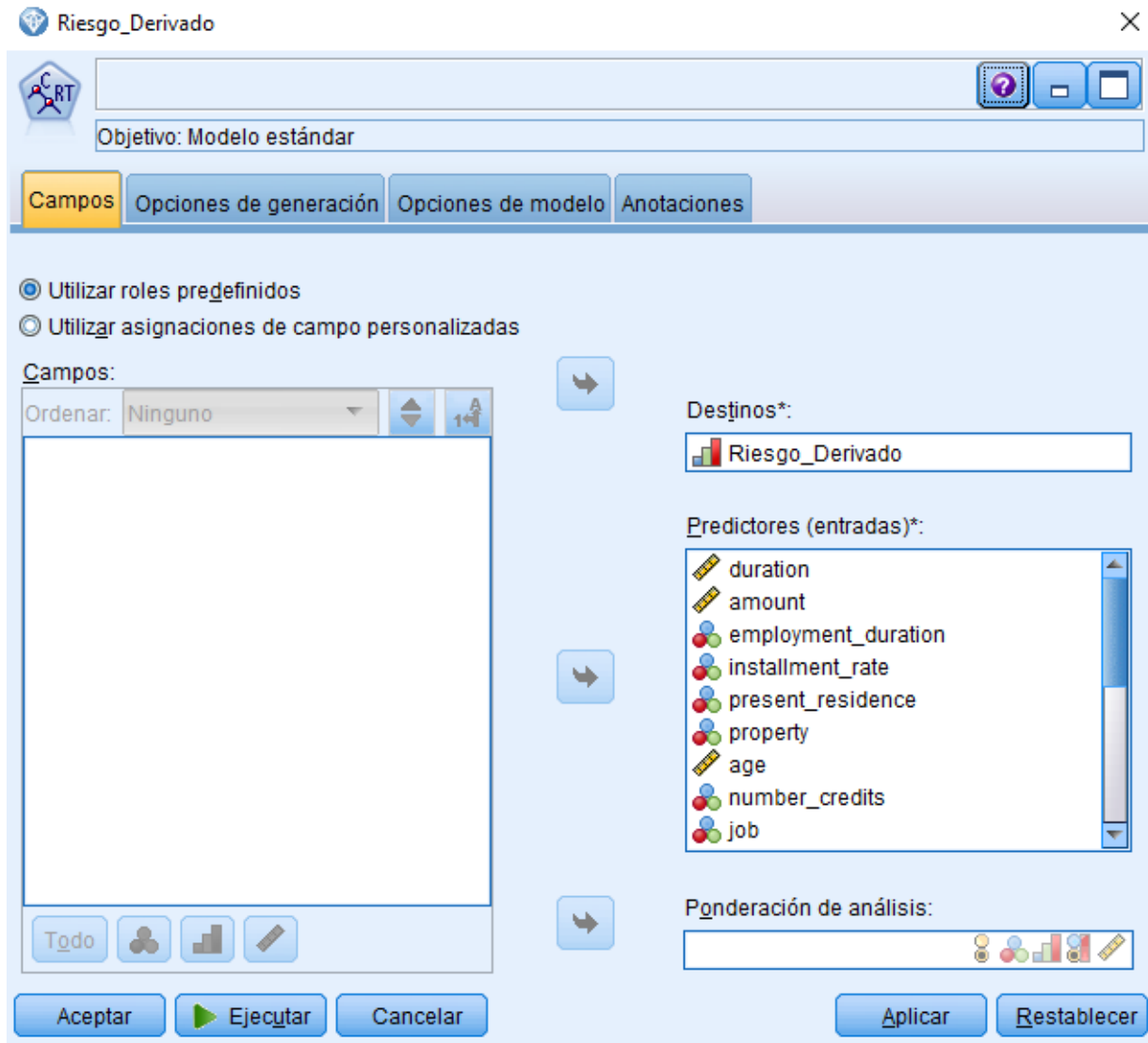


Figura 4.27: *IBM SPSS Modeler*, Generación del modelo árbol de decisión C&R.

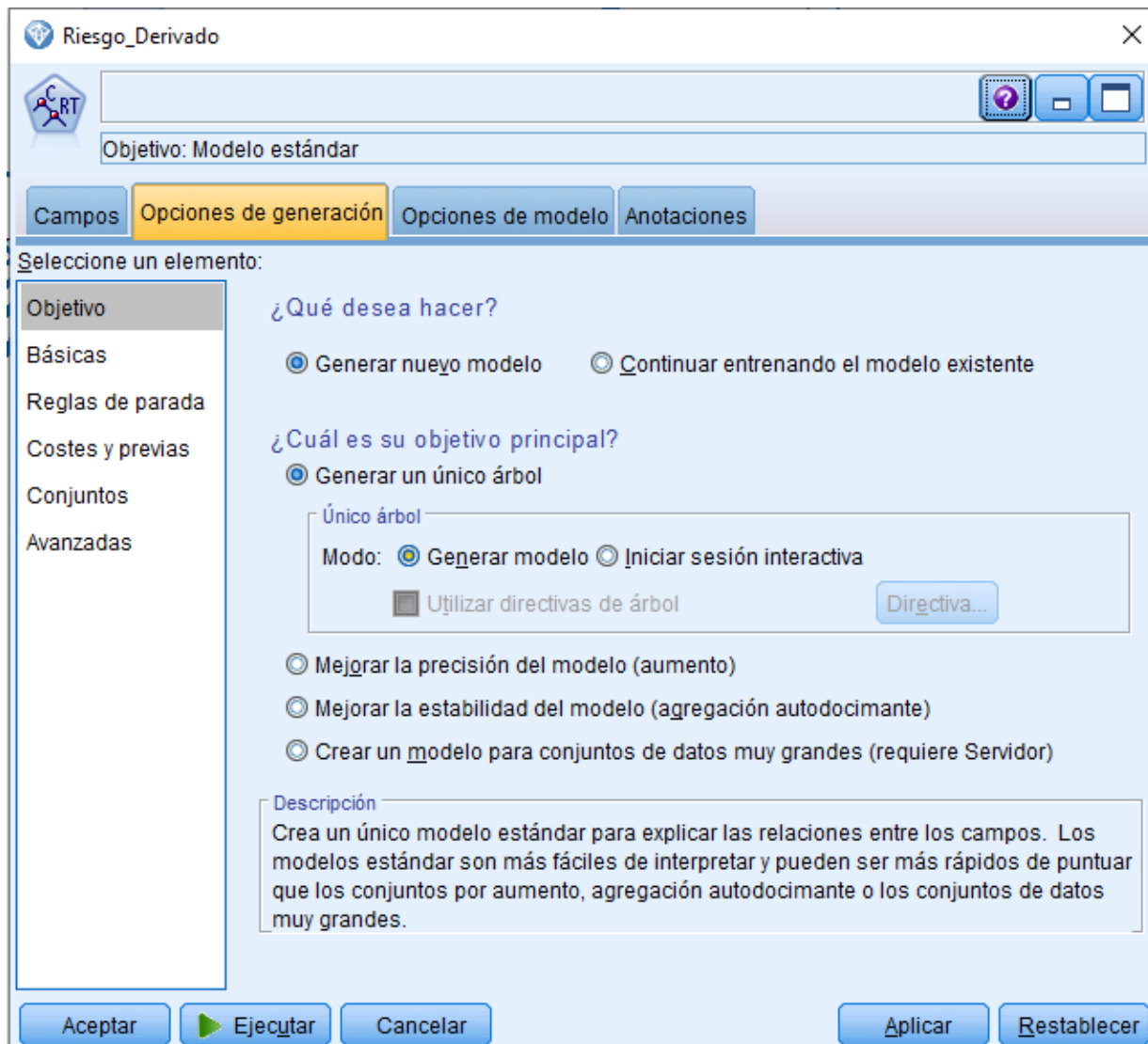


Figura 4.28: IBS SPSS Modeler, Opciones para la Generación del modelo árbol de decisión C&R.

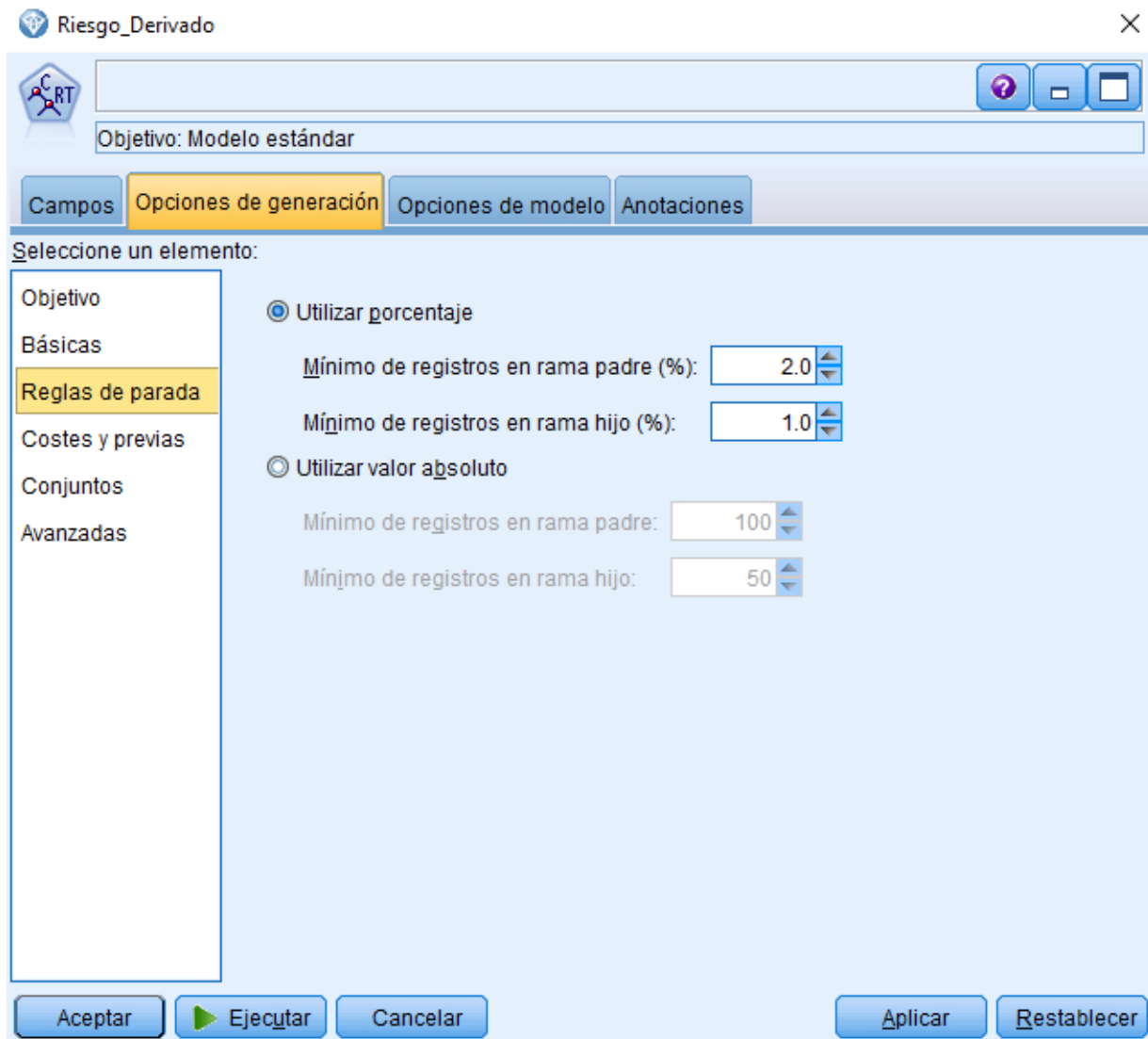


Figura 4.29: IBS SPSS Modeler, Opciones para la Generación del modelo árbol de decisión C&R 2.

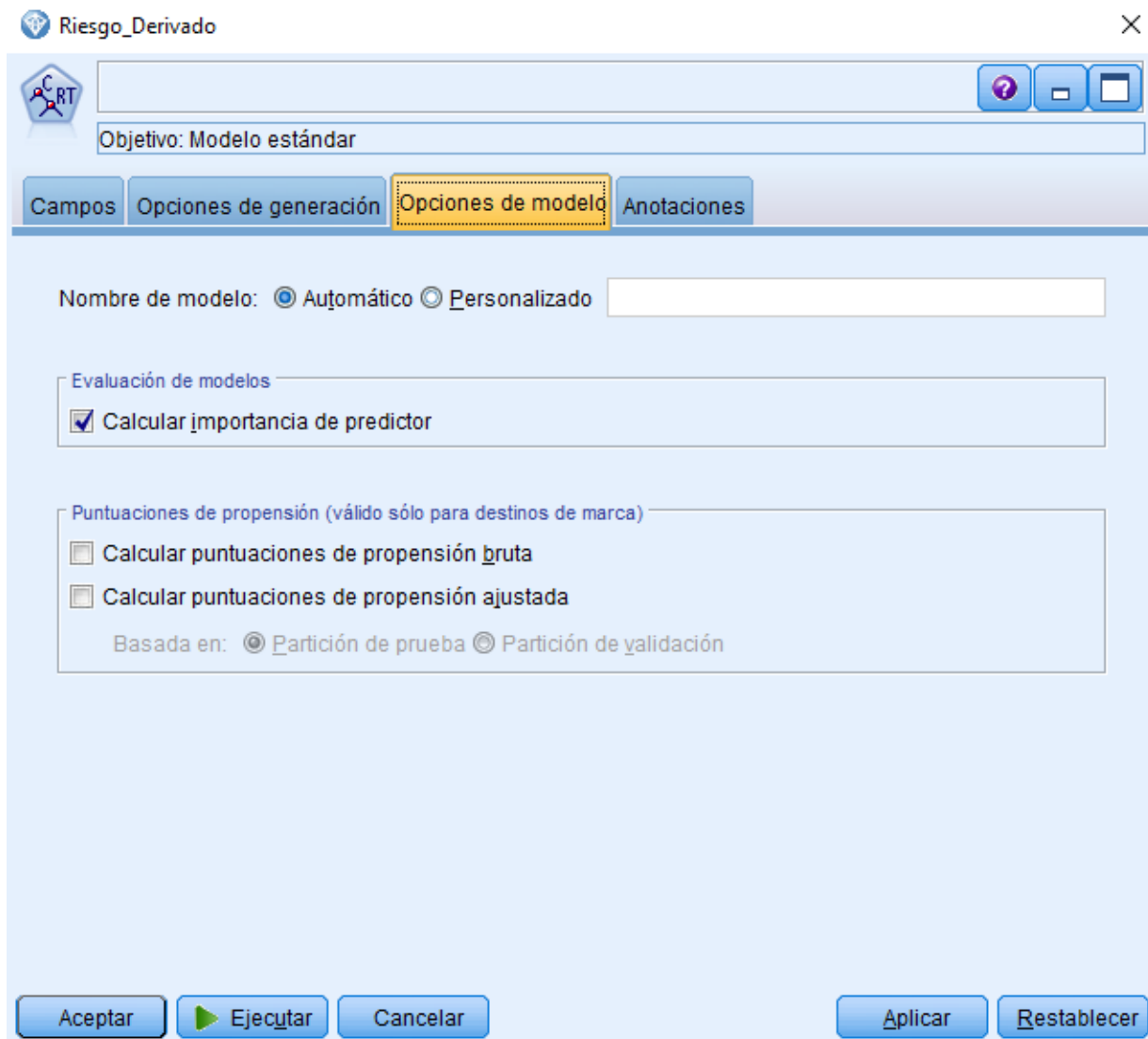


Figura 4.30: *IBS SPSS Modeler*, Opciones para la Generación del modelo árbol de decisión C&R 3.

4.4.7. Ejecución del modelo: Árbol de decisión C&R

En esta sección se mostrarán todos los datos generados por el modelo. Para el árbol de decisión se mostrará la importancia del predictor y el árbol generado por el modelo.

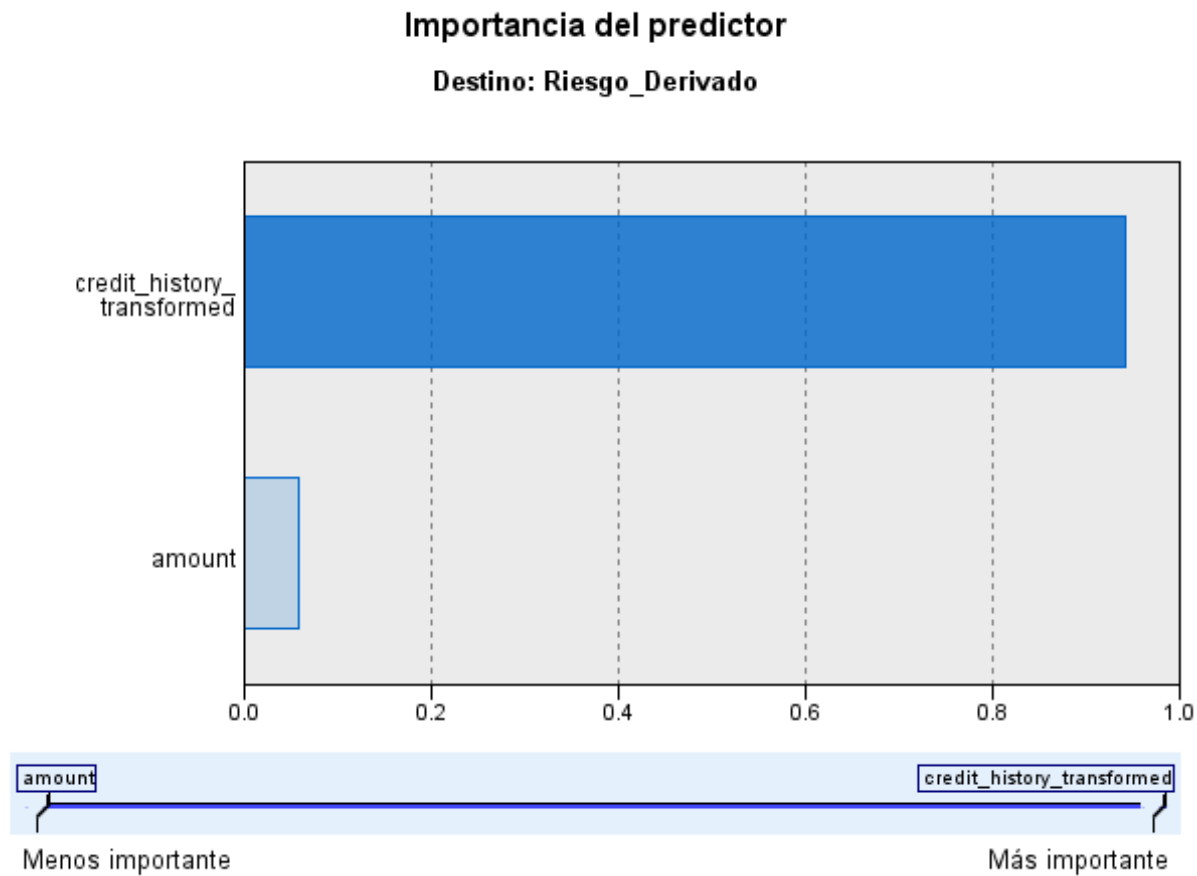


Figura 4.31: *IBS SPSS Modeler*, Importancia del predictor del modelo árbol de decisión C&R.

