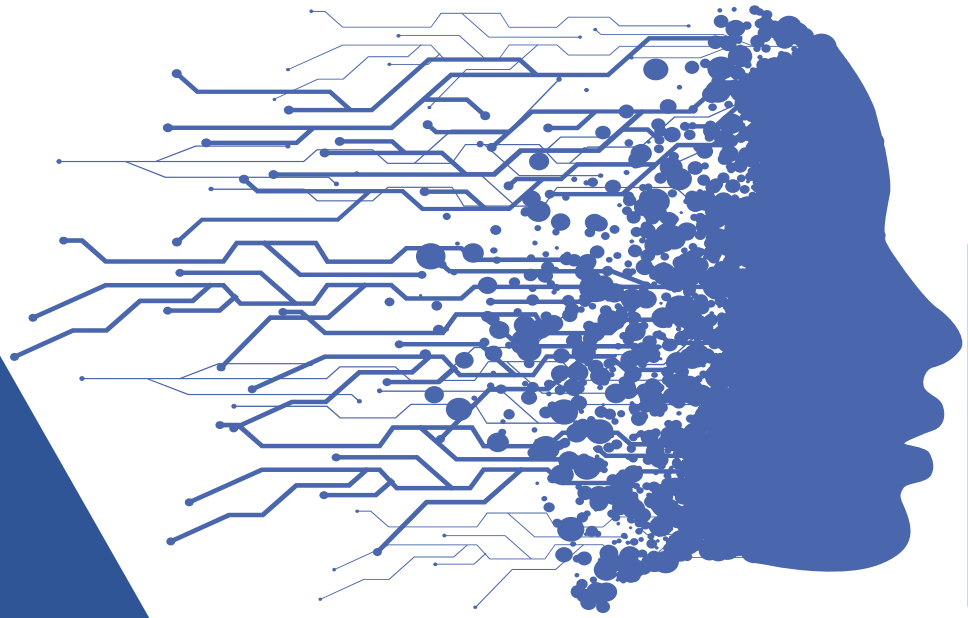
MAESTRÍA EN GESTIÓN EN SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Diseño de modelos de minería de datos para la prospección de compra en seguros de vida



Autor: Ing. José Quintana
Director: PhD. Freddy Tapia

Ecuador

Diciembre 2022

Contenido

01 Introducción

Antecedentes

Empresa de Seguros

Analizar información

Toma decisiones

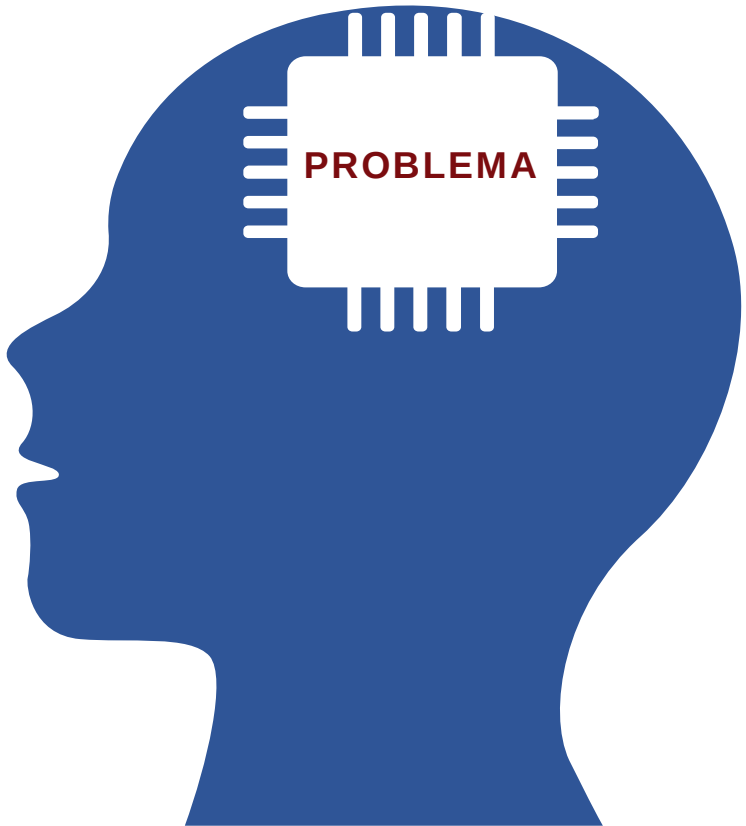
Pérdida de oportunidades

Problema

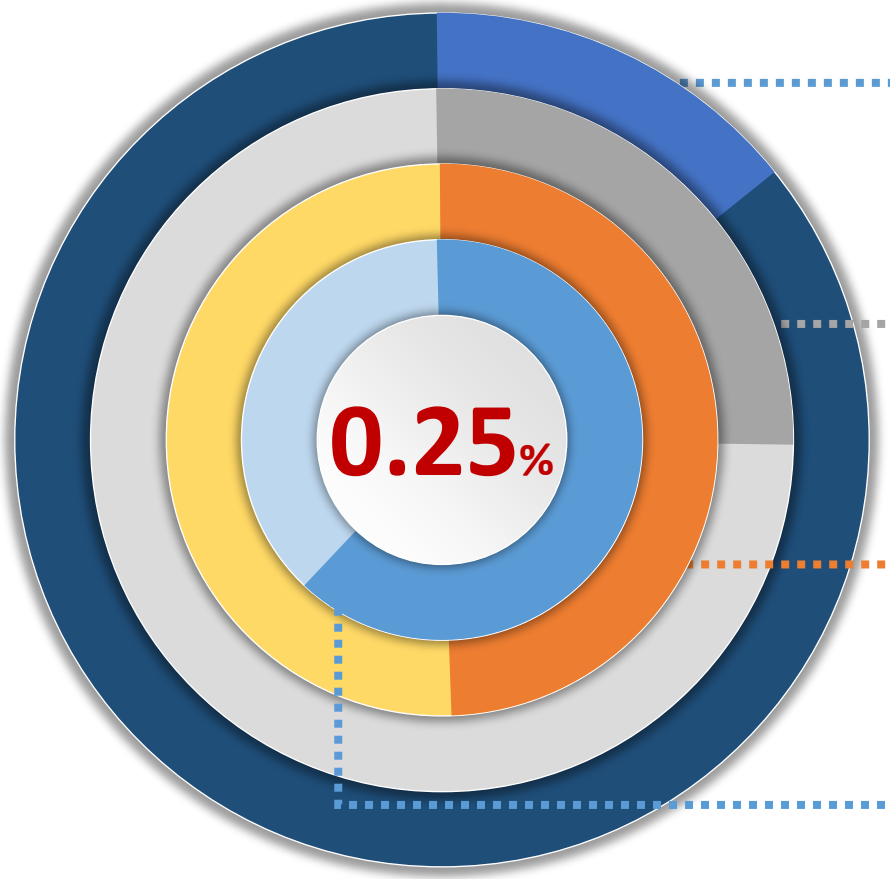
Año 2019 inicia Proceso de Prospección

845 mil prospectos entregados vs 1812 ventas.

Baja efectividad de venta de las campañas, menos del 1%.



Porcentaje de Conversión



Campañas
69

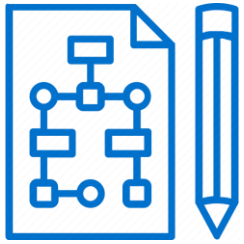
Prospectos entregados
845,660

Prospectos gestionados
736,293

Clientes adquiridos
1,812

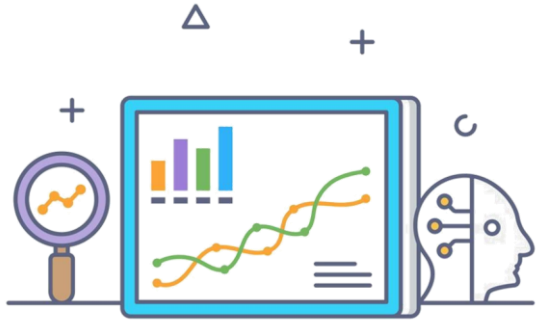
- Desde noviembre 2019.
- 12k x mes entregado.
- Prospectos reciclados cada 3 meses.