En este modelo se volvió a considerar al campo “credit_history_transformed” como el más importante, a diferencia con el modelo de la Red Neuronal, este modelo consideró importante el campo del monto solicitado.

La siguiente figura (Figura 4.32) muestra el árbol de decisión C&R generado después del análisis de los datos.

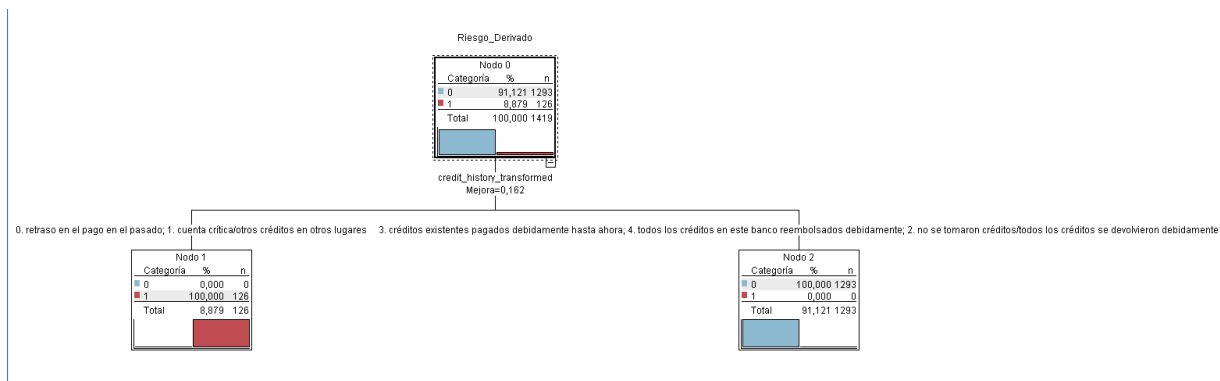


Figura 4.32: *IBS SPSS Modeler*, Modelo árbol de decisión C&R.

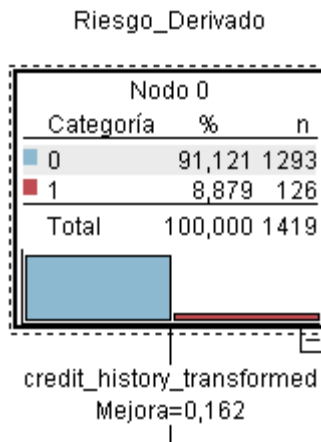


Figura 4.33: *IBS SPSS Modeler*, Nodo 0 del árbol de decisión C&R.

En la Figura 4.33 se puede observar que el nodo 0 tiene toda la cuenta de los datos derivados y las decisiones se tomarán según el predictor “credit_history_transformed”, donde los valores 0 y 1 del predictor serán las decisiones para el nodo 1, y los valores restantes son las decisiones para derivar el Nodo 2.

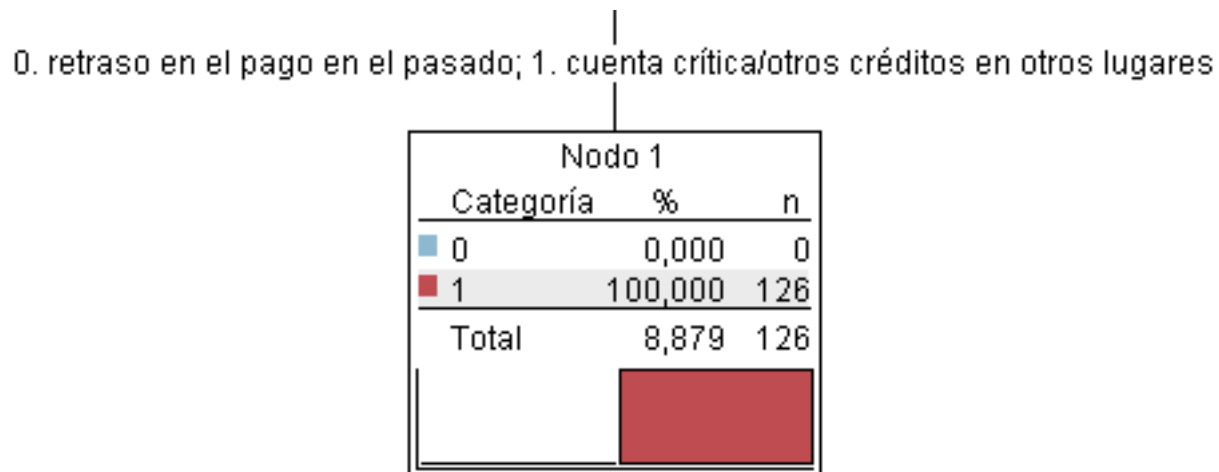


Figura 4.34: *IBS SPSS Modeler*, Nodo 1 del árbol de decisión C&R.

Como al Nodo 1 le llegan los problemas con respecto al cumplimiento de los pagos de la cuenta del cliente, la decisión según el modelo es marcar el indicador de riesgo con 1, lo que indica que la otorgación del crédito para ese cliente es de riesgo.

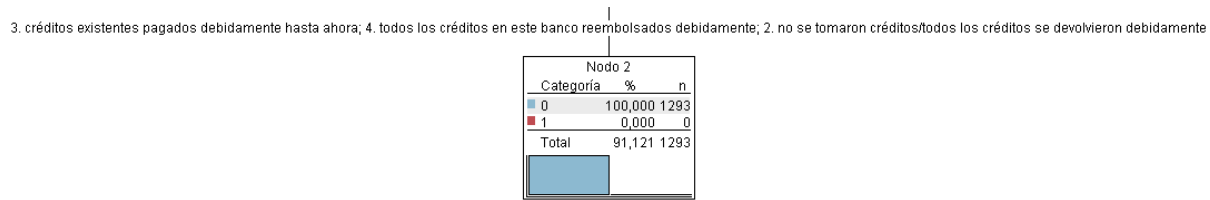


Figura 4.35: *IBS SPSS Modeler*, Nodo 2 del árbol de decisión C&R.

Al contrario, como se puede observar en el Nodo 2 (Figura 4.35) recaen todos aquellos clientes que no habían solicitado algún crédito o que no han tenido problemas con el cumplimiento de sus créditos. Por lo tanto el modelo considera que el cliente es apto para la otorgación del crédito solicitante y marca con el valor 0 el indicador de riesgo.

4.4.8. Generación del modelo: Regresión logística

A continuación se mostrarán los pasos a seguir para generar el modelo de Regresión Logística en la herramienta *IBM SPSS Modeler*. Como ya se planteó, se utilizarán los parámetros por defecto propuestos por cada modelo.

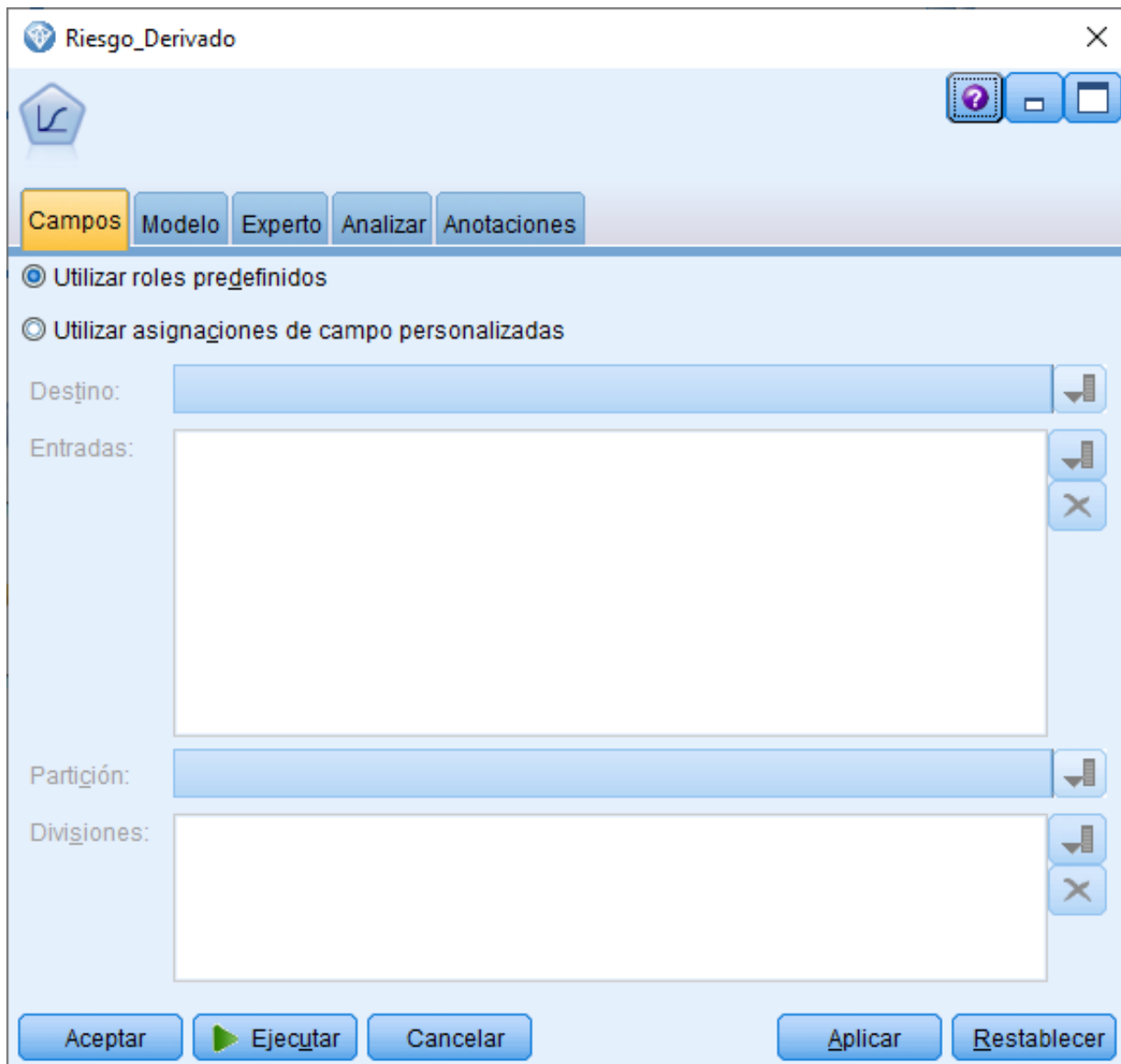


Figura 4.36: *IBS SPSS Modeler*, Generación del modelo Regresión logística.

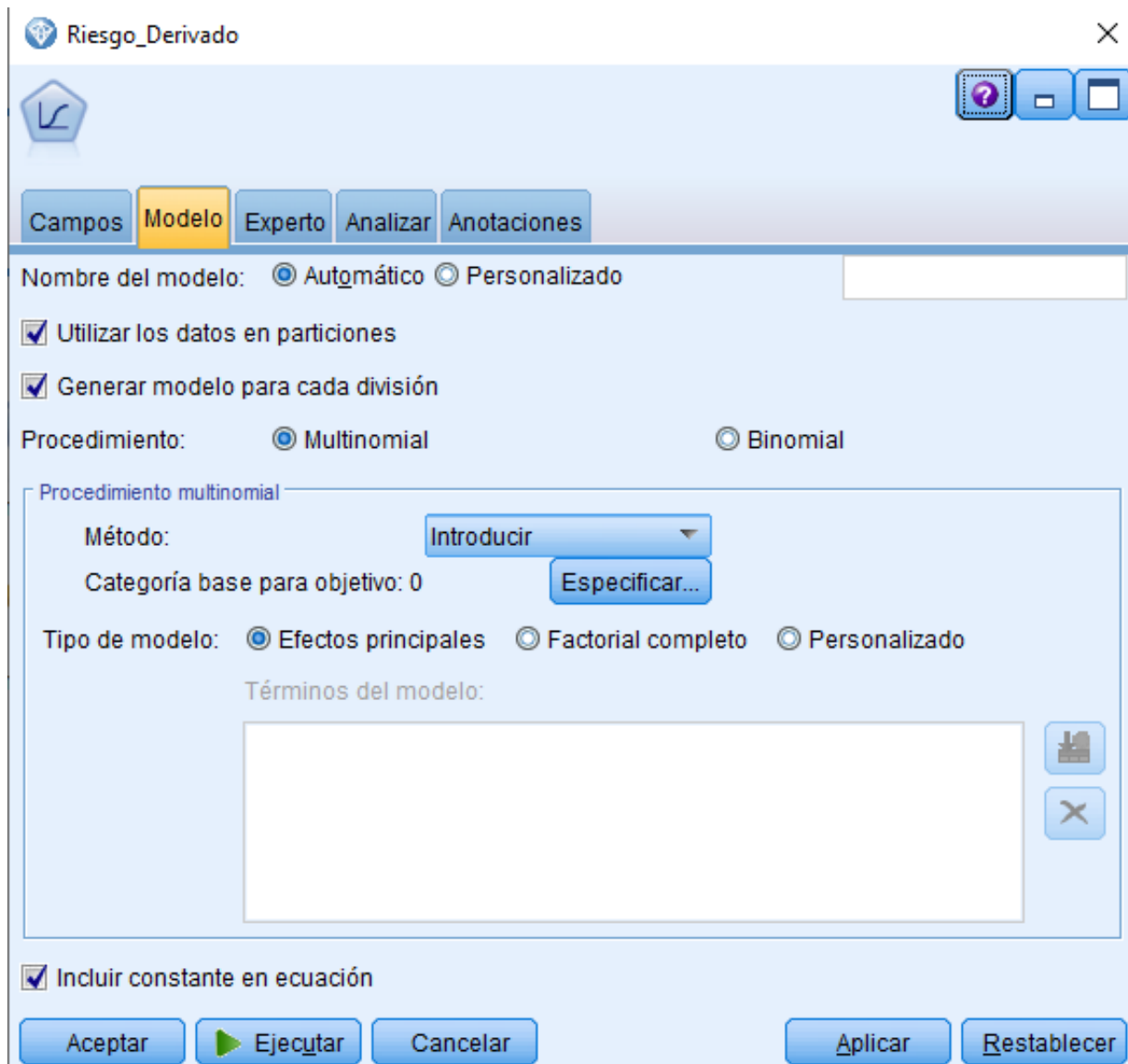


Figura 4.37: *IBS SPSS Modeler*, Opciones para la Generación del modelo Regresión Logística.

4.4.9. Ejecución del modelo: Algoritmo de Regresión Logística

En esta sección se mostrarán todos los datos generados por el modelo. Para el Algoritmo de Regresión Logística se mostrará el análisis del modelo.

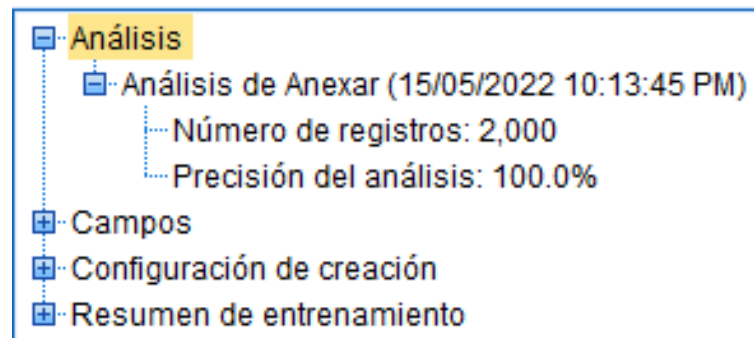
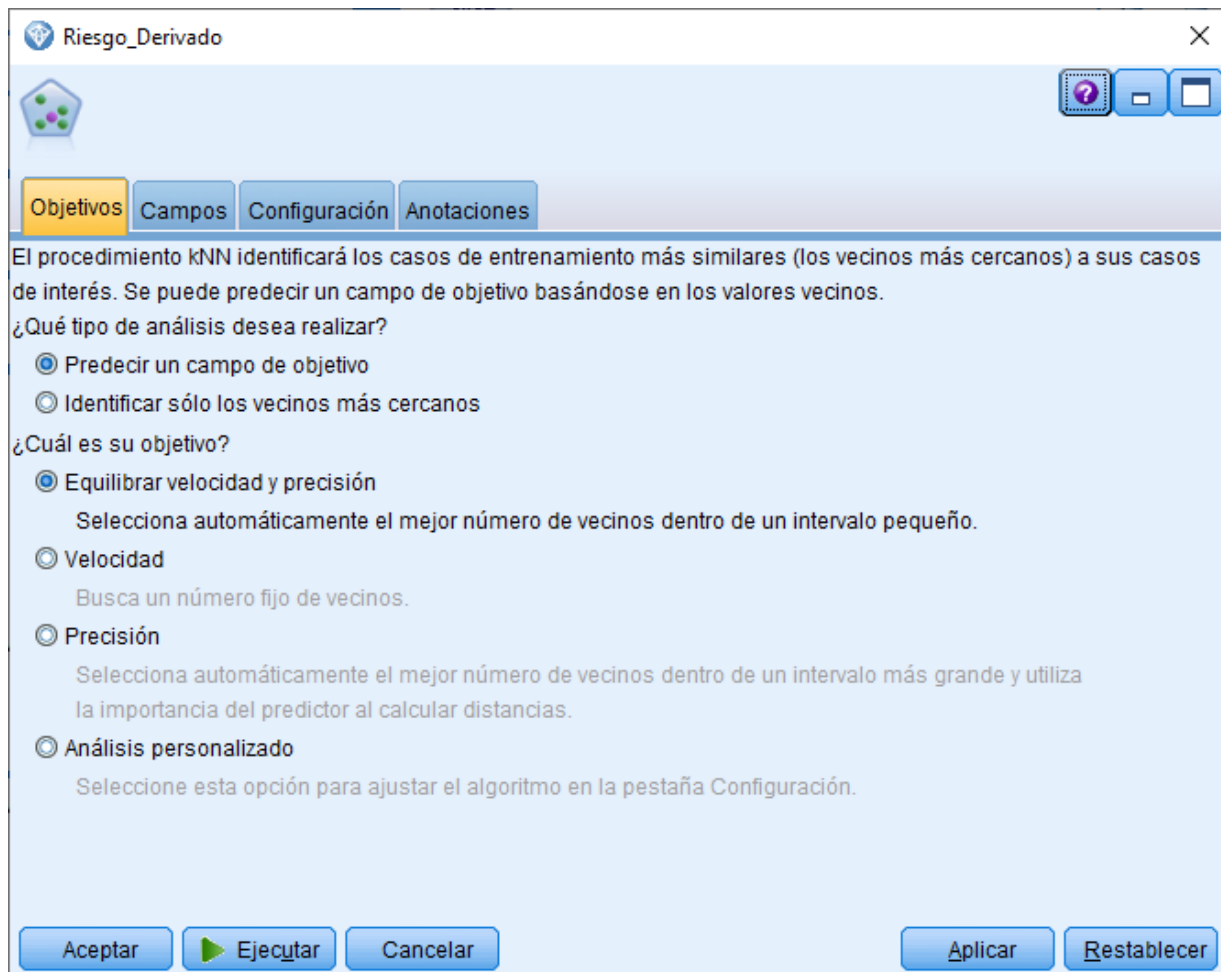


Figura 4.38: *IBS SPSS Modeler*, Análisis del modelo Regresión Logística.

De acuerdo con la Figura 4.38 se han analizado los 2,000 registros resultantes de la preparación de los datos obteniendo una precisión del análisis del 100 %, otro caso que es ideal y que se esperaba en este modelo, ya que es principalmente utilizado en la clasificación de datos.

4.4.10. Generación del modelo: Algoritmo KNN

A continuación se mostrarán los pasos a seguir para generar el modelo KNN en la herramienta *IBM SPSS Modeler*. Como ya se planteó, se utilizarán los parámetros por defecto propuestos por cada modelo.

Figura 4.39: *IBS SPSS Modeler*, Generación del modelo KNN.

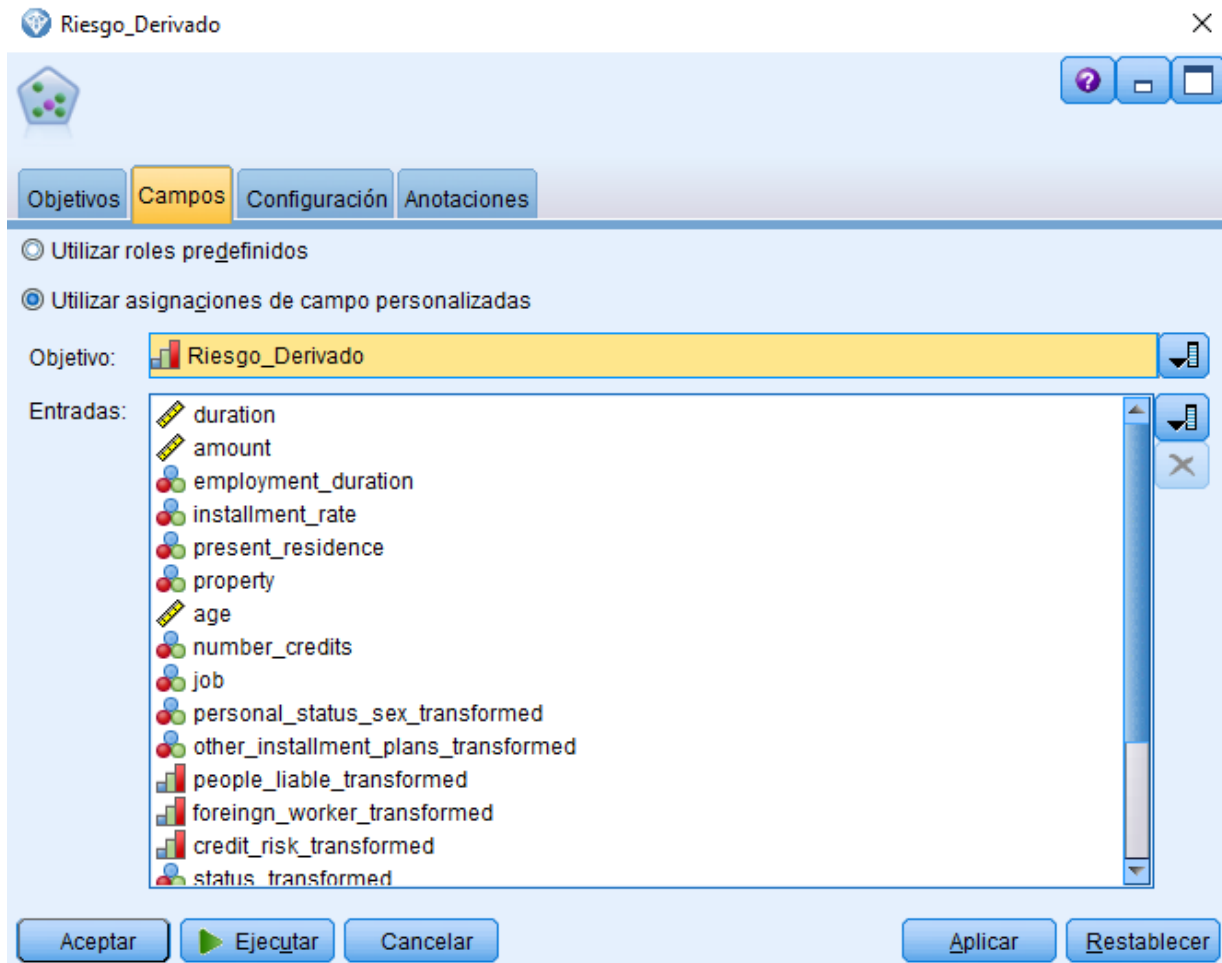


Figura 4.40: *IBS SPSS Modeler*, Selección del objetivo para el modelo KNN.

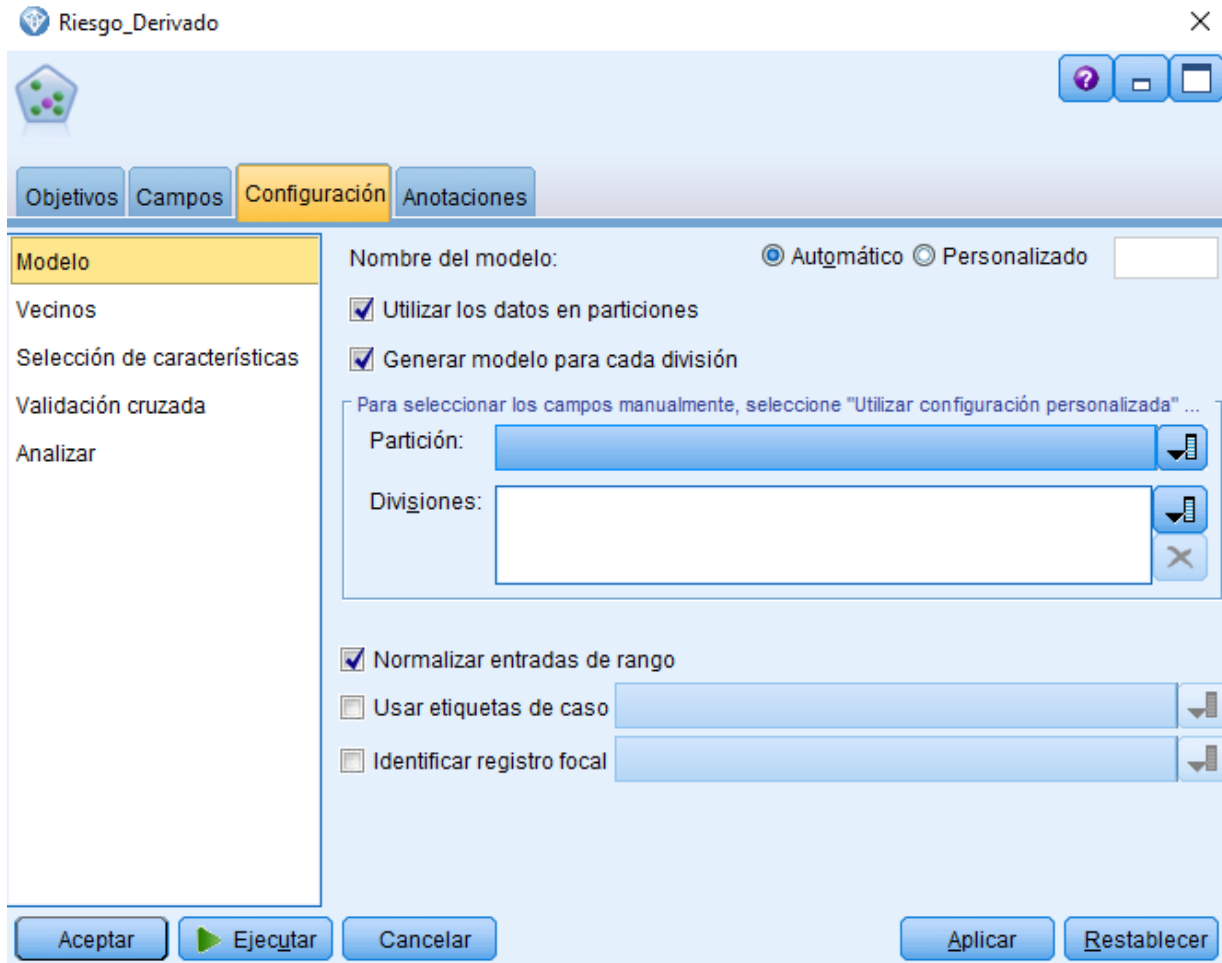


Figura 4.41: *IBS SPSS Modeler*, Opciones para la generación del modelo KNN.

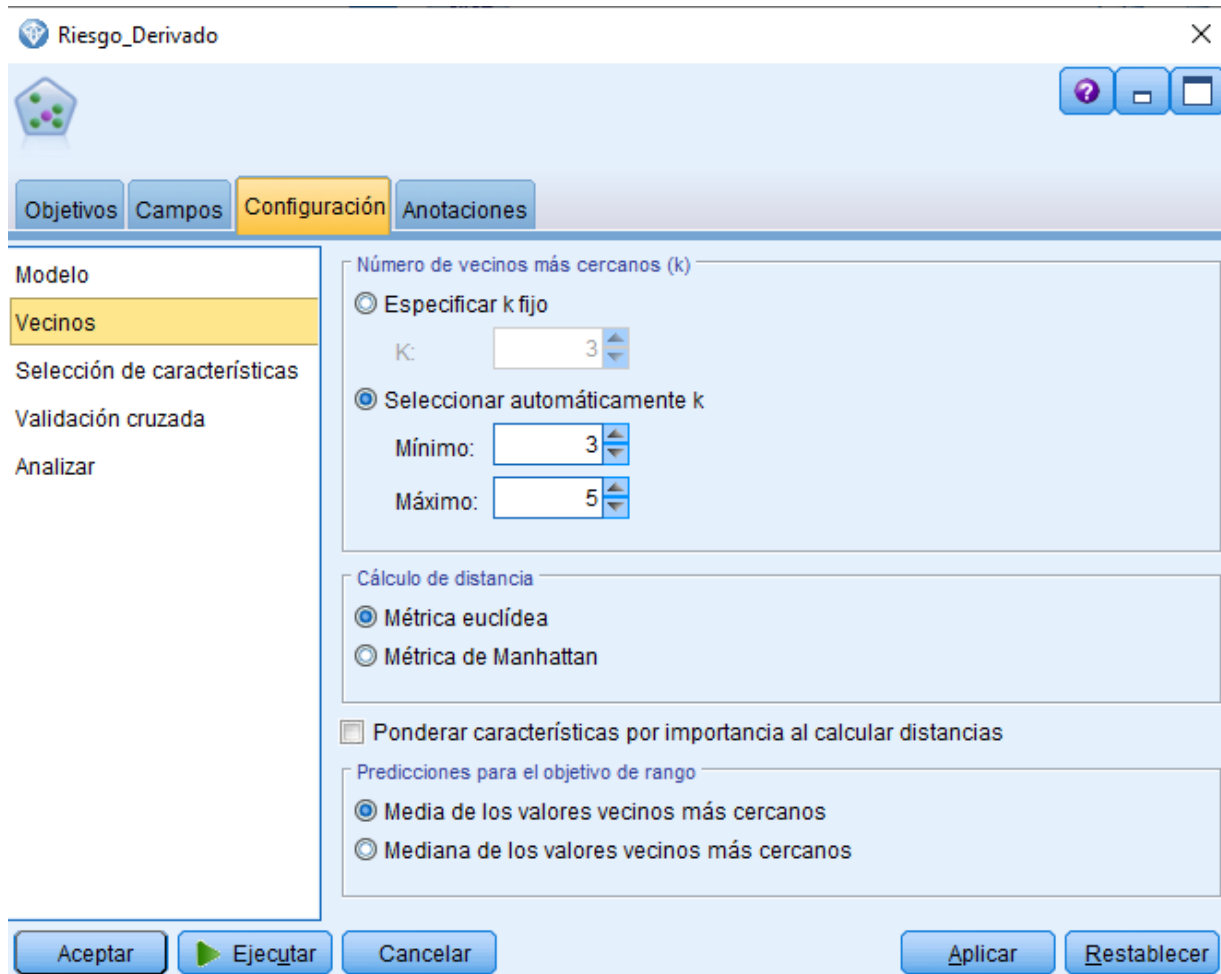


Figura 4.42: IBS SPSS Modeler, Opciones para la generación del modelo KNN 2.

4.4.11. Ejecución del modelo: Algoritmo KNN

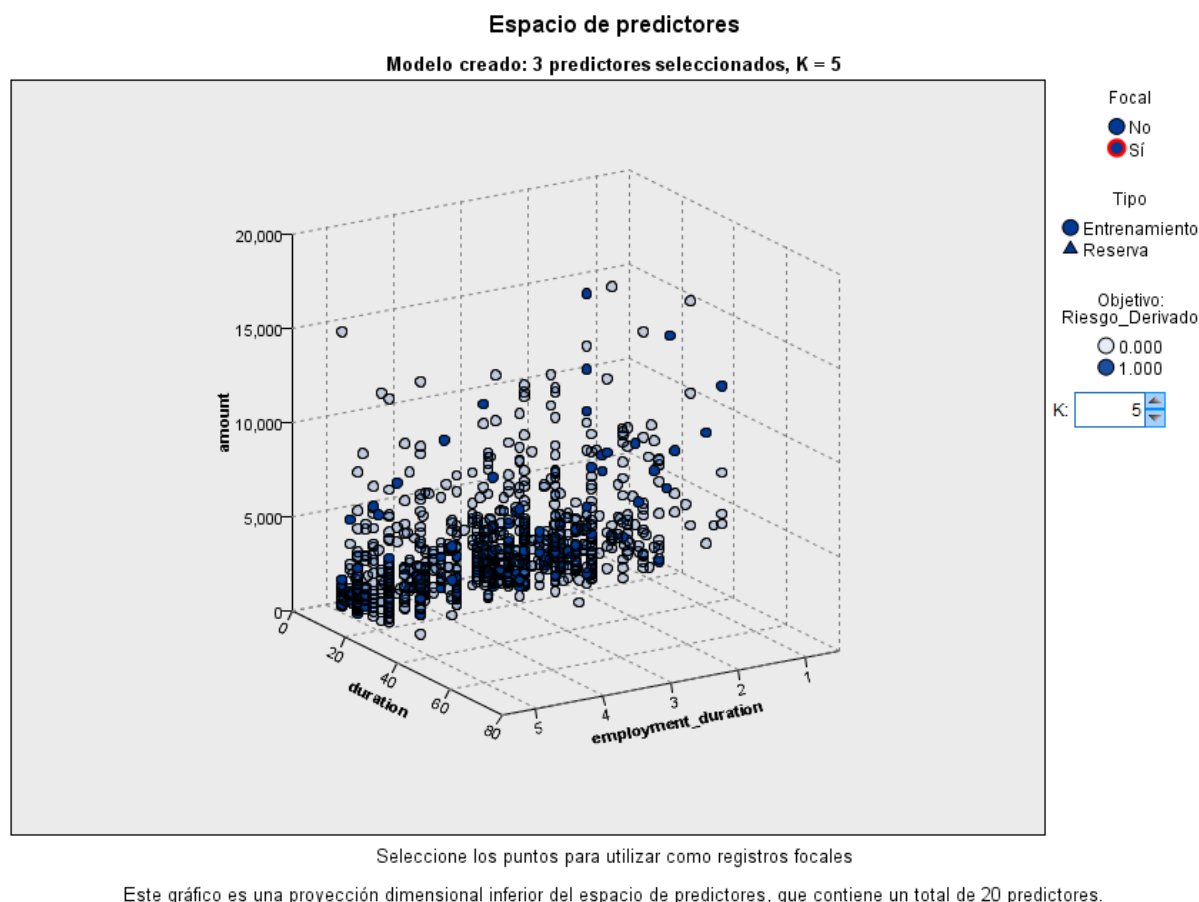


Figura 4.43: *IBS SPSS Modeler*, Resultados del modelo KNN.