Objetivo

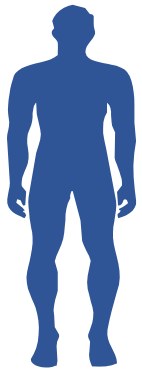


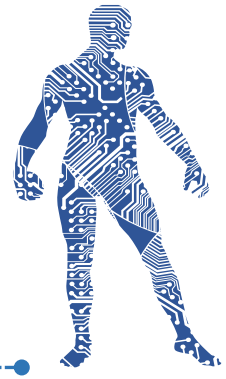
Diseñar Marco de Trabajo

Encontrar uno o varios modelos



Predecir quien comprará mi producto





02- Configuración Experimental y Metodología

Objetivo de la Minería de Datos

Buscar en los datos

Perspectivas analíticas para facilitar los procesos de **TOMA DE DECISIÓN**

Nivel de análisis según la ETAPA de madurez y OBJETIVOS del negocio

Conocer lo que ha pasado

Análisis Descriptivo

Analiza e interpreta lo que ha sucedido

- Qué pasó
- Cómo pasó

Estimar los futuros resultados

Análisis Predictivo

A partir de lo sucedido intenta predecir lo que sucederá

- Por qué pasó
- Qué podría pasar

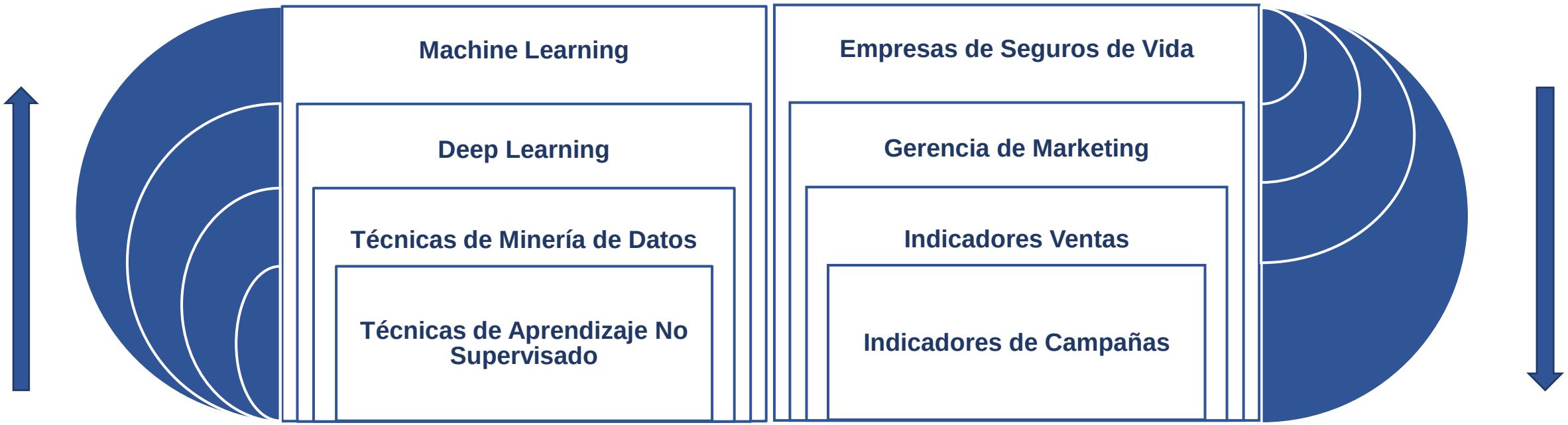
Mejorar los posibles resultados

Análisis Prescriptivo

Identifica estrategias y acciones que Mejoren los resultados previstos

- Qué hacer
- Cómo hacerlo

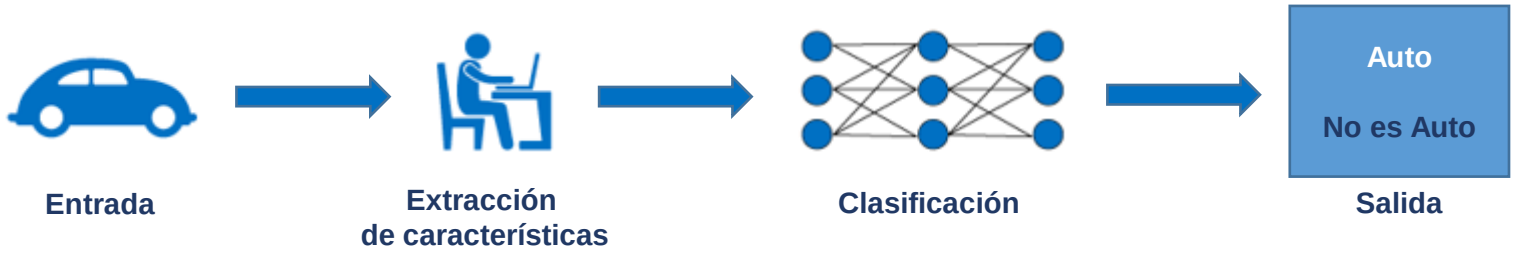
Jerarquía de las Variables del Problema



Variable Independiente ⟷ Variable Dependiente

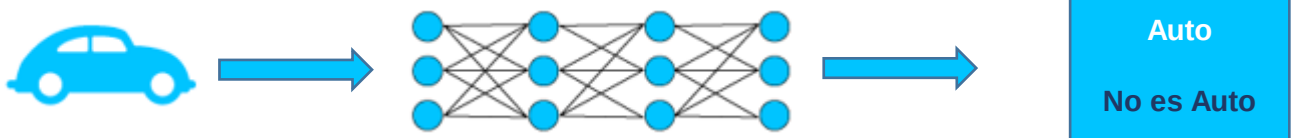
Tipos de Aprendizaje

Machine Learning



Entrada → Extracción de características → Clasificación → Salida

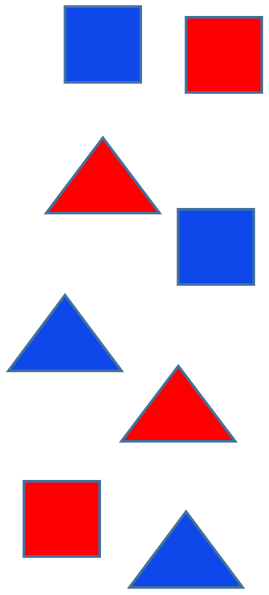
Deep Learning



Entrada → Extracción de características y Clasificación → Salida

Tipos de Machine Learning

Conjunto de datos



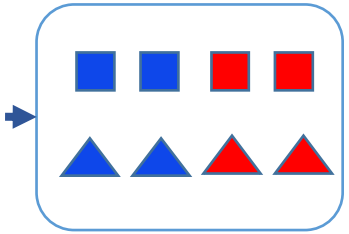
Aprendizaje Supervisados o predictivos (Clasificación, Regresión)

Conjunto de datos

Etiquetas

Aprendizaje No Supervisados o del descubrimiento del conocimiento (Cluster, K-Means)