4.4.12. Generación del modelo: Algoritmo de Regresión Lineal

A continuación se mostrarán los pasos a seguir para generar el modelo de Regresión Lineal en la herramienta *IBM SPSS Modeler*. Como ya se planteó, se utilizarán los parámetros por defecto propuestos por cada modelo.

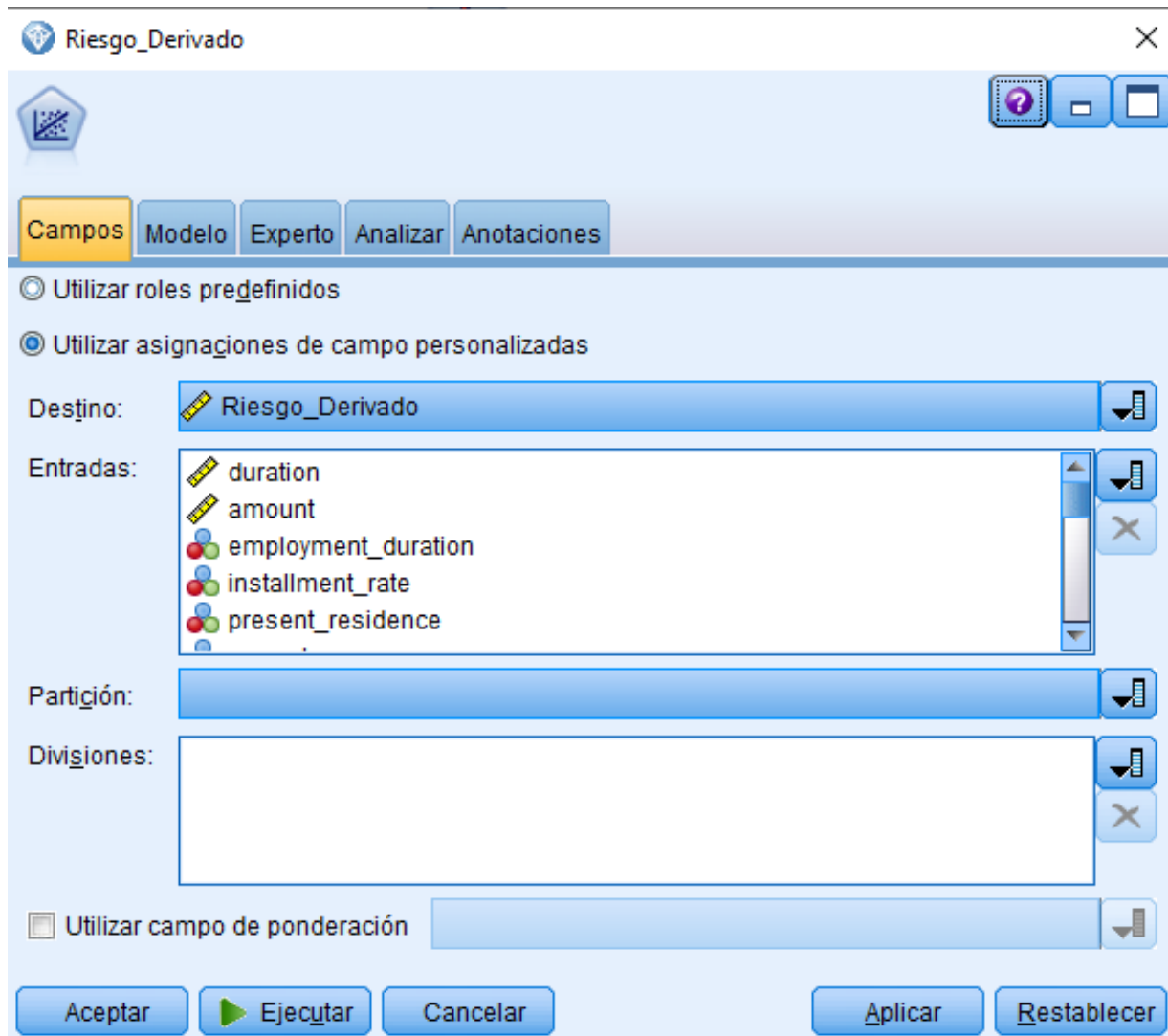


Figura 4.44: *IBS SPSS Modeler*, Selección del objetivo para el modelo Algoritmo de Regresión Lineal.

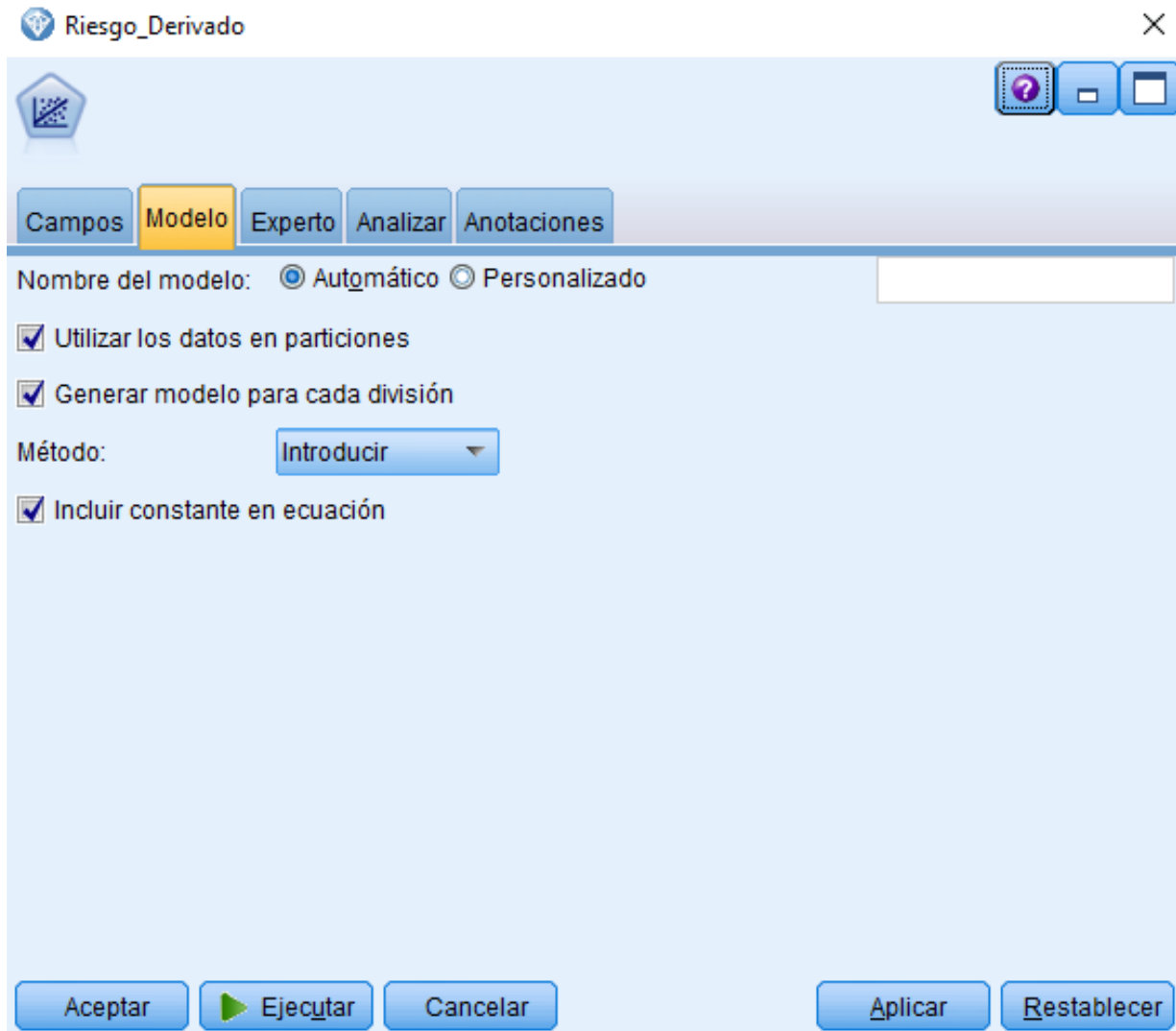


Figura 4.45: *IBS SPSS Modeler*, Opciones para la generación del modelo Algoritmo de Regresión Lineal.

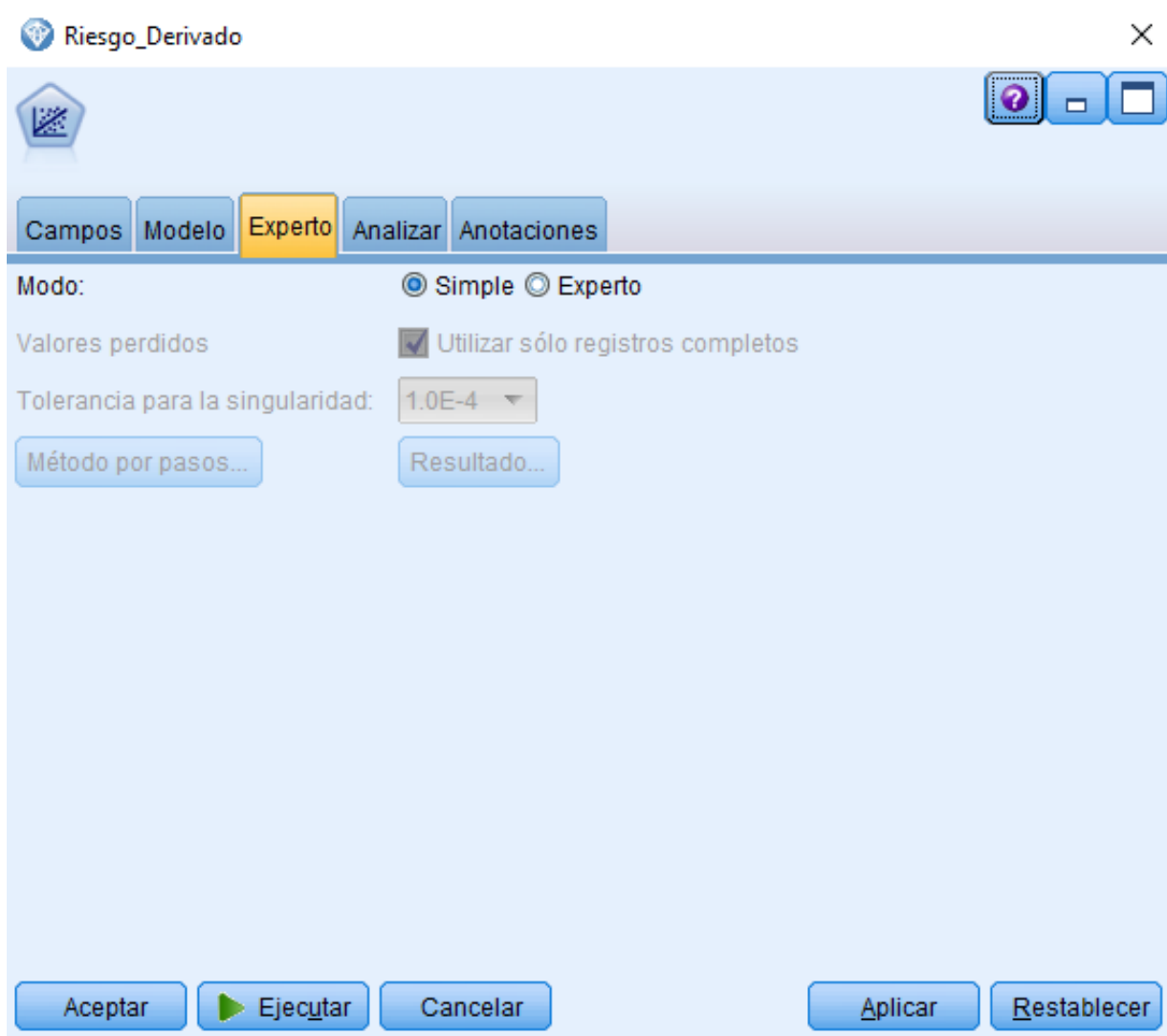


Figura 4.46: *IBS SPSS Modeler*, Opciones para la generación del modelo Algoritmo de Regresión Lineal 2.

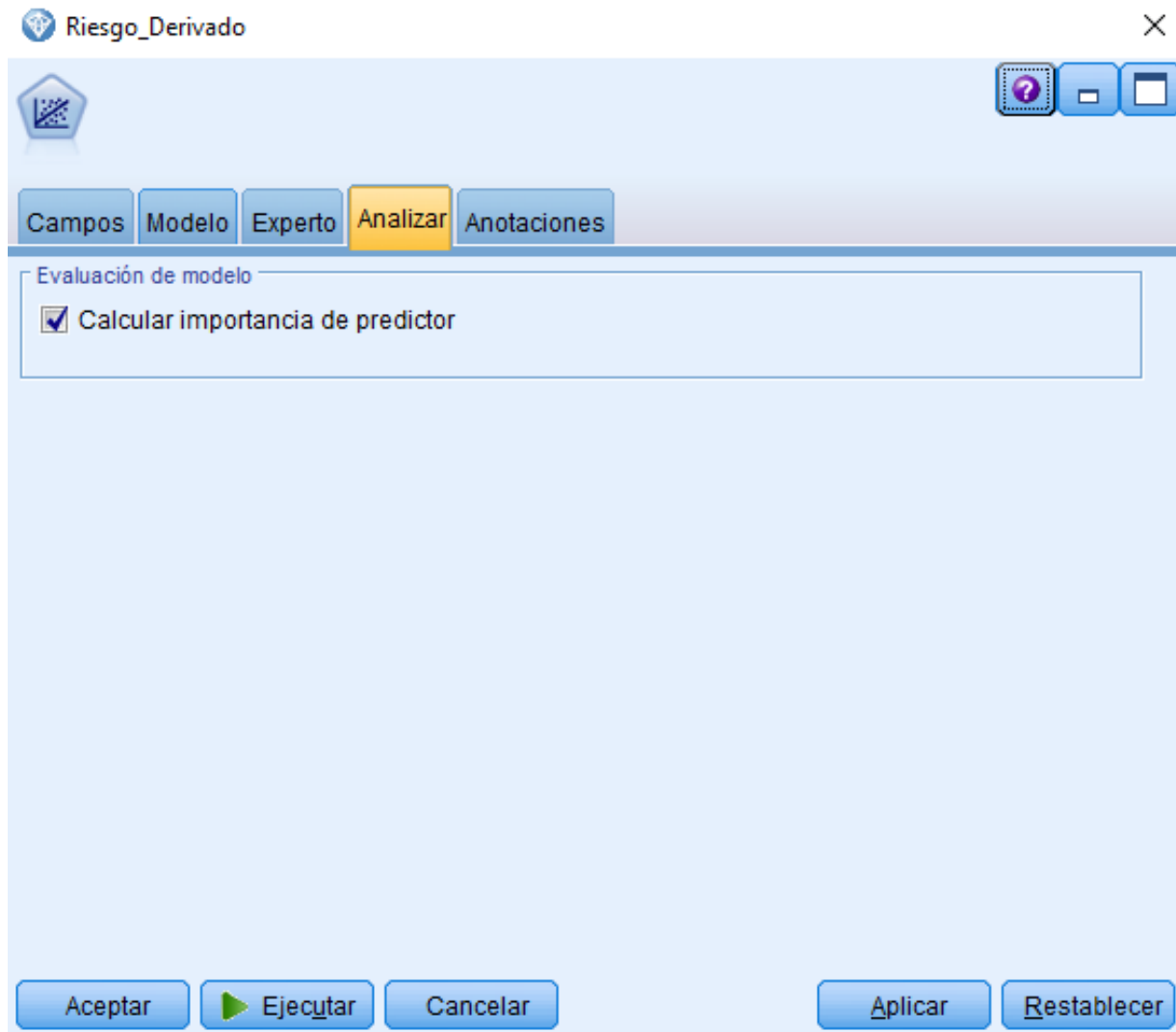


Figura 4.47: *IBS SPSS Modeler*, Opciones para la generación del modelo Algoritmo de Regresión Lineal 3.

4.4.13. Ejecución del modelo: Algoritmo de Regresión Lineal

En esta sección se mostrarán todos datos generados por el modelo. Para el algoritmo de Regresión Lineal se mostrará la importancia del predictor.