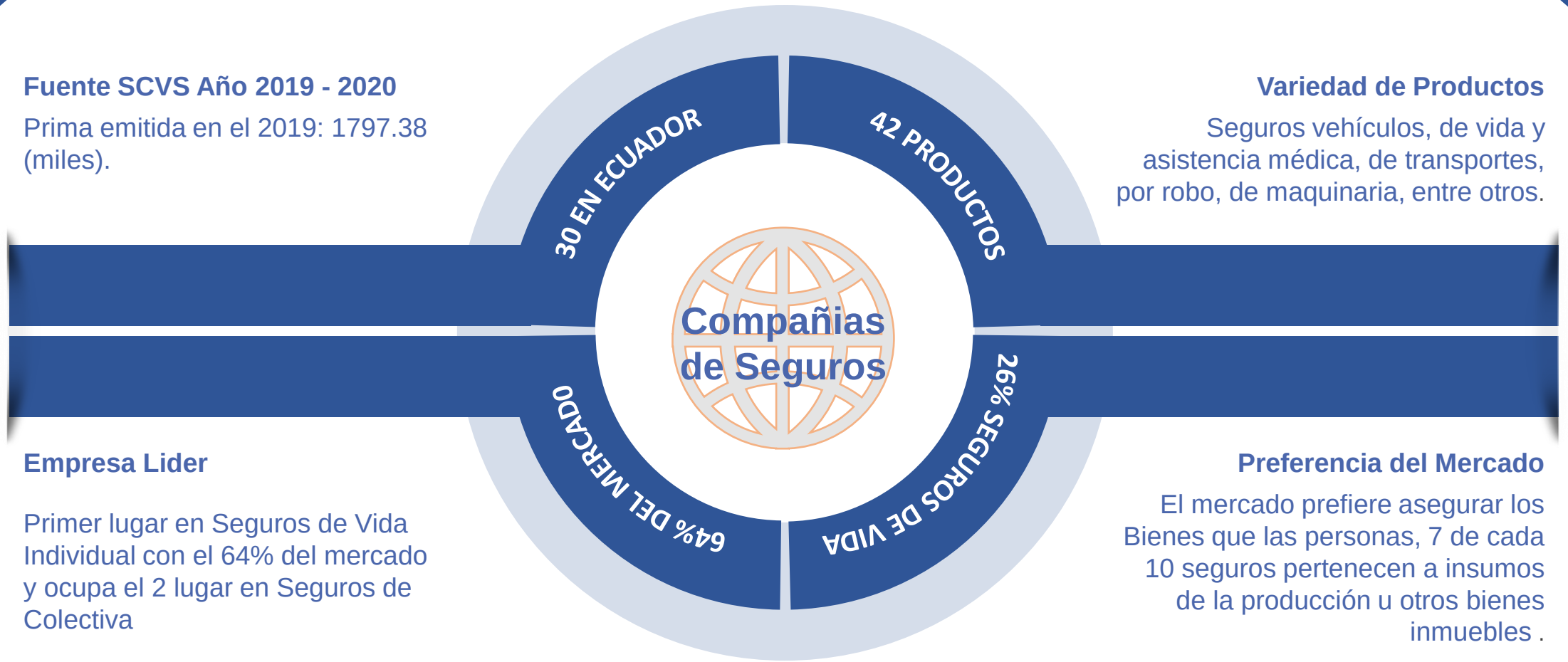
Conjunto de datos



Presencia en el Mercado Asegurador

Fuente SCVS Año 2019 - 2020

Prima emitida en el 2019: 1797.38 (miles).



Variedad de Productos

Seguros vehículos, de vida y asistencia médica, de transportes, por robo, de maquinaria, entre otros.

Empresa Lider

Primer lugar en Seguros de Vida Individual con el 64% del mercado y ocupa el 2 lugar en Seguros de Colectiva

Preferencia del Mercado

El mercado prefiere asegurar los Bienes que las personas, 7 de cada 10 seguros pertenecen a insumos de la producción u otros bienes inmuebles .

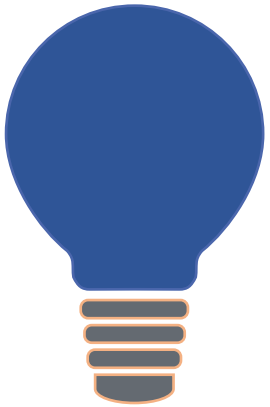
Campañas de Ventas

Gerencia de Marketing

La gerencia de Marketing se encarga de la relación directa con el cliente

Indicadores de Ventas

La fuerza de ventas se enfoca en la venta y su rendimiento es evaluado con el cumplimiento de su presupuesto mensual.

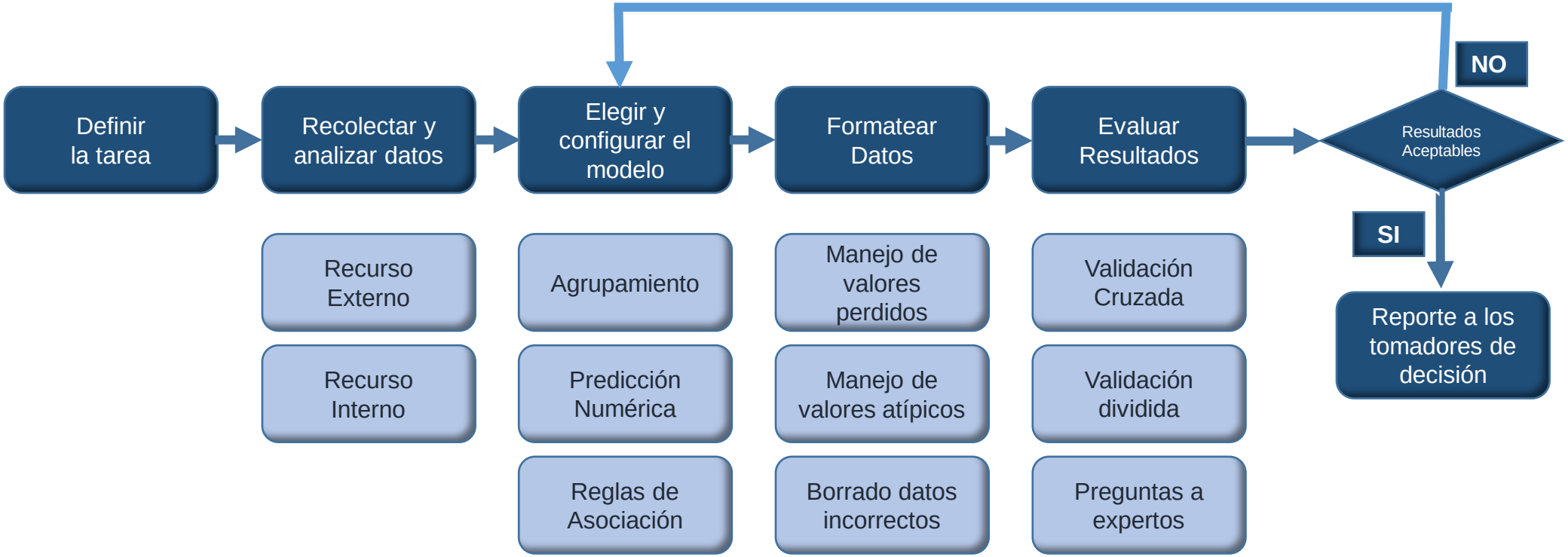


Indicadores de Campañas

Se compara la venta real vs prospectos entregados

SME

Pequeñas y Medianas Empresas (PYME), brinda las pautas para poder realizar trabajos de análisis de datos de manera estructurada y metodológica en seis pasos



Steps of SME (Dittert M., Härting RC., Reichstein C., Bayer C., 2018)

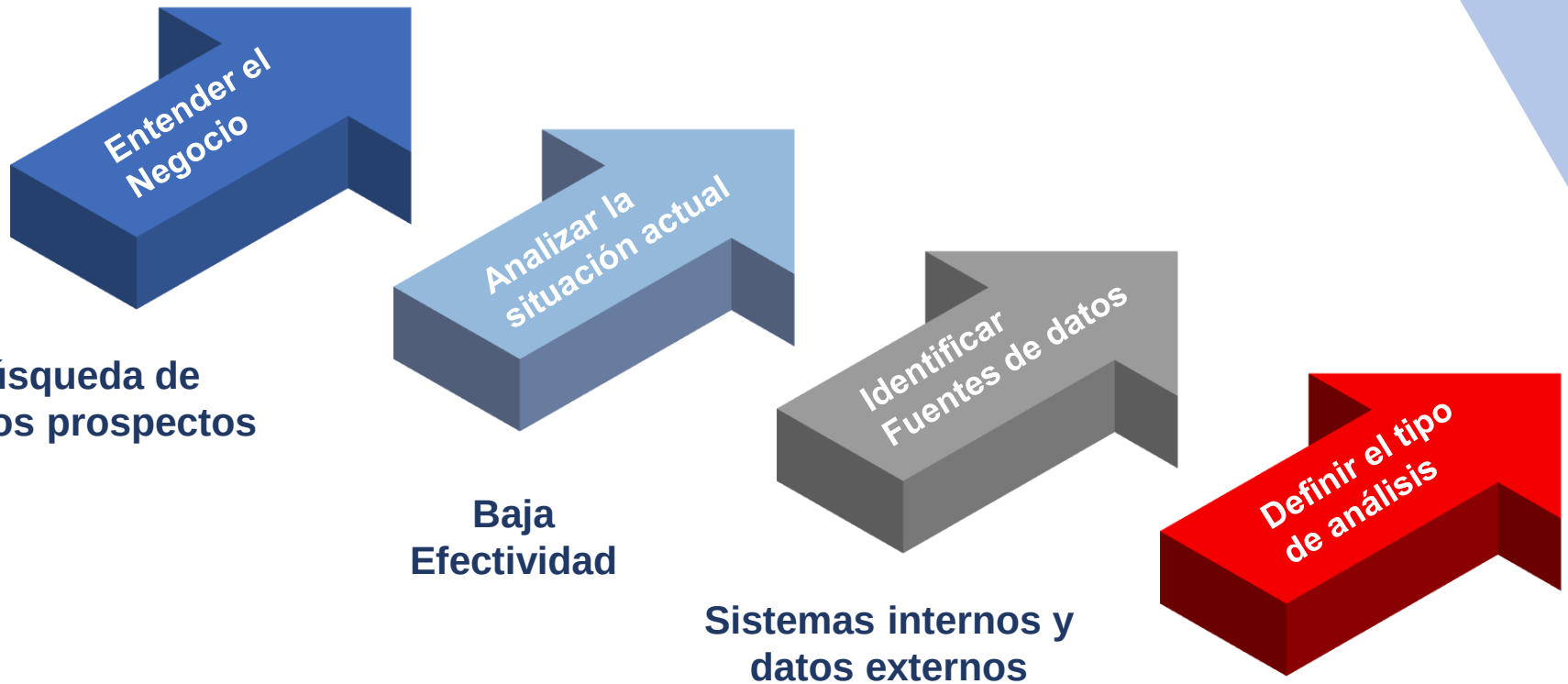
Herramientas de Data Mining

- Permiten interactuar con datos.
- Respaldan el ciclo completo(crear, implementar y administrar modelos de análisis).
- Se uso Alteryx y R

	Alteryx	Anaconda	Databricks	Datarobot	Knime	Azure ML	Rapid Miner
Capacidades del producto							
Exploración y visualización de datos	⚠	✓	⚠	✓	⚠	⚠	⚠
Rendimiento y escalabilidad	✓	⚠	✓	✓	✓	✓	✓
Acceso a los datos	✓	⚠	✓	✓	✓	✓	✓
Preparación de datos	✓	✗	✓	⚠	✓	⚠	✓
Automatización	✓	⚠	⚠	✓	✓	✓	✓
Interfaz de usuario	✓	⚠	⚠	✓	✓	✓	✓
Aprendizaje automático	✓	✓	✓	✓	✓	✓	✓
Otras analíticas avanzadas	✓	⚠	✓	⚠	✓	⚠	⚠
Flexibilidad y apertura	✓	✓	⚠	⚠	✓	⚠	✓
Despliegue	✓	⚠	✓	✓	✓	✗	✓
Gestión de modelos	⚠	⚠	✗	✓	✓	⚠	✓
Soluciones preenvasadas	⚠	✗	✗	✓	⚠	⚠	✓
Colaboración	⚠	✗	✓	✗	⚠	⚠	✓
Requiere personal especializado	⚠	✗	✗	✓	⚠	✓	⚠
Experiencia del cliente							
Precios	✓	⚠	⚠	✗	✗	⚠	✗
Facilidad de implementación	✓	✓	✓	✓	✓	✓	✓
Oportunidad de la respuesta del proveedor	✓	✓	✓	✓	✓	✓	✓
Calidad del soporte técnico	✓	✓	✓	✓	✓	✓	✓

03-Análisis de Datos

Definir tareas



Búsqueda de nuevos prospectos

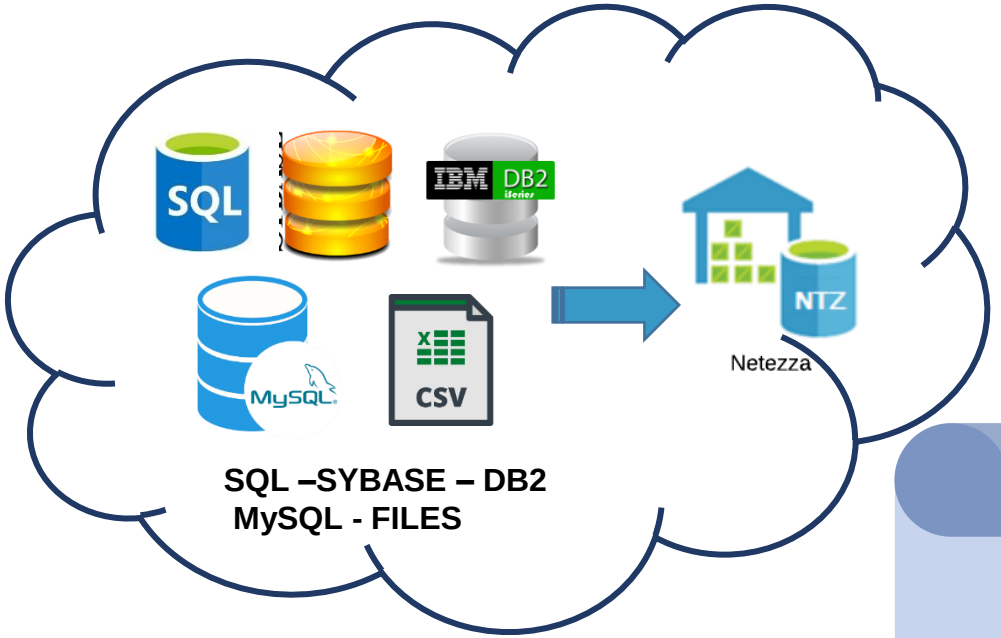
Baja Efectividad

Sistemas internos y datos externos

Predicción de compra

Obtener y Analizar los datos

Fuentes



840k

Total de registros

280

Variables encontradas

91

Variables analizadas,
35 de 90 al 100% sin datos existentes

37

Variables usadas
Excluyen repetidas y únicos.