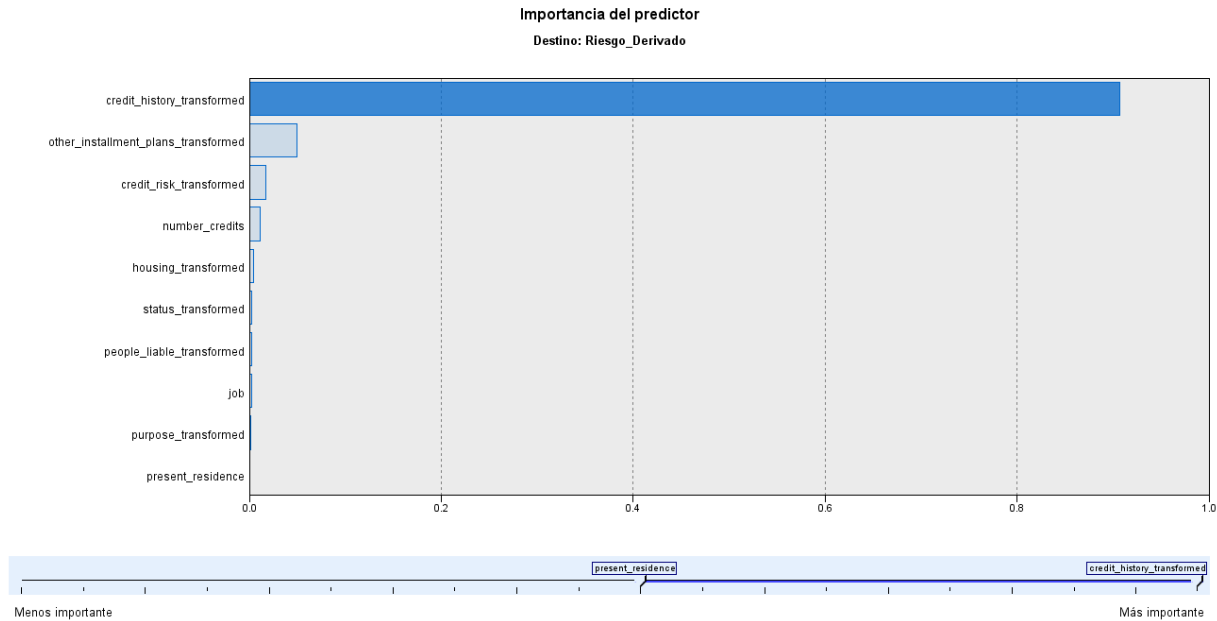


Figura 4.48: *IBS SPSS Modeler*, Resultados del modelo Algoritmo de Regresión Lineal.

Como se puede observar en la imagen generada por el modelo (Figura 4.48), el predictor de más importancia es el de “credit_history_transformed”. Este es otro modelo más que toma este predictor como el más importante, llegando a la conclusión de que si un cliente que solicita un crédito tiene un mal historial es muy probable que se le rechace la solicitud del crédito, por otra parte si el cliente cuenta con un buen historial crediticio los modelos arrojan que es posible otorgar el crédito, aunque se puede tener en consideración algún otro filtro que haga pesar más a otro tipo de predictor con el fin de estar 100 % precavidos de que el cliente cumplirá con el pago del crédito.

4.5. Evaluación de los Modelos

En esta sección se evaluará cada uno de los modelos generados por la herramienta *IBM SPSS Modeler*.

Evaluación del modelo Red Neuronal Perceptron Backpropagation

Resultados para el campo de resultado Riesgo_Derivado		
Comparando \$N-Riesgo_Derivado con Riesgo_Derivado		
Correctos	2,000	100%
Erróneos	0	0%
Total	2,000	
Evaluación del rendimiento		
0	0.093	
1	2.419	

Figura 4.49: *IBS SPSS Modeler*, Evaluación de los resultados del modelo Red Neuronal Perceptron Backpropagation.

Evaluación del modelo Árbol de Decisión C&R

Resultados para el campo de resultado Riesgo_Derivado		
Comparando \$R-Riesgo_Derivado con Riesgo_Derivado		
Correctos	2,000	100%
Erróneos	0	0%
Total	2,000	
Evaluación del rendimiento		
0	0.093	
1	2.419	

Figura 4.50: *IBS SPSS Modeler*, Evaluación de los resultados del modelo Árbol de Decisión C&R.

Evaluación del modelo Algoritmo de Regresión Logística

Resultados para el campo de resultado Riesgo_Derivado		
Comparando \$L-Riesgo_Derivado con Riesgo_Derivado		
Correctos	2,000	100%
Erróneos	0	0%
Total	2,000	
Evaluación del rendimiento		
0	0.093	
1	2.419	

Figura 4.51: *IBS SPSS Modeler*, Evaluación de los resultados del modelo Algoritmo de Regresión Logística.

Evaluación del modelo Algoritmo KNN

Resultados para el campo de resultado Riesgo_Derivado

Comparando \$KNN-Riesgo_Derivado con Riesgo_Derivado

Correctos	1,928	96.4%
Erróneos	72	3.6%
Total	2,000	

Evaluación del rendimiento

0	0.055
1	2.401

Figura 4.52: *IBS SPSS Modeler*, Evaluación de los resultados del modelo Algoritmo KNN.

Evaluación del modelo Algoritmo de Regresión Lineal

Resultados para el campo de resultado Riesgo_Derivado

Comparando \$E-Riesgo_Derivado con Riesgo_Derivado

Error mínimo	-0.536
Error máximo	0.477
Error promedio	-0.0
Error absoluto promedio	0.108
Desviación estándar	0.152
Correlación lineal	0.845
Ocurrencias	2,000

Figura 4.53: *IBS SPSS Modeler*, Evaluación de los resultados del modelo Algoritmo de Regresión Lineal.

Observaciones de la Evaluación de los Modelos

Como se puede observar, los modelos como Red Neuronal Perceptron Backpropagation, Árbol de decisión C&R y el modelo de Regresión Logística nos otorga resultados muy favorables en la pronosticación y clasificación de nuestra variable destino. Teniendo un rendimiento del 100 %, este resultado es gracias a que estos modelos son especializados en clasificación, los ideales para este proyecto.

Por otra parte el Algoritmo KNN y el Algoritmo de Regresión Lineal aunque no nos ofrecen malos resultados, estos tienen cierto porcentaje de error, el caso del algoritmo KNN es de 3.6 % y el de la regresión lineal es de 0.155 debido a que el tipo de los datos no es el adecuado para estos modelos, ya que la mayoría de los datos con los que se trabajaron son de tipo Nominal y Ordinal, dificultando el cálculo de los Modelos.

Para complementar esta evaluación a cada modelo se le presentará una serie de datos los cuales fueron extraídos y eliminados del Dataset después de realizar la limpieza automática de los datos, con el fin de comprobar la eficiencia y eficacia de los modelos. Estos datos estan almacenados en formato ".csv"

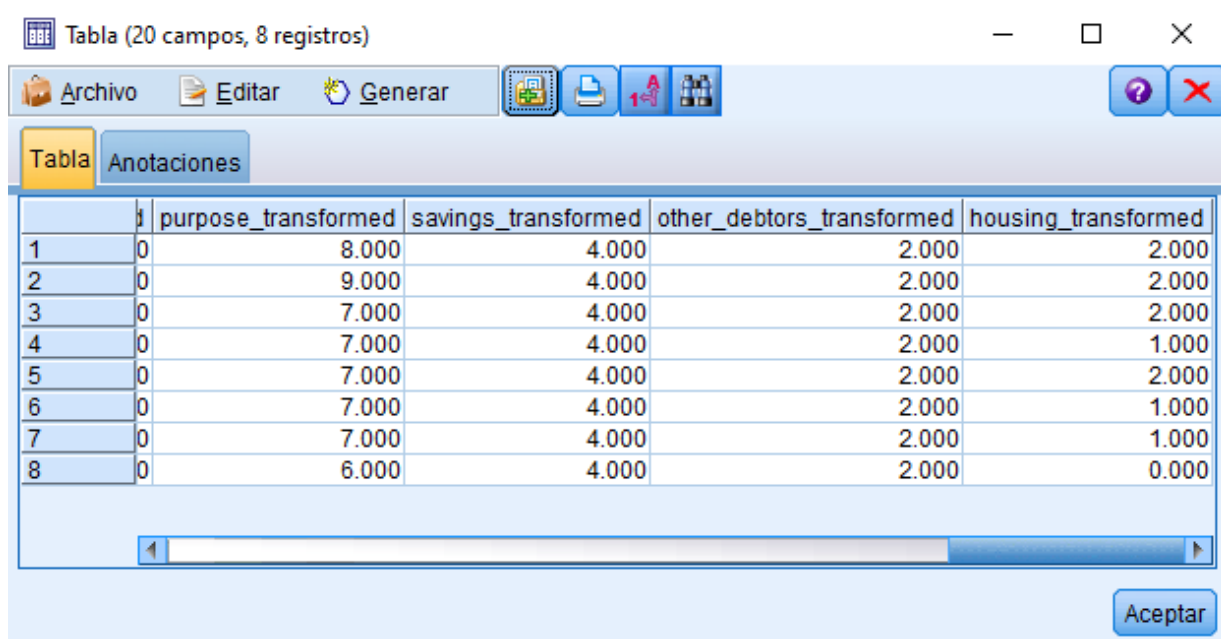


Tabla (20 campos, 8 registros)

Archivo Editar Generar

Tabla Anotaciones

		purpose_transformed	savings_transformed	other_debtors_transformed	housing_transformed
1	0	8.000	4.000	2.000	2.000
2	0	9.000	4.000	2.000	2.000
3	0	7.000	4.000	2.000	2.000
4	0	7.000	4.000	2.000	1.000
5	0	7.000	4.000	2.000	2.000
6	0	7.000	4.000	2.000	1.000
7	0	7.000	4.000	2.000	1.000
8	0	6.000	4.000	2.000	0.000

Aceptar

Figura 4.54: Datos para la evaluación.

Para evitar problemas con la pronosticación de los modelos, antes de las pruebas se estableció el tipo de cada campo de estos datos de prueba, la Figura 4.55 nos muestra el resultado:

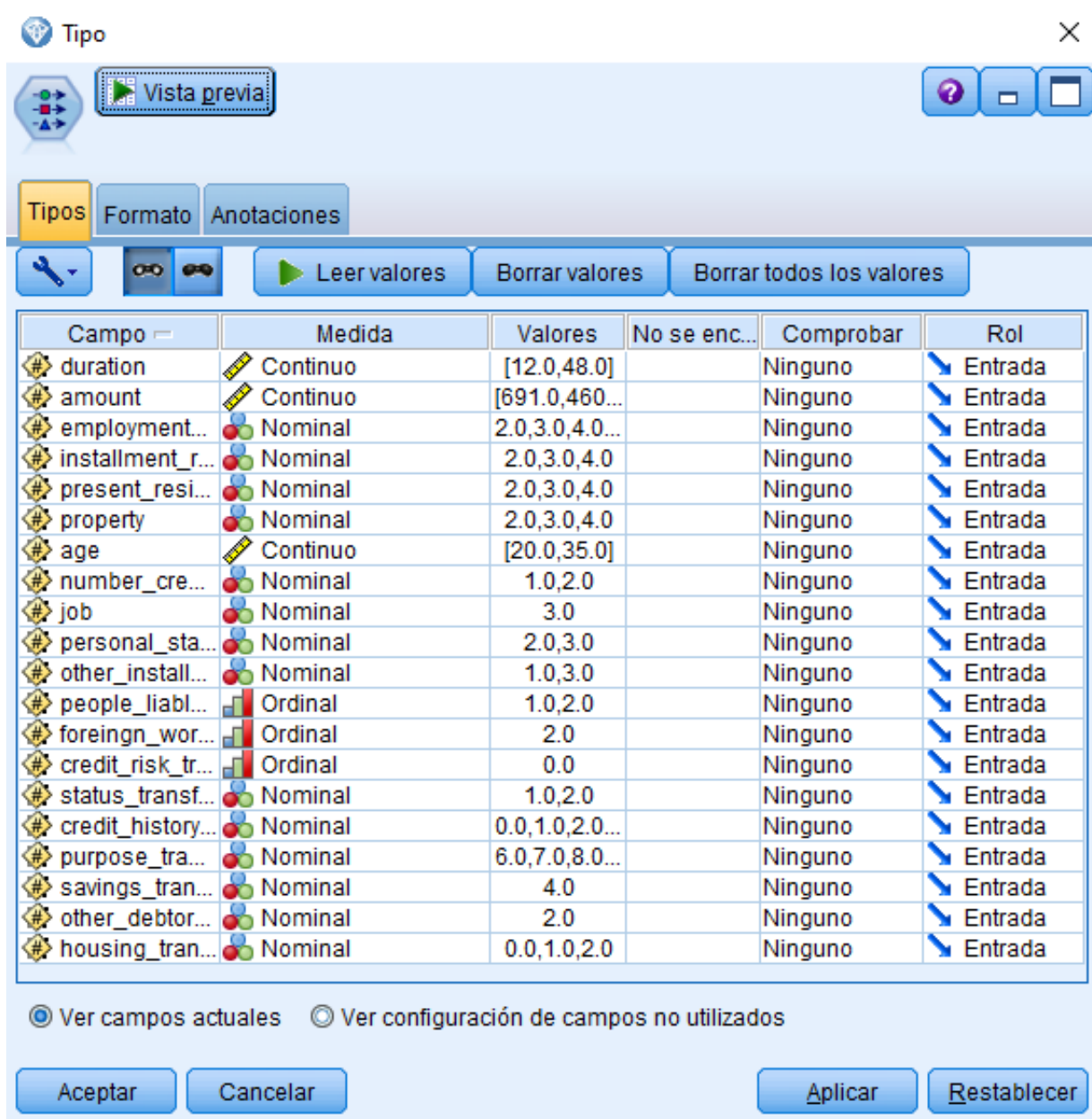


Figura 4.55: Tipos de los campos de los datos de prueba.

Una vez tipificados los datos están listos para ser probados por los modelos generados (Ver Sección 4.4.3). La Figura 4.56 es la ruta creada en la herramienta *IBM SPSS Modeler* para esta evaluación.

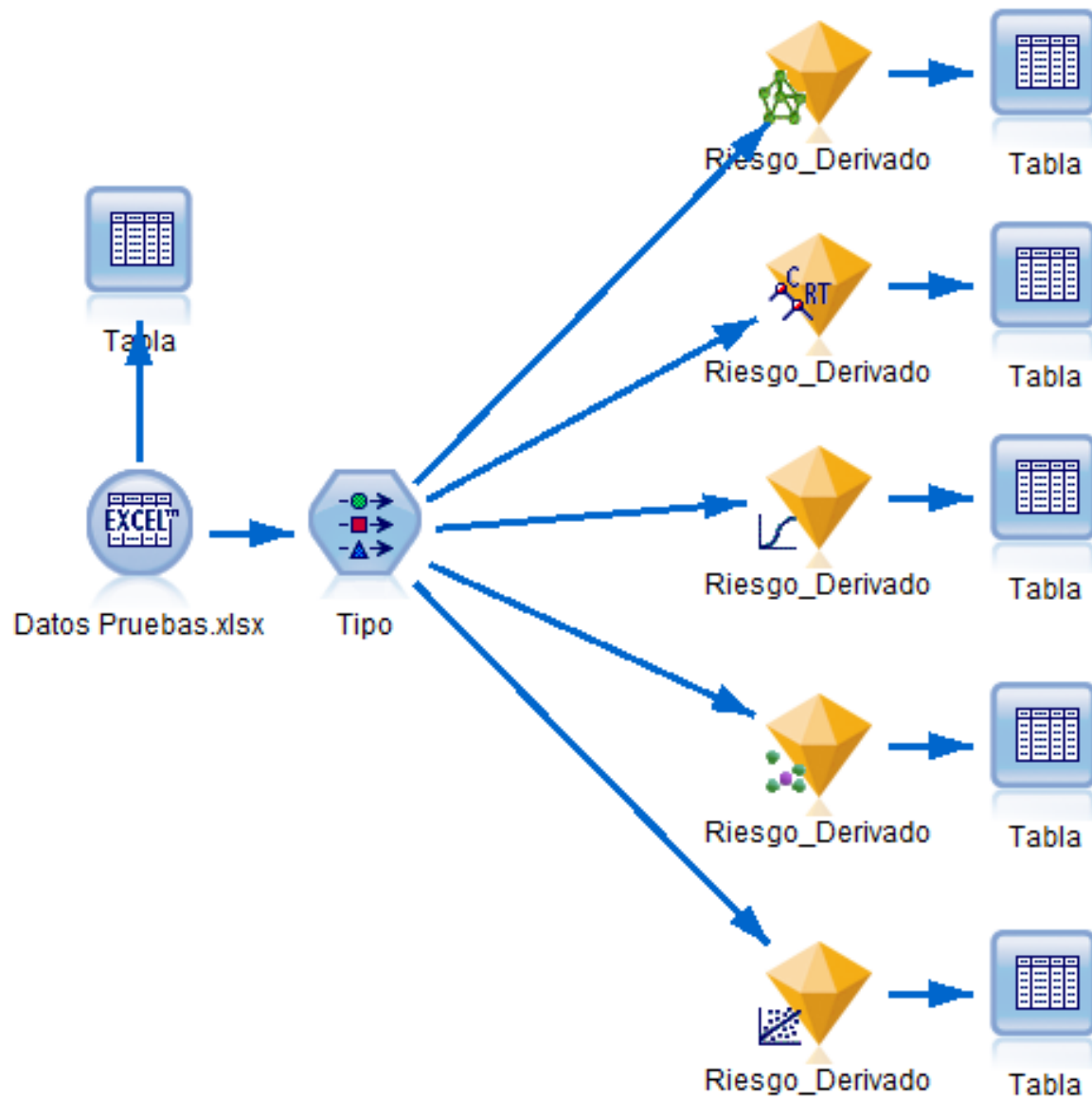


Figura 4.56: Tipos de los campos de los datos de prueba.

La siguiente figura (Figura 4.57) muestra la salida esperada para estos datos:

	A
1	Riesgo_Derivado
2	0
3	0
4	1
5	1
6	0
7	0
8	1
9	1

Figura 4.57: Salida esperada en la evaluación de los modelos.

Evaluación con nuevos datos del modelo Red Neuronal Perceptron Back-propagation



Figura 4.58: IBM SPSS Modeler, Evaluación del modelo Red Neuronal Perceptron Back-propagation con nuevos datos.

Tabla (22 campos, 8 registros) #4						
	transformed	other_debtors_transformed	housing_transformed	\$N-Riesgo_Derivado	\$NC-Riesgo_Derivado	
1	4.000	2.000	2.000	0	0.999	
2	4.000	2.000	2.000	0	0.997	
3	4.000	2.000	2.000	1	0.999	
4	4.000	2.000	1.000	1	0.999	
5	4.000	2.000	2.000	0	1.000	
6	4.000	2.000	1.000	0	0.998	
7	4.000	2.000	1.000	1	0.998	
8	4.000	2.000	0.000	1	0.998	

Figura 4.59: IBM SPSS Modeler, Salida generada por el modelo Red Neuronal Perceptron Backpropagation.

Evaluación con nuevos datos del modelo Árbol de Decisión C&R



Figura 4.60: IBM SPSS Modeler, Evaluación del modelo Árbol de Decisión C&R con nuevos datos.

Tabla (22 campos, 8 registros) #6

Archivo Editar Generar

Tabla Anotaciones

	transformed	other_debtors_transformed	housing_transformed	\$R-Riesgo_Derivado	\$RC-Riesgo_Derivado
1	4.000	2.000	2.000	0	1.000
2	4.000	2.000	2.000	0	1.000
3	4.000	2.000	2.000	1	1.000
4	4.000	2.000	1.000	1	1.000
5	4.000	2.000	2.000	0	1.000
6	4.000	2.000	1.000	0	1.000
7	4.000	2.000	1.000	1	1.000
8	4.000	2.000	0.000	1	1.000

Aceptar

Figura 4.61: IBM SPSS Modeler, Salida generada por el modelo Árbol de Decisión C&R.

Evaluación con nuevos datos del modelo Algoritmo de Regresión Logística



Figura 4.62: IBM SPSS Modeler, Evaluación del modelo Algoritmo de Regresión Logística.

Tabla (22 campos, 8 registros) #7

Archivo Editar Generar

Tabla Anotaciones

	s_transformed	other_debtors_transformed	housing_transformed	\$L-Riesgo_Derivado	\$LP-Riesgo_Derivado
1	4.000	2.000	2.000	0	1.000
2	4.000	2.000	2.000	0	1.000
3	4.000	2.000	2.000	1	1.000
4	4.000	2.000	1.000	1	1.000
5	4.000	2.000	2.000	0	1.000
6	4.000	2.000	1.000	0	1.000
7	4.000	2.000	1.000	1	1.000
8	4.000	2.000	0.000	1	1.000

Aceptar

Figura 4.63: IBM SPSS Modeler, Salida generada por el modelo Algoritmo de Regresión Logística.

Evaluación con nuevos datos del modelo Algoritmo KNN



Figura 4.64: IBM SPSS Modeler, Evaluación del modelo Algoritmo KNN con nuevos datos.

Tabla (22 campos, 8 registros) #8

Archivo Editar Generar

Tabla Anotaciones

	formed	other_debtors_transformed	housing_transformed	\$KNN-Riesgo_Derivado	\$KNNP-Riesgo_Derivado
1	4.000	2.000	2.000	0	0.857
2	4.000	2.000	2.000	0	0.857
3	4.000	2.000	2.000	1	0.857
4	4.000	2.000	1.000	1	0.857
5	4.000	2.000	2.000	0	0.857
6	4.000	2.000	1.000	0	0.857
7	4.000	2.000	1.000	1	0.714
8	4.000	2.000	0.000	1	0.857

Aceptar

Figura 4.65: IBM SPSS Modeler, Salida generada por el modelo Algoritmo KNN.

Evaluación con nuevos datos del modelo Algoritmo de Regresión Lineal



Figura 4.66: IBM SPSS Modeler, Evaluación del modelo Algoritmo de Regresión Lineal con nuevos datos.

Tabla (21 campos, 8 registros) #1

Archivo Editar Generar

Tabla Anotaciones

	transformed	savings_transformed	other_debtors_transformed	housing_transformed	\$E-Riesgo_Derivado
1	8.000	4.000	2.000	2.000	0.076
2	9.000	4.000	2.000	2.000	0.370
3	7.000	4.000	2.000	2.000	0.725
4	7.000	4.000	2.000	1.000	0.862
5	7.000	4.000	2.000	2.000	-0.021
6	7.000	4.000	2.000	1.000	0.014
7	7.000	4.000	2.000	1.000	0.794
8	6.000	4.000	2.000	0.000	0.770

Aceptar

Figura 4.67: IBM SPSS Modeler, Salida generada por el modelo Algoritmo de Regresión Lineal.

Observaciones de la evaluación con nuevos datos de los modelos

Como se puede observar en las Figuras 4.59, 4.61, 4.63, 4.65 y 4.67 todos los modelos pronosticaron perfectamente la salida esperada, pero los resultados de la Figura 4.65 aunque están correctos, hay que recordar que este modelo tiene una pequeña tasa de error, por lo que no es recomendable confiar ciegamente en este modelo.

Por otra parte, el modelo que no se apega al tipo de dato que manejamos en el campo destino es el modelo de Algoritmo de Regresión Lineal de la Figura 4.67, llegando a la conclusión que para este tipo de problema las mejores opciones de modelos fueron: Red Neuronal Perceptron Backpropagation, Árbol de Decisión C&R y el Algoritmo de Regresión Logística.

Capítulo 5

Herramienta Intelligent Data Analysis Tool

Esta herramienta fue diseñada con el propósito de gestionar un dataset, haciendo la limpieza de los datos y mostrar gráficas que nos ayuden a comprender de manera más clara los datos. Además de implementar algún modelo de predicción según los datos ofrecidos por el dataset seleccionado.

La herramienta está desarrollada en el lenguaje de Python y cuenta con una interfaz de introducción (Figura 5.1) en la cual al presionar el botón “Go to main” se podrá acceder a la interfaz principal.

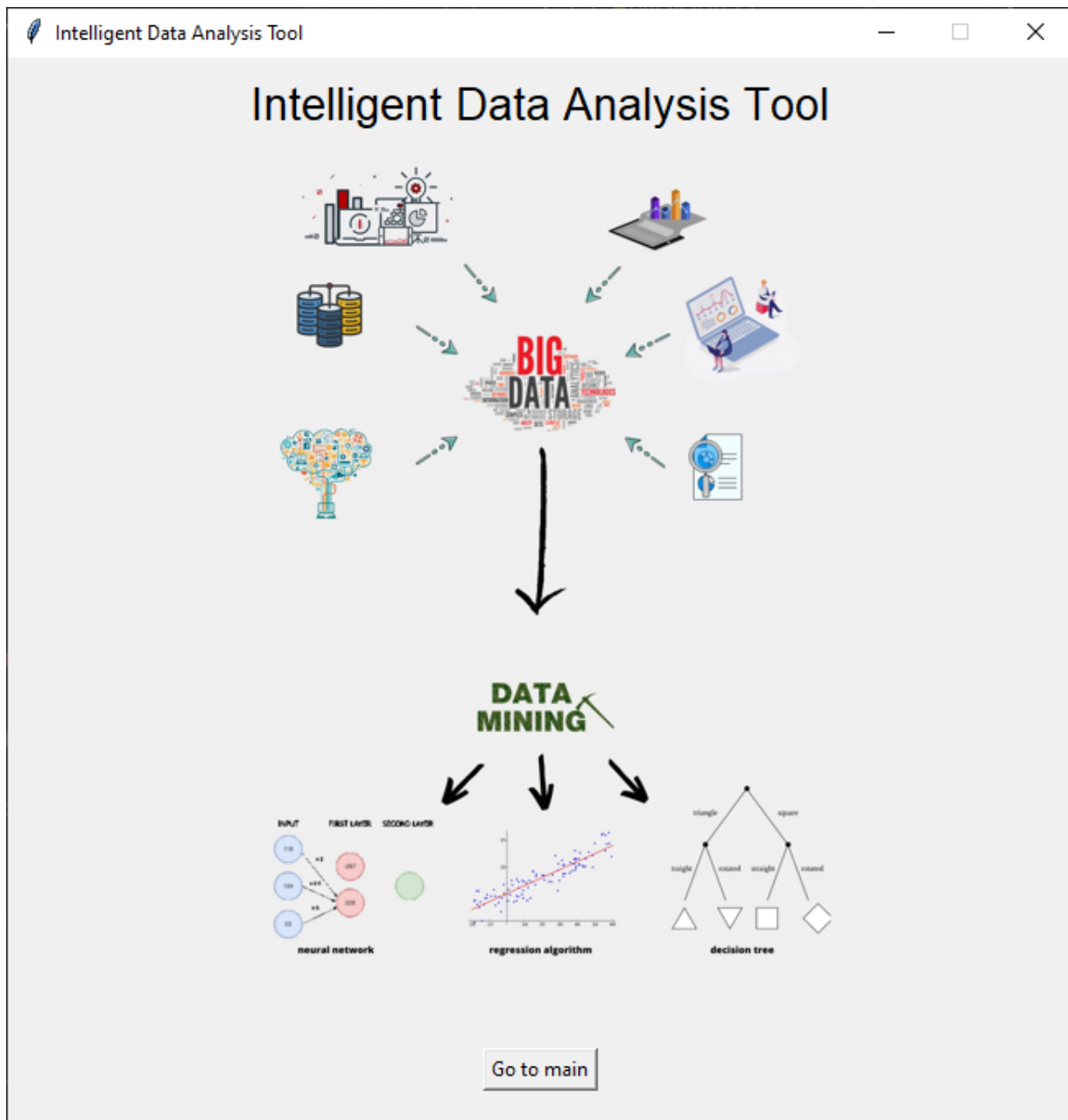


Figura 5.1: Interfaz de introducción.

La interfaz principal (Figura 5.2) consta de un visualizador de datos y un menú de pestañas el cual nos proporciona distintas funcionalidades.

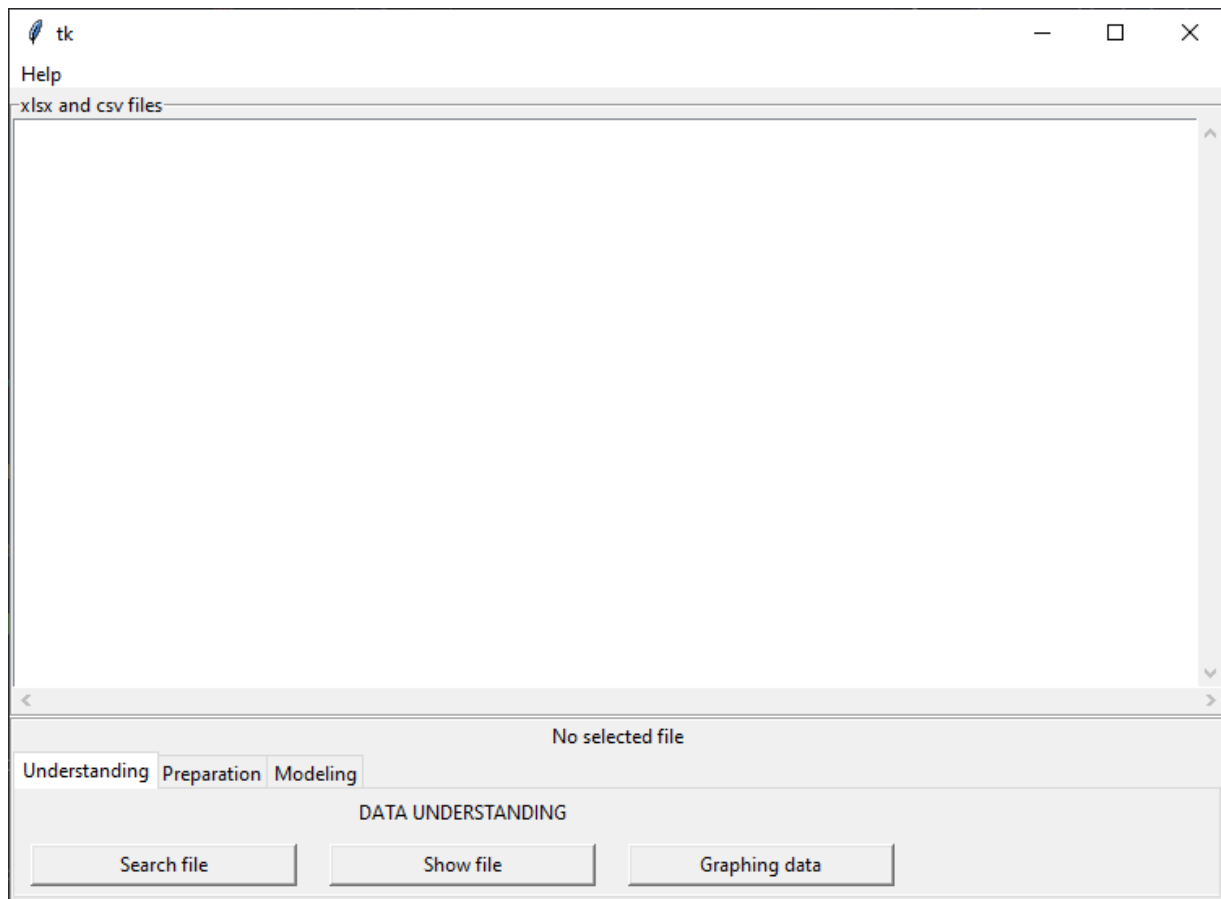
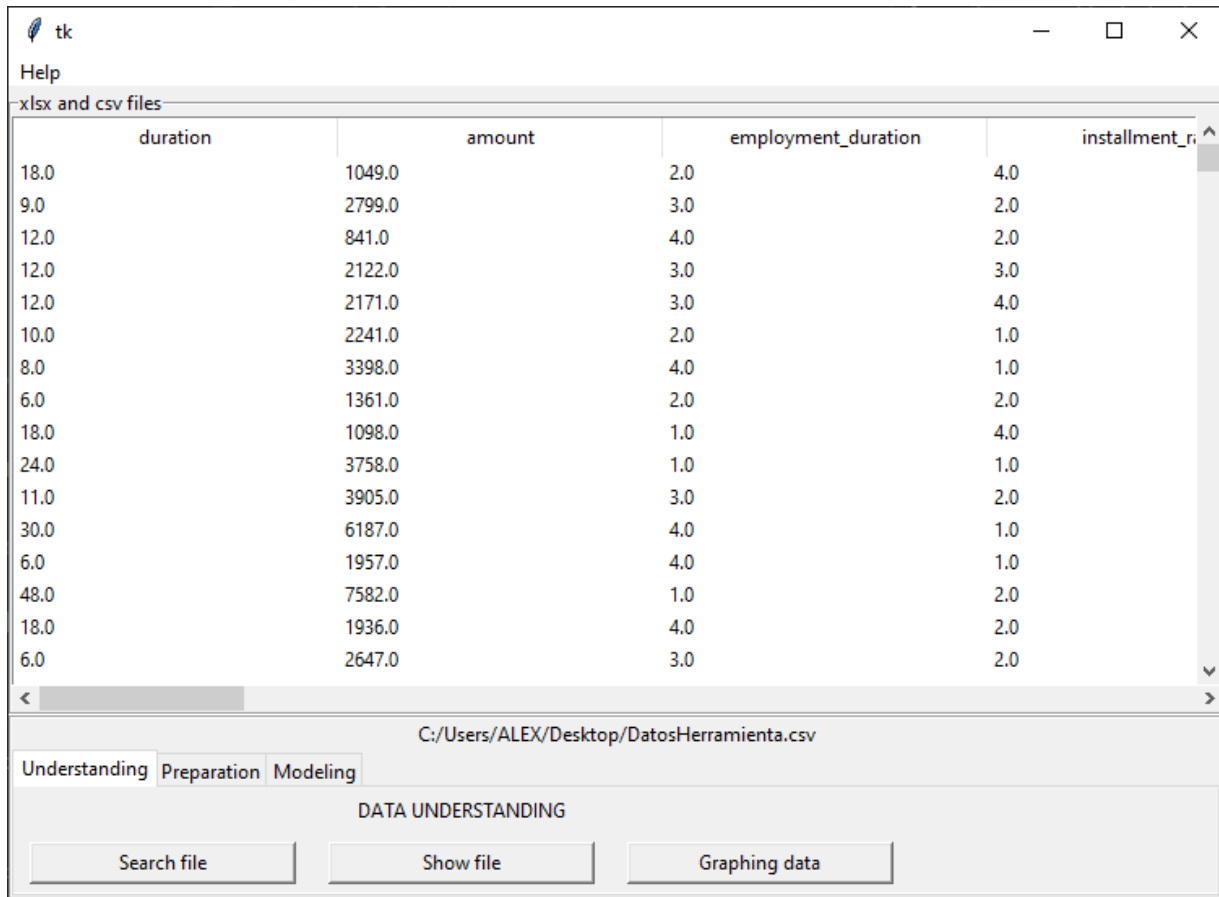


Figura 5.2: Interfaz principal.