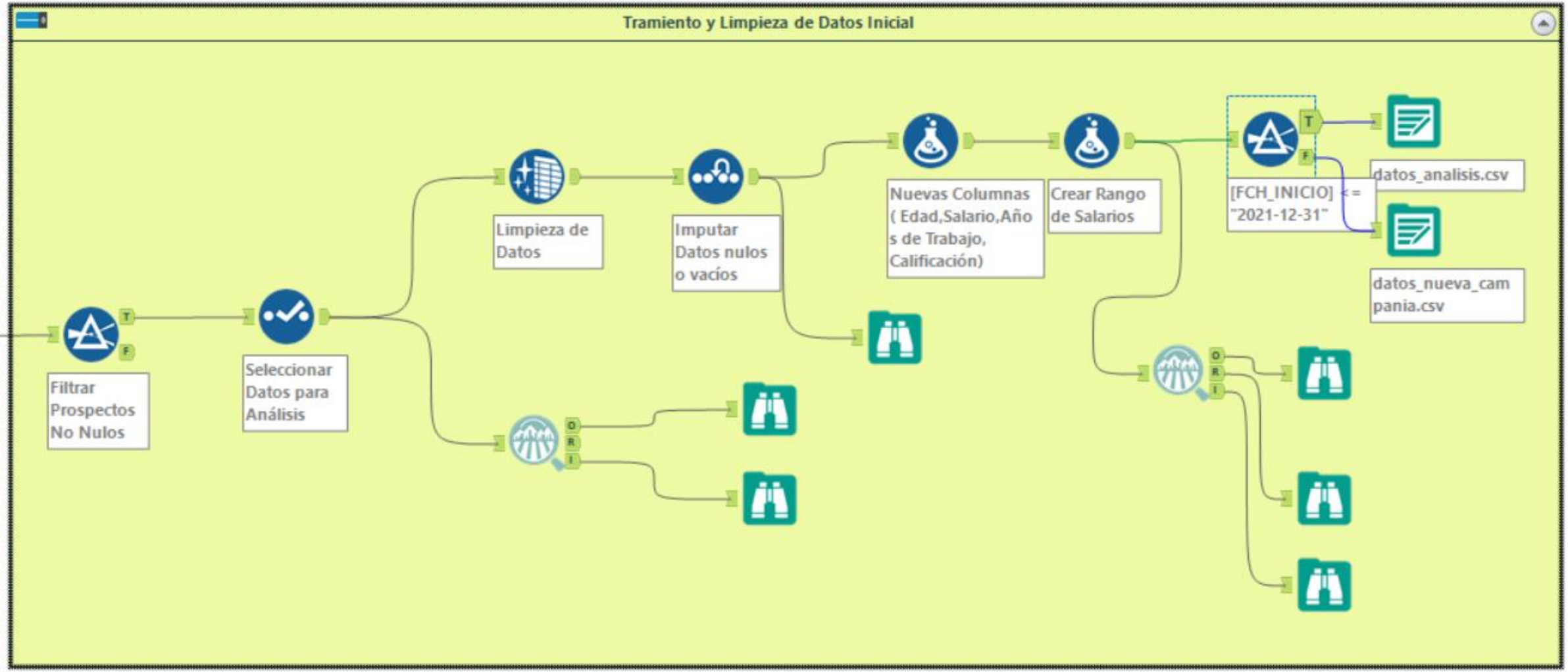
591k entrenar modelos

224k validar modelos

27k ventas efectivas, 3.20%

11k predicción

Formateo de Datos
(seleccionar, limpieza y construcción de datos)



Análisis descriptivo

Table

Numero de Hijos

Value	Frequency	Percent
0	236,088	31.28
1	226,444	30.00
2	208,676	27.65
3	66,775	8.85
4	12,741	1.69
5	2,861	0.38
6	798	0.11
7	284	0.04
8	98	0.01
9	50	0.01
10	12	0.00
13	4	0.00
11	1	0.00
12	1	0.00
21	1	0.00

Estado Civil

Value	Frequency	Percent
CASADO	440,938	58.42
SOLTERO	221,862	29.39
DIVORCIADO	77,852	10.31
UNION LIBRE	6,949	0.92
VIUDO	4,386	0.58
ND	2,847	0.38

PEP

Value	Frequency	Percent
0	751,132	99.51
1	3,702	0.49

This field has a small number of unique values, and appears to be categorical. Consider changing the field data type to "string" if

PERFIL_PRODUCTIVIDAD

Value	Frequency	Percent
Productivos Maduros	379,766	50.31

Registro	Name	Field Category	Min	Max	Median
1	EXPLICACION	String	[Null]	[Null]	[Null]
2	PLAN_ACCION	String	[Null]	[Null]	[Null]
3	FCH_FALLECIMIENTO	Date Time	[Null]	[Null]	[Null]
4	EQSALDOCOMERCIAL	String	[Null]	[Null]	[Null]
5	MNT_INGRESO_INDEPENDIENTE	Numeric	0	121.241,67	1.380
6	ESTADO_INDEPENDIENTE	String	[Null]	[Null]	[Null]
7	EQSALDOMICROCREDITO	String	[Null]	[Null]	[Null]
8	FCH_SALIDA	Date Time	[Null]	[Null]	[Null]
9	EQSALDOVIVIENDA	Numeric	25,79	1.373.736,52	46.766,54
10	CALCULO	String	[Null]	[Null]	[Null]
11	METRICA	String	[Null]	[Null]	[Null]
12	MNT_SALARIO_MAXIMO	Numeric	700	6.000	3.000
13	MNT_SALARIO_MINIMO	Numeric	400,01	6.000,01	1.400,01
14	SUBSEGMENTO2	String	[Null]	[Null]	[Null]

Table Barc... | Frequency | Frequency | Rango.Edad | Compra.SI_NO

Compra.SI_NO \ Rango.Edad	18 A 24 AÑOS	25 A 35 AÑOS	36 A 45 AÑOS	46 A 55 AÑOS	56 A 64 AÑOS	MAYOR O IGUAL A 65 AÑOS	MENOR A 18 AÑOS	ND	Totals
NO	3,487	218,416	317,538	175,224	19,169	835	506	2,792	**737,967**
SI	726	6,415	6,602	2,587	390	95	35	17	**16,867**
Totals	**4,213**	**224,831**	**324,140**	**177,811**	**19,559**	**930**	**541**	**2,809**	**754,834**

SALARIO

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
854394	0	9754	0.999	1771	1316	0	0	1078	1500	2180	3030	4000

lowest : -12100.00 -6000.00 -1152.72 -300.00 0.00,
highest: 98546.00 105483.00 130851.00 136974.00 257580.00

Perfil de Cliente

Rango Edad

36 A 45 AÑOS	38.70%
25 A 35 AÑOS	37.99%

Estado Civil

CASADO	53.48%
SOLTERO	31.30%

Salario

DE 0 A 700	28.12%
DE 701 A 1500	27.46%
DE 1500 A 2500	23.06%

Deudas

Al día	83.13%

Número Hijos

1 A 2 HIJOS	50.16%
NO TIENE HIJOS	34.97%

Relación de Dependencia

SI	72.99%

Sexo

MASCULINO	53.04%
FEMENINO	46.13%

Canal Entrada

REFERIDO	55.70%
CAMPAÑA MKT	37.40%

Ciclo de Vida

PAREJA CON HIJOS	46.35%
MONOPARENTAL	26.01%
SOLOS	18.39%

Resultado Experto

AA	32.93%
C	28.55%
A	14.19%
B	13.91%

Segmento

Ejecutivo Medio	29.30%
Ejecutivo Alto	22.76%
Profesional Independiente	11.64%
Ejecutivo VIP	9.80%

Bancarizado

SI	97.52%

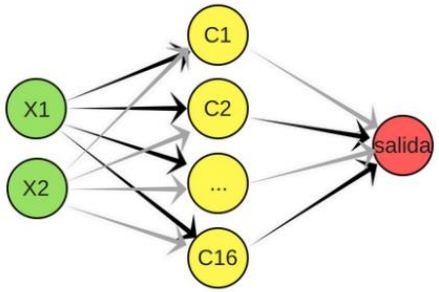
Generación

MILLENNIALS	35.31%
XENNIALS	31.71%
GENERACIÓN X	25.62%

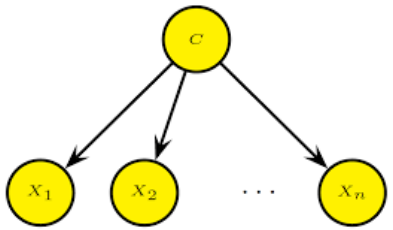
Grupo Etáreo

ADULTO JOVEN	61.65%
ADULTO MEDIO	30.01%

Selección de modelos

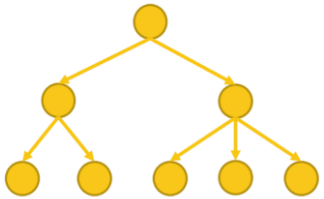


Redes Neurales: Por lo general, las neuronas se agregan en capas. Diferentes capas pueden realizar diferentes transformaciones en sus entradas.

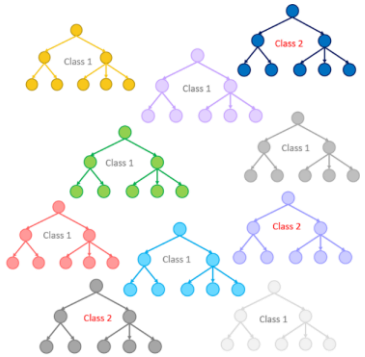


Clasificación Bayesiana: Crea un modelo de clasificación probabilística binomial o multinomial de la relación entre un conjunto de variables predictoras y una variable objetivo categórica.

Selección de modelos

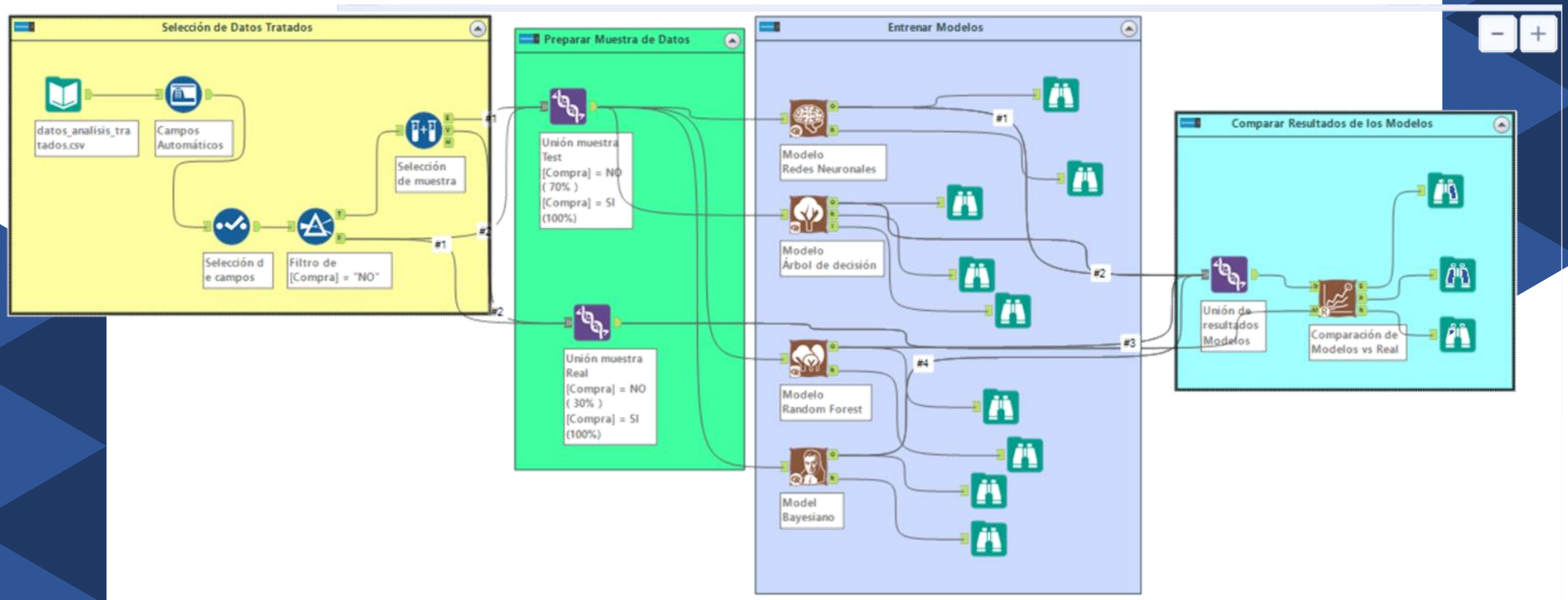


Árbol de decisión: Se utiliza para crear un conjunto de reglas de división si-entonces para optimizar los criterios de creación de modelos en función de los métodos de aprendizaje del árbol de decisiones.



Random Forest: crea un modelo que genera un conjunto de modelos de árboles de decisión para predecir una variable de destino en función de una o más variables predictoras.

Selección y configuración de modelos



Resultado de los modelos

Model Comparison Report

Fit and error measures

Model	Accuracy	Accuracy_No	F1	AUC	Accuracy_SI
ModelRedNeuronal	0.9177	0.9958	0.9559	0.7685	0.2446
ModelForest	0.9243	0.9989	0.9594	0.8136	0.2825
ModelBayesiano	0.9174	0.9791	0.9550	0.8030	0.3859
Árbol_de_decisión_77	0.9448	0.9940	0.9699	0.8399	0.5212

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of ModelBayesiano

	Actual_NO	Actual_SI
Predicted_NO	228342	16631
Predicted_SI	4872	10451

Confusion matrix of ModelForest

	Actual_NO	Actual_SI
Predicted_NO	232948	19431
Predicted_SI	266	7651

Confusion matrix of ModelRedNeuronal

	Actual_NO	Actual_SI
Predicted_NO	232238	20459
Predicted_SI	976	6623

Confusion matrix of Árbol_de_decisión_77

	Actual_NO	Actual_SI
Predicted_NO	231816	12966
Predicted_SI	1398	14116

Evaluación de resultados

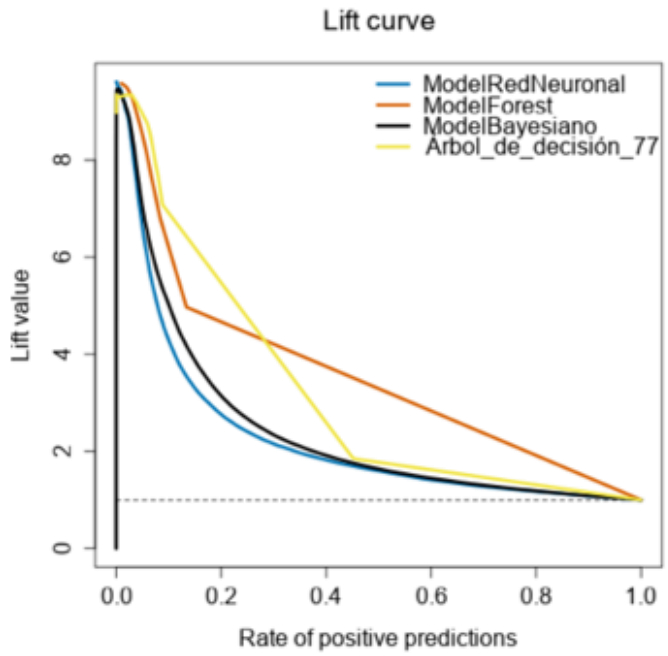
Para evaluar un modelo, una de las medidas más importantes a considerar es la Matriz de Confusión.

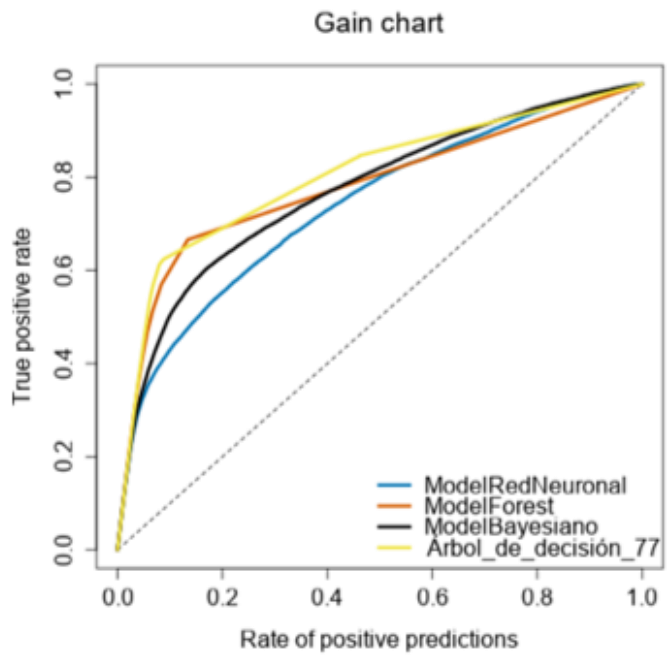
		Predicted Condition	
		Positive (PP)	Negative (PN)
Actual Condition	Positive (P)	True Positive (TP) hit	False Negative (FN) miss, underestimation
	Negative (N)	False Positive (FP) false alarm, overestimation	True Negative (TN) correct rejection

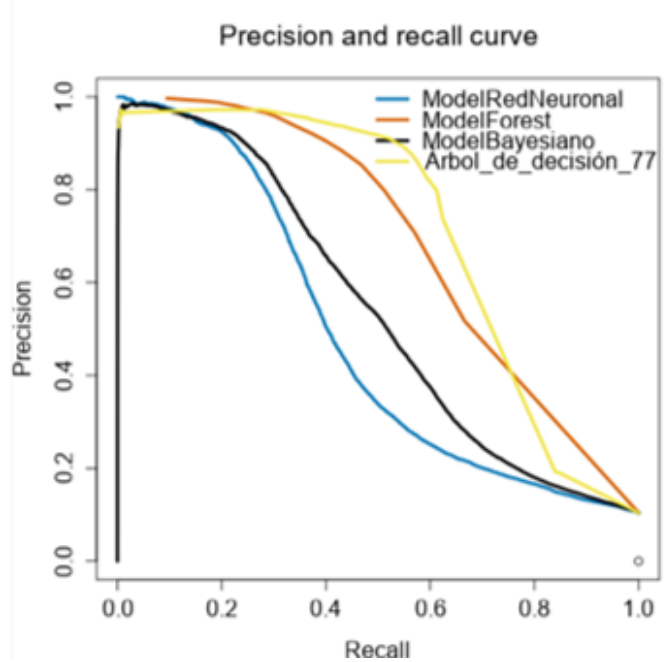
MODEL		ACTUAL	
	PREDICTION	NO	YES
Bayesian Method	NO	228,342	16,631
	YES	4,872	10,451
Neural Network	NO	232,238	20,459
	YES	976	6,623
Random Forest	NO	232,948	19,431
	YES	266	7,651
Decision Tree	NO	231,816	12,966
	YES	1,398	14,116

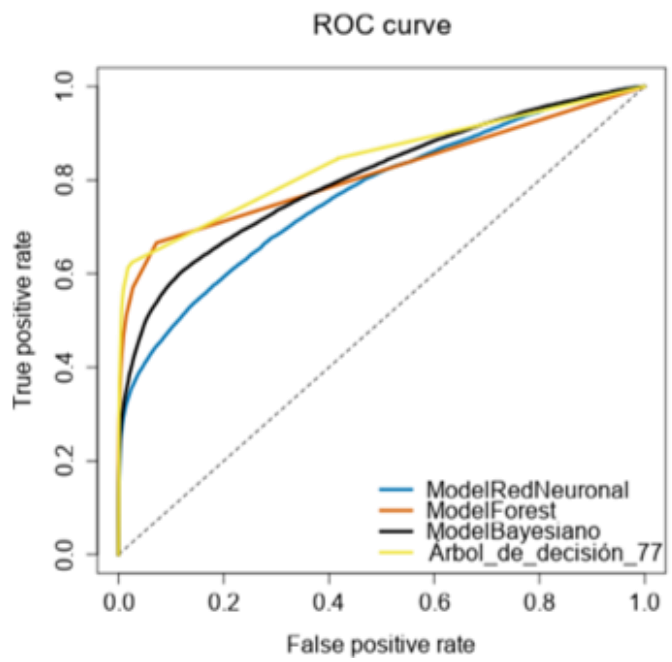
Model	Accuracy	Accuracy NO	F1	AUC	Accuracy YES
Bayesian Method	0.9174	0.9791	0.9550	0.8030	0.3859
Neuronal Network	0.9177	0.9958	0.9559	0.7685	0.2446
Random Forest	0.9243	0.9989	0.9594	0.8136	0.2825
Decision Tree	0.9448	0.9940	0.9699	0.8399	0.5212

Resultado de los modelos

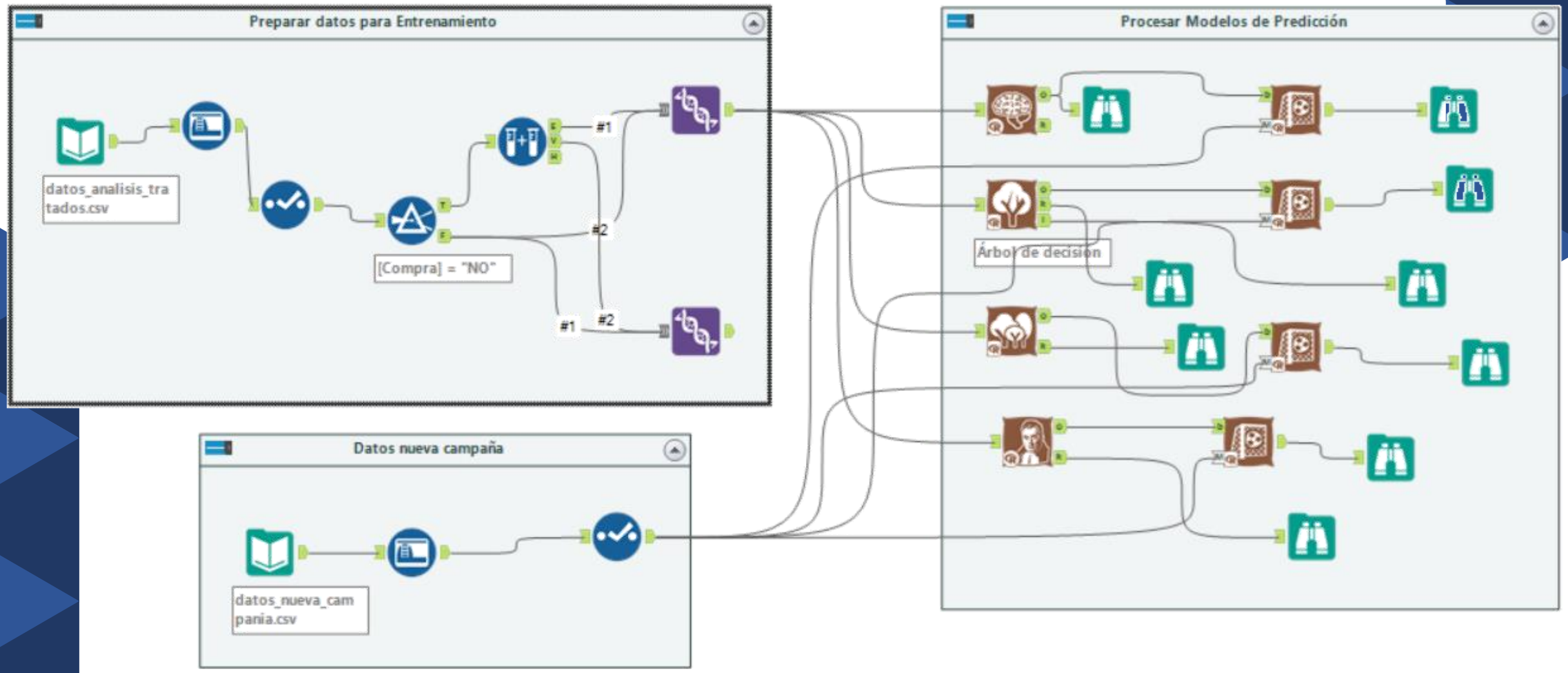








Correr la Predicción



Reportar a los tomadores de decisión los resultados

Precisión SI: se define como el número de casos que se pronosticaron correctamente como Clase SI dividido por el número total de casos que realmente pertenecen a Clase SI, esta medida también se conoce como Accuracy.

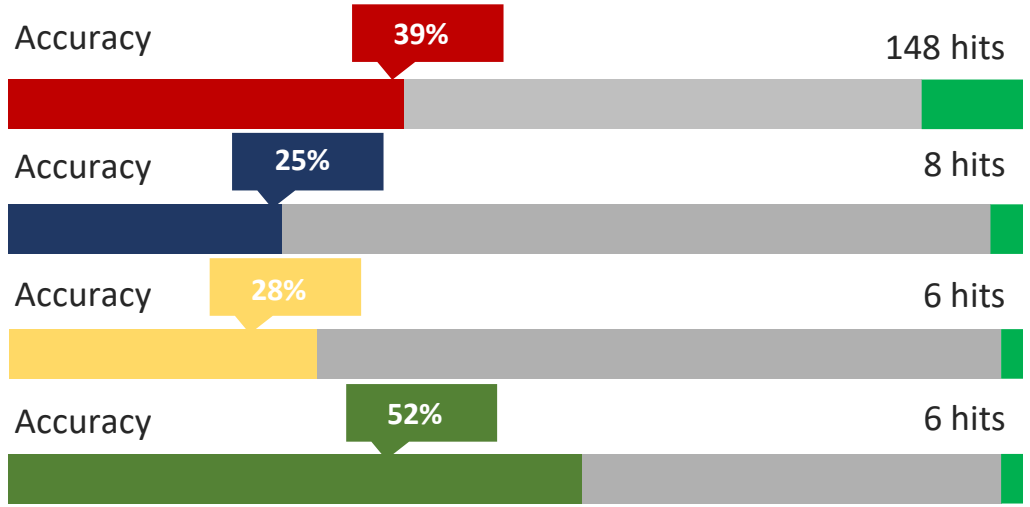
Bayesian Method

Random Forest

Neuronal Network

Decision Tree

Presición SI/Número de Predicciones



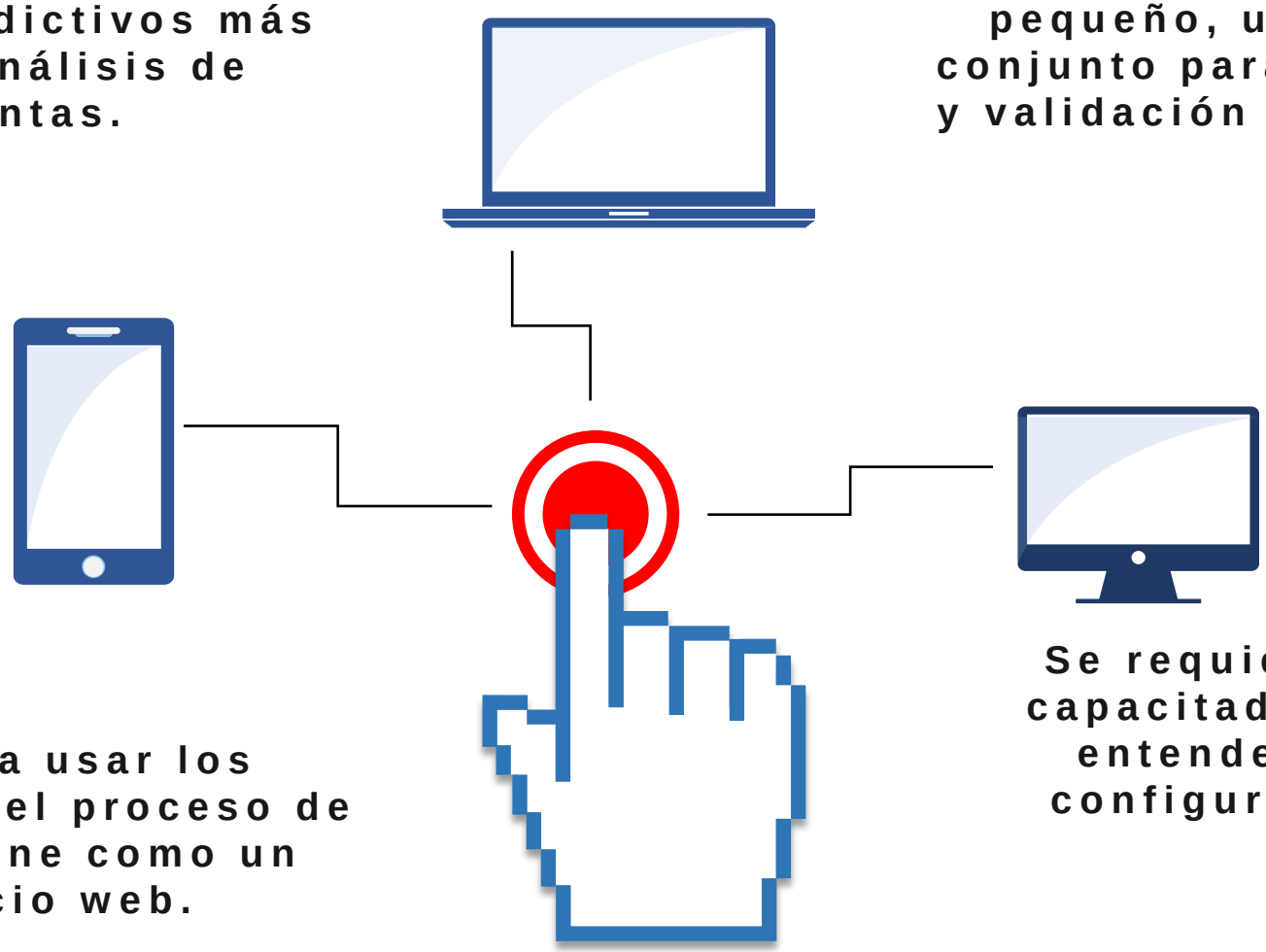
04-Discusión, conclusiones y recommendaciones

Conclusiones y Recomendaciones

Guía metodológica para iniciar un proceso de análisis de datos.

Tiene las principales metodologías, los modelos predictivos más utilizados y análisis de herramientas.

Importante cuando el número de datos positivos es pequeño, utilizar todo el conjunto para entrenamiento y validación de los modelos.



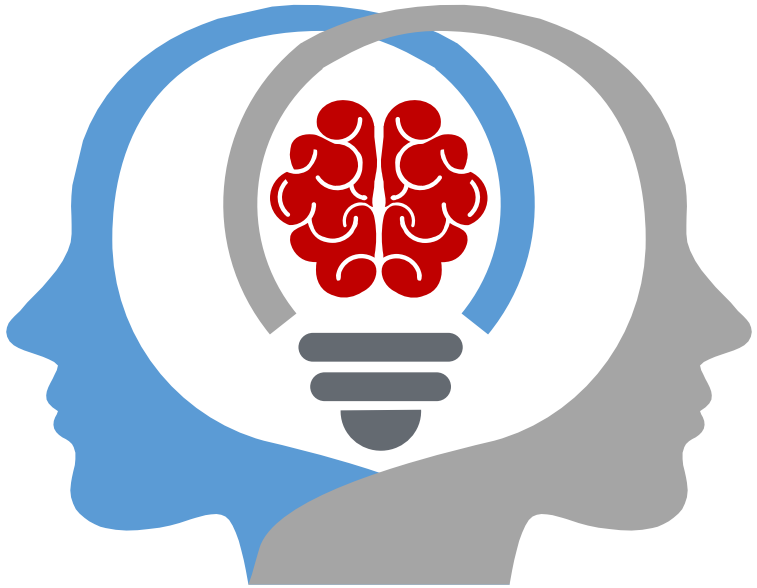
Se desea usar los modelos en el proceso de venta online como un servicio web.

Se requiere de personas capacitadas, para definir, entender resultados y configurar los modelos.

Conclusiones y Recomendaciones

Con la implementación se desea elevar ventas en un 30% al incluir enriquecimiento de datos demográficos y socioeconómicos.

Mejor modelo: árbol de decisión con un 52%.



El acceso a los datos es un tema crítico y el proceso de obtención, limpieza y carga de datos es un proceso largo

Se logró conocer y segmentar a nuestros clientes al agruparlos por edad, número de hijos, sexo, etc.

Conclusiones y Recomendaciones

Es importante contar con personas que conozcan el negocio y donde se graban los datos en los diferentes sistemas de la empresa.

Se pudo determinar la calidad de los datos.

Se deben usar varios modelos y comparar sus resultados.

El uso de herramientas reduce los tiempos de creación de modelos lo que permite mayor enfoque en la preparación y el análisis de datos.

Artículo publicado en Springer

Presentado en el 11th Congreso Internacional de Mejora de Procesos de Software 2002

Octubre 19-21

Acapulco, Guerrero, México

Springer

Congratulations!

Dear José Quintana Cruz,

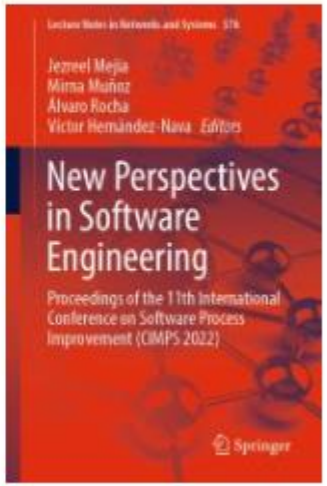
Congratulations! The eBook you contributed to has now been published on SpringerLink – our content platform, which offers customers and library patrons access to your publication at their workplace, home or wherever they want to read it. Readers who prefer a printed edition of your book will be able to order it from SpringerLink shortly.