La pestaña “Understanding” cuenta con tres botones, el botón “Search File” es para buscar el dataset con el que se desea trabajar, una vez seleccionado se deberá presionar el botón de “Show File” para mostrar los datos. Como se puede observar en la Figura 5.3.



duration	amount	employment_duration	installment_r
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

C:/Users/ALEX/Desktop/DatosHerramienta.csv

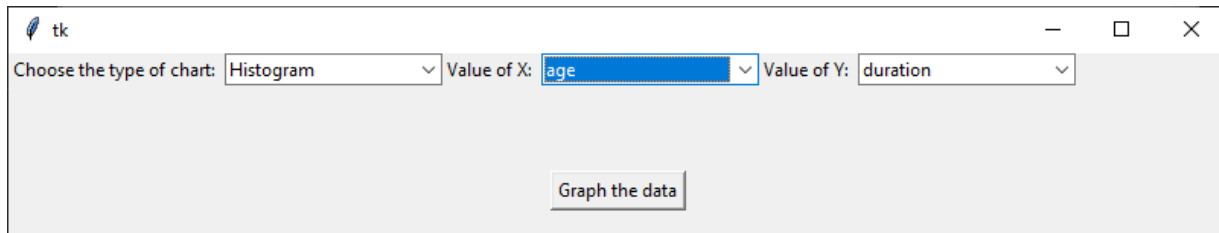
Understanding Preparation Modeling

DATA UNDERSTANDING

Search file Show file Graphing data

Figura 5.3: Visualización de los datos en la herramienta.

El botón de “Graphing data” despliega una ventana emergente la cual cuenta con distintos tipos de gráficos así como el ajuste del eje X y Y del gráfico. Un ejemplo es la Figura 5.4.



tk

Choose the type of chart: Histogram Value of X: age Value of Y: duration

Graph the data

Figura 5.4: Interfaz para graficar los datos.

Al presionar el botón “Graph the data” se mostrará el gráfico seleccionado. La Figura 5.5 es un ejemplo de cómo se muestran los gráficos.

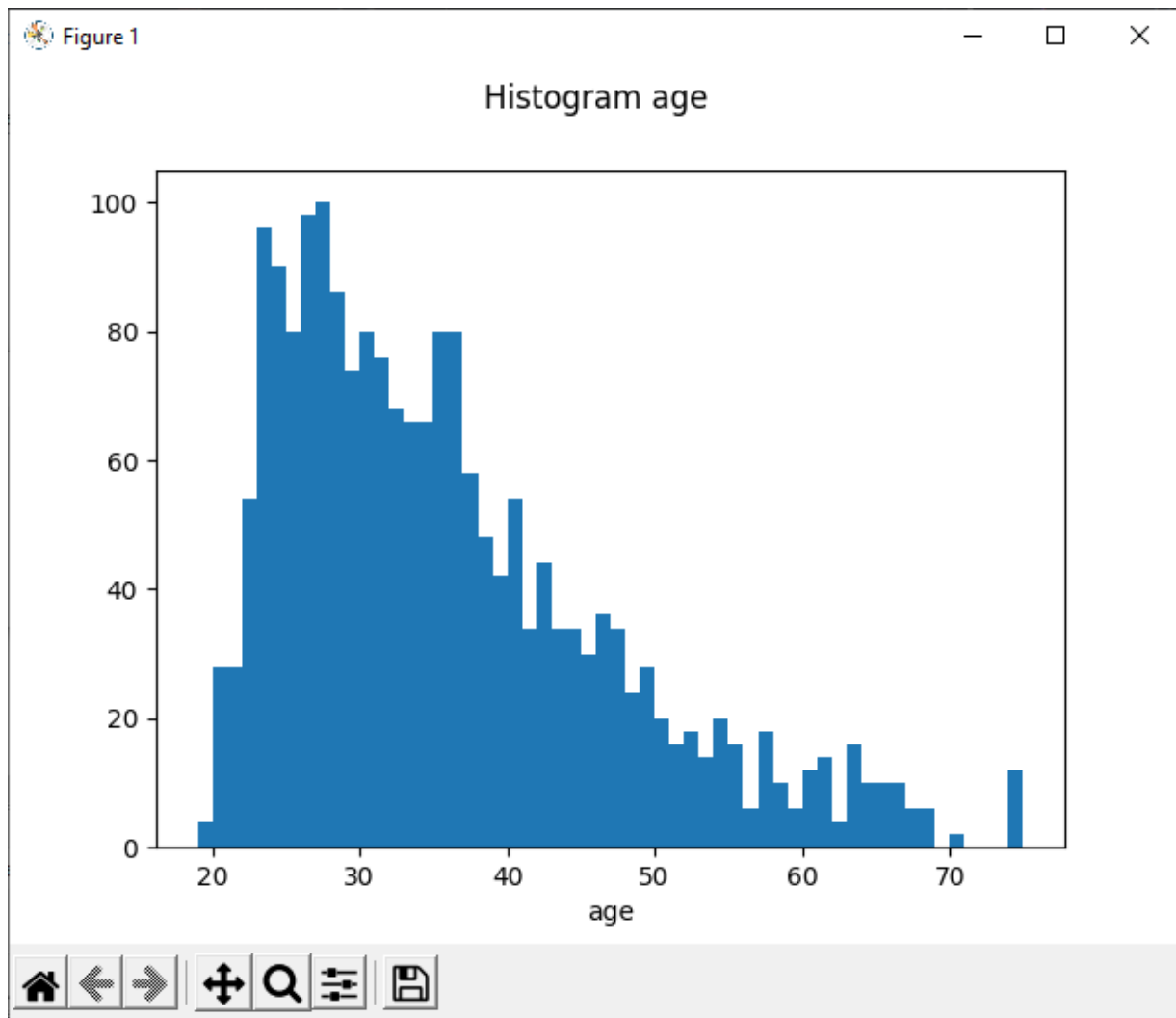


Figura 5.5: Gráfico.

Continuando con la interfaz principal de la aplicación, la pestaña “Preparation” cuenta con 4 botones los cuales son opciones de limpieza de los datos. Ver 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.
- Botón All: esta opción de limpieza combina las tres técnicas anteriores.

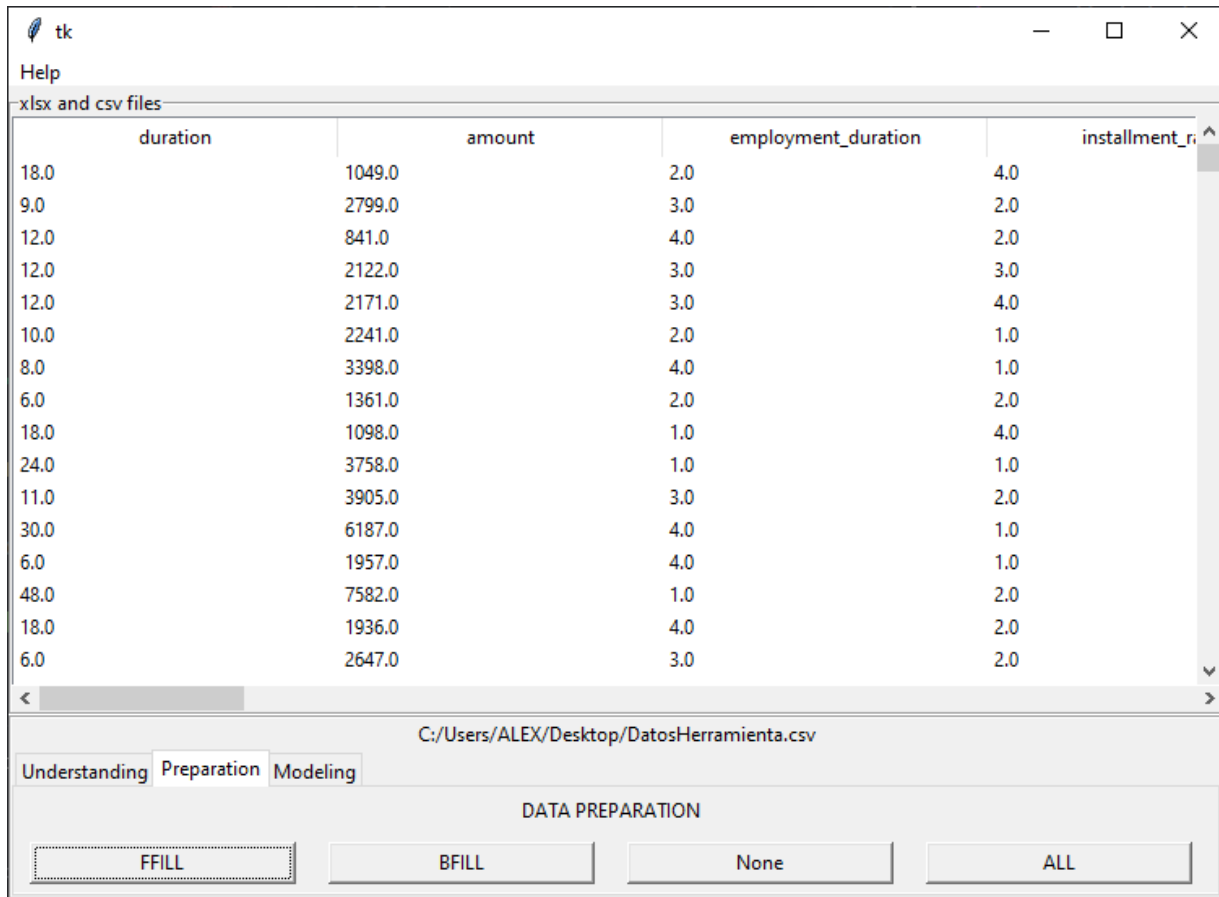


Figura 5.6: Opciones de limpieza de los datos.

La pestaña “Modeling” nos permite hacer la implementación de algún Modelo de predicción como Red Neuronal Artificial, Regresión Lineal, Árbol de Decisión o algoritmo KNN (hasta este punto sólo se tiene parcialmente implementado el modelo de Red Neuronal, permitiéndonos predecir únicamente campos binarios). La Figura 5.7 muestra la interfaz con la selección de la pestaña “Modeling”.

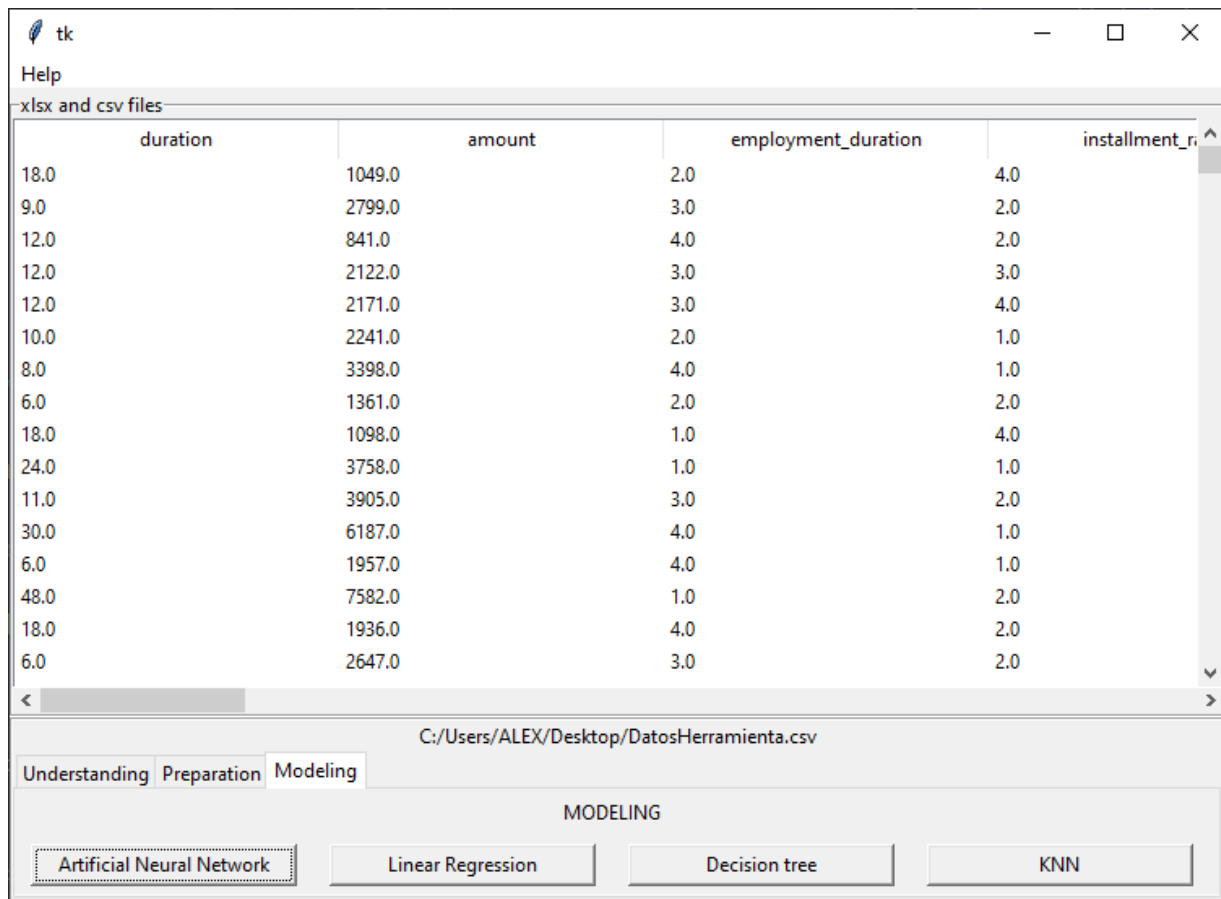


Figura 5.7: Opciones de técnicas de modelado.

Al seleccionar el modelo “Artificial Neural Network” se abrirá una ventana emergente que cuenta con dos secciones, en la primera sección “INPUT” se deberán seleccionar todos los campos de entrada para el modelo, en la sección “TARGET” se seleccionará el campo destino (Figura 5.8).

The screenshot shows a window titled "tk" with two main sections: "INPUT" and "TARGET". Each section contains a list of variables with checkboxes. In the "INPUT" section, all variables except "Riesgo_Derivado" are checked. In the "TARGET" section, only "Riesgo_Derivado" is checked. A "Build model" button is located at the bottom center.

INPUT	TARGET
<input checked="" type="checkbox"/> duration	<input type="checkbox"/> duration
<input checked="" type="checkbox"/> amount	<input type="checkbox"/> amount
<input checked="" type="checkbox"/> employment_duration	<input type="checkbox"/> employment_duration
<input checked="" type="checkbox"/> installment_rate	<input type="checkbox"/> installment_rate
<input checked="" type="checkbox"/> present_residence	<input type="checkbox"/> present_residence
<input checked="" type="checkbox"/> property	<input type="checkbox"/> property
<input checked="" type="checkbox"/> age	<input type="checkbox"/> age
<input checked="" type="checkbox"/> number_credits	<input type="checkbox"/> number_credits
<input checked="" type="checkbox"/> job	<input type="checkbox"/> job
<input checked="" type="checkbox"/> personal_status_sex_transformed	<input type="checkbox"/> personal_status_sex_transformed
<input checked="" type="checkbox"/> other_installment_plans_transformed	<input type="checkbox"/> other_installment_plans_transformed
<input checked="" type="checkbox"/> people_liable_transformed	<input type="checkbox"/> people_liable_transformed
<input checked="" type="checkbox"/> foreignn_worker_transformed	<input type="checkbox"/> foreignn_worker_transformed
<input checked="" type="checkbox"/> credit_risk_transformed	<input type="checkbox"/> credit_risk_transformed
<input checked="" type="checkbox"/> status_transformed	<input type="checkbox"/> status_transformed
<input checked="" type="checkbox"/> credit_history_transformed	<input type="checkbox"/> credit_history_transformed
<input checked="" type="checkbox"/> purpose_transformed	<input type="checkbox"/> purpose_transformed
<input checked="" type="checkbox"/> savings_transformed	<input type="checkbox"/> savings_transformed
<input checked="" type="checkbox"/> other_debtors_transformed	<input type="checkbox"/> other_debtors_transformed
<input checked="" type="checkbox"/> housing_transformed	<input type="checkbox"/> housing_transformed
<input type="checkbox"/> Riesgo_Derivado	<input checked="" type="checkbox"/> Riesgo_Derivado

Build model

Figura 5.8: Selección de campos de entrada y salida del modelo.

Al presionar el botón “Build Model” nuevamente se abrirá una ventana emergente (Figura 5.9) en la cual se deberá seleccionar los datos de prueba para el modelo.

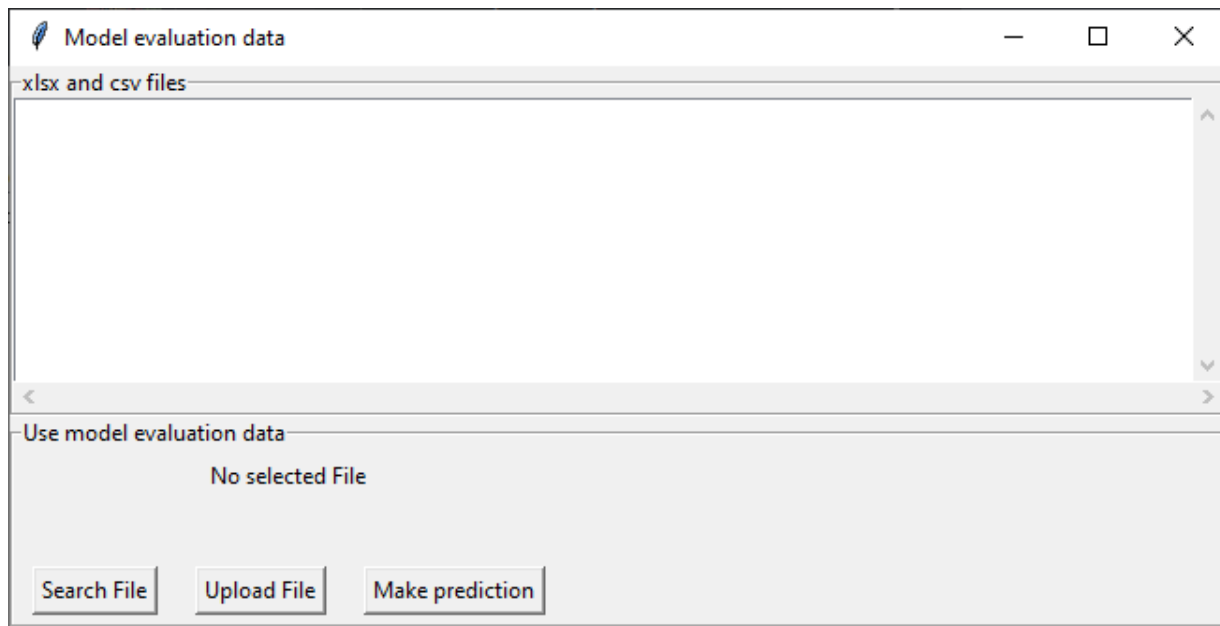


Figura 5.9: Selección de los datos de prueba para el modelo.

Una vez cargados (Figura 5.10) para comenzar con la predicción se deberá presionar el botón “Make prediction”.

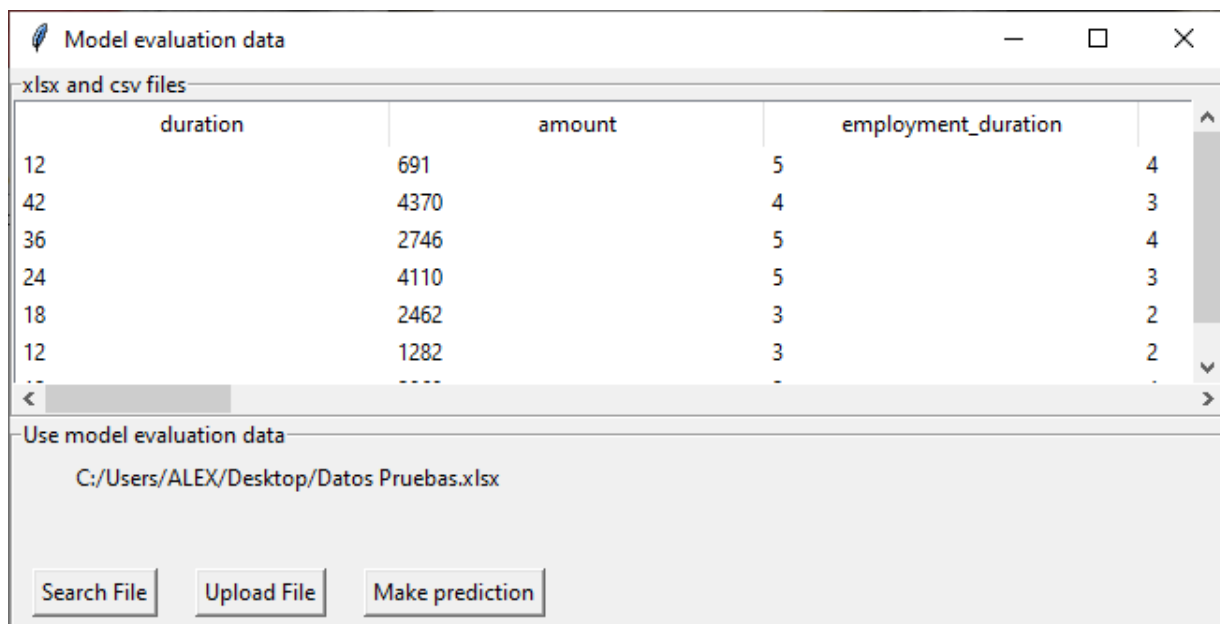


Figura 5.10: Comenzar con la predicción.

Terminada la predicción, se mostrará el gráfico de evolución en la precisión del modelo. La Figura 5.11 es un ejemplo de como el modelo evoluciona según las generaciones que pasan.

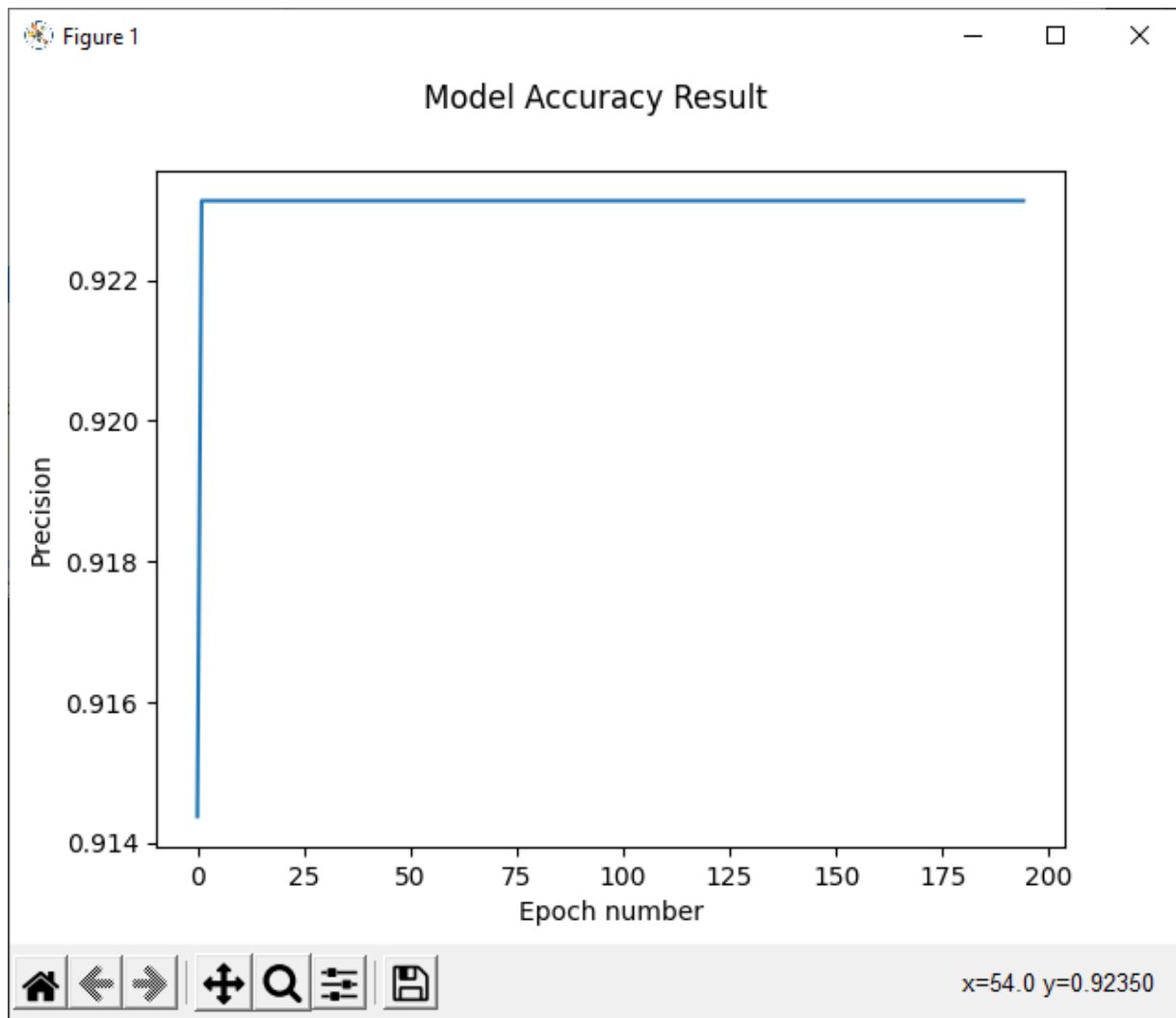


Figura 5.11: Gráfico de la evolución en la precisión.

También se mostrará la predicción y la tasa de efectividad del modelo. La Figura 5.12 muestra la predicción generada por el modelo y su tasa de efectividad.

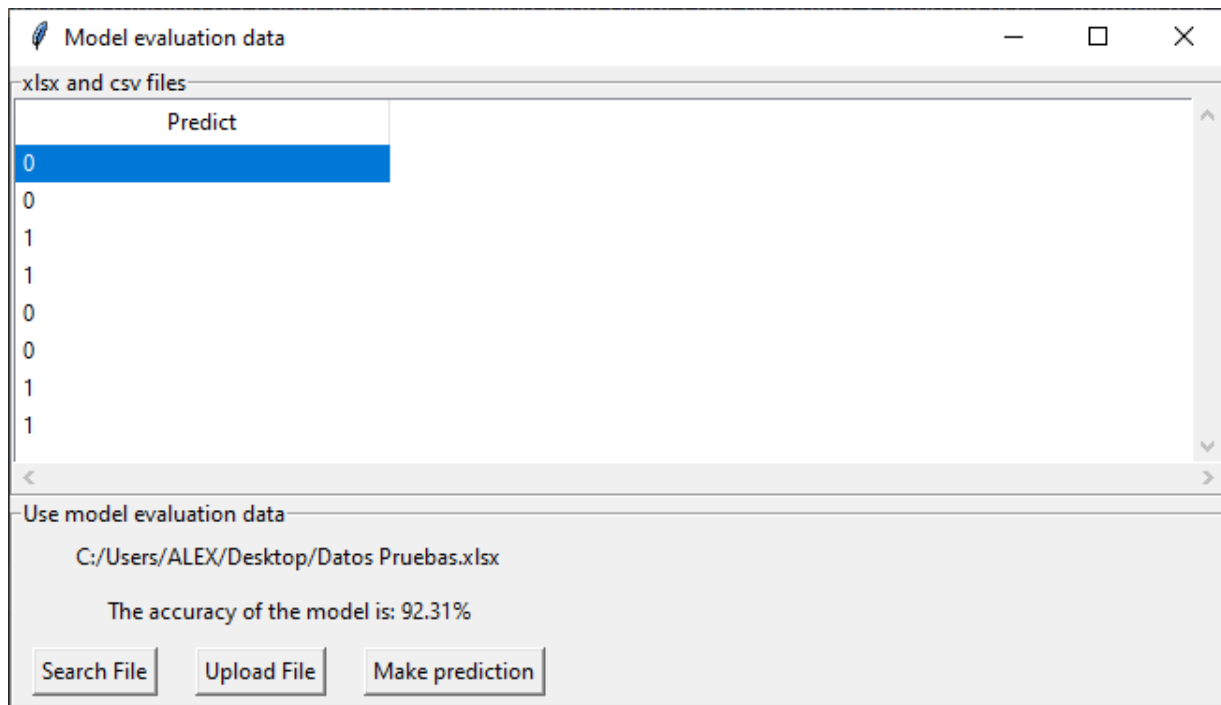


Figura 5.12: Predicción y tasa de efectividad del modelo.

Como se puede observar en la Figura 5.12 la salida es la misma que en la Figura 4.59, de esta manera comprobamos que nuestro modelo está funcionando a la perfección.

El proyecto se encuentra en un repositorio de GitHub abierto para todo el público interesado, el link al proyecto: <https://github.com/Diego-Cansino/Intelligent-Data-Analysis-Tool>

Conclusiones

En el Capítulo 1 de este proyecto se intentó abordar todo lo relacionado sobre los datos a gran escala, partiendo de una definición propia derivada de la investigación y documentación sobre qué es el big data, también se mencionaron y explicaron las siete V's principales de los Datos a Gran Escala que son fundamentales y características en los datos masivos, seguido de una explicación sobre los tipos de datos que existen dentro del big data así como su procedencia y cómo estos pueden ser clasificados ya sea en datos estructurados, semi-estructurados o no estructurados. También se explica cómo estos datos pueden ser procesados para que tengan un valor significativo haciendo uso de un ciclo de vida que se basa en distintas etapas. Se planteó un ciclo de vida genérico con el fin de conocer las etapas y principales actividades a realizar en un proyecto de minería de datos, no obstante existen metodologías especializadas para el procesamiento de los datos, las metodologías dominantes en este campo son: KDD, CRISP-DM, SMART y SEMMA.

En el Capítulo 2 se hace una explicación sobre las técnicas de modelado comúnmente utilizadas en la Ciencia de los Datos donde se plantean sus aspectos técnicos y sus campos de aplicación de las Redes Neuronales Artificiales, Algoritmos de Regresión Lineal, Algoritmos de Regresión Logística, Algoritmo KNN y Árboles de Decisión. Actualmente las Redes Neuronales Artificiales son la técnica de moldeado más utilizadas en proyectos de minería de datos gracias a los múltiples beneficios que otorgan, sin embargo es una de las técnicas más difíciles de comprender y dominar.

El Capítulo 3 es una introducción al área de estudio del proyecto, en este capítulo se da la definición de las Finanzas y se explican los cuatro tipos de Finanzas que existen, seguido de algunos casos de uso de los datos a gran escala aplicados a las finanzas corporativas y personales, con el fin de comprender el alcance de estos datos masivos.

El Capítulo 4 es la parte fundamental del proyecto, ya que en este se hace la aplicación de la metodología CRISP-DM para evaluar la solvencia de los clientes que solicitan algún crédito, con el fin de obtener un indicador de riesgo que nos permita saber si el cliente es o no apto para recibir el crédito. A lo largo de este capítulo se van siguiendo las distintas etapas de la metodología así como cada una de sus actividades principales, lo que le da riqueza al proyecto.

Los resultados obtenidos por los modelos aplicados fueron excelentes, obteniendo un mínimo de 90% de efectividad por modelo. Un aspecto que ayudó a este porcentaje fue que los datos estaban muy bien clasificados, no tenían errores de datos nulos o datos

fuera de rango evitando cálculos erróneos en los modelos; otro factor que ayudó fue que se consiguió la continuación del dataset asegurando así la calidad y el volumen de los datos.

En el Capítulo 5 se presentó una herramienta desarrollada en Python con el fin de ayudarnos en este proceso de comprender las etapas de la metodología CRISP-DM así como las técnicas de modelado de los datos, reforzando los conceptos básicos de un proyecto de minería de datos.

Aún hay mucho que desarrollar en esta aplicación, especialmente en las técnicas de modelado, ya que al momento sólo se puede hacer uso del modelo de red neuronal para predecir datos binarios, falta complementar esta función para que pueda predecir datos continuos, a su vez hay que desarrollar otros modelos como el de regresión lineal, árboles de decisión, algoritmo KNN, entre otros. Otro campo de mejora es en el diseño de las interfaces, estas podrían ser más intuitivas y más amigables al usuario.

En general, de este proyecto me llevo grandes aprendizajes empezando por conocer más a fondo el mundo del big data e incluso atreviéndome a dar una definición propia. También me llevo la experiencia de saber cómo llevar acabo un proyecto de minería de datos haciendo uso de la metodología CRISP-DM y de distintas herramientas, pero en especial la herramienta IBM SPSS Modeler, la cual me fué de mucha ayuda a lo largo del Capítulo 4.

Aún queda un largo camino por recorrer y descubrir, pero este es un gran avance que me permite seguir desarrollándome como Ingeniero en Computación.

Bibliografía

- [1] “El modelo de redes neuronales.” [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model>
- [2] “Notas de la UEA: "Datos a gran escala" impartida en la Universidad Autónoma Metropolitana.”
- [3] “¿Qué es la regresión lineal?” [Online]. Available: <https://la.mathworks.com/discovery/linear-regression.html>
- [4] “Regresión Logística.” [Online]. Available: <https://aprendeia.com/regresion-logistica-multiple-machine-learning-teoria/>
- [5] “K-Nearest Neighbors.” [Online]. Available: <https://aprendeia.com/regresion-logistica-multiple-machine-learning-teoria/>
- [6] Instituto de Ingeniería del Conocimiento., “Las 7 V del Big Data: Características más importantes.” [Online]. Available: <https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/>
- [7] —, “Infografía Big Data: las 7 V.” [Online]. Available: <https://www.iic.uam.es/innovacion/big-data-infografia-7-v/>
- [8] Oracle, “¿Qué es el big data? | Oracle México.” [Online]. Available: <https://www.oracle.com/mx/big-data/what-is-big-data/>
- [9] “Tipos de datos: datos estructurados, semiestructurados y no estructurados. Blog de Tecnología - IMF Smart Education.” [Online]. Available: <https://blogs.imf-formacion.com/blog/tecnologia/tipos-de-datos-datos-estructurados-semiestructurados-y-no-estructurados-202006/>
- [10] Patricio Rodríguez, Norma Palomino, Javier Mondaca , *El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe*.
- [11] “Casos de uso del Big Data.” [Online]. Available: <https://evaluandocloud.com/casos-uso-del-big-data/>
- [12] “Redes Neuronales Artificiales y sus Aplicaciones.” [Online]. Available: <https://docplayer.es/83329-Redes-neuronales-artificiales-y-sus-aplicaciones.html>

- [13] “Notas de la Materia: .Análisis de Decisiones impartida en el Instituto Politécnico Nacional.”
- [14] “Finanzas. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/finanzas.html>
- [15] “¿En qué consisten las finanzas corporativas? Business Class: Trends & Insights | American Express. .” [Online]. Available: <https://www.americanexpress.com/es-mx/negocios/trends-and-insights/articles/consisten-finanzas-corporativas>
- [16] “Finanzas Personales. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/finanzas-personales.html>
- [17] “Finanzas públicas. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/finanzas-publicas.html>
- [18] “Ingreso público. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/ingreso-publico.html>
- [19] “Gasto público. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/gasto-publico.html>
- [20] “Finanzas internacionales. Economipedia.” [Online]. Available: <https://economipedia.com/definiciones/finanzas-internacionales.html>
- [21] “Big Data en las finanzas. Usos, beneficios y casos de éxito.” [Online]. Available: <https://iat.es/tecnologias/big-data/finanzas/>
- [22] “Conceptos básicos de ayuda de CRISP-DM.” [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview>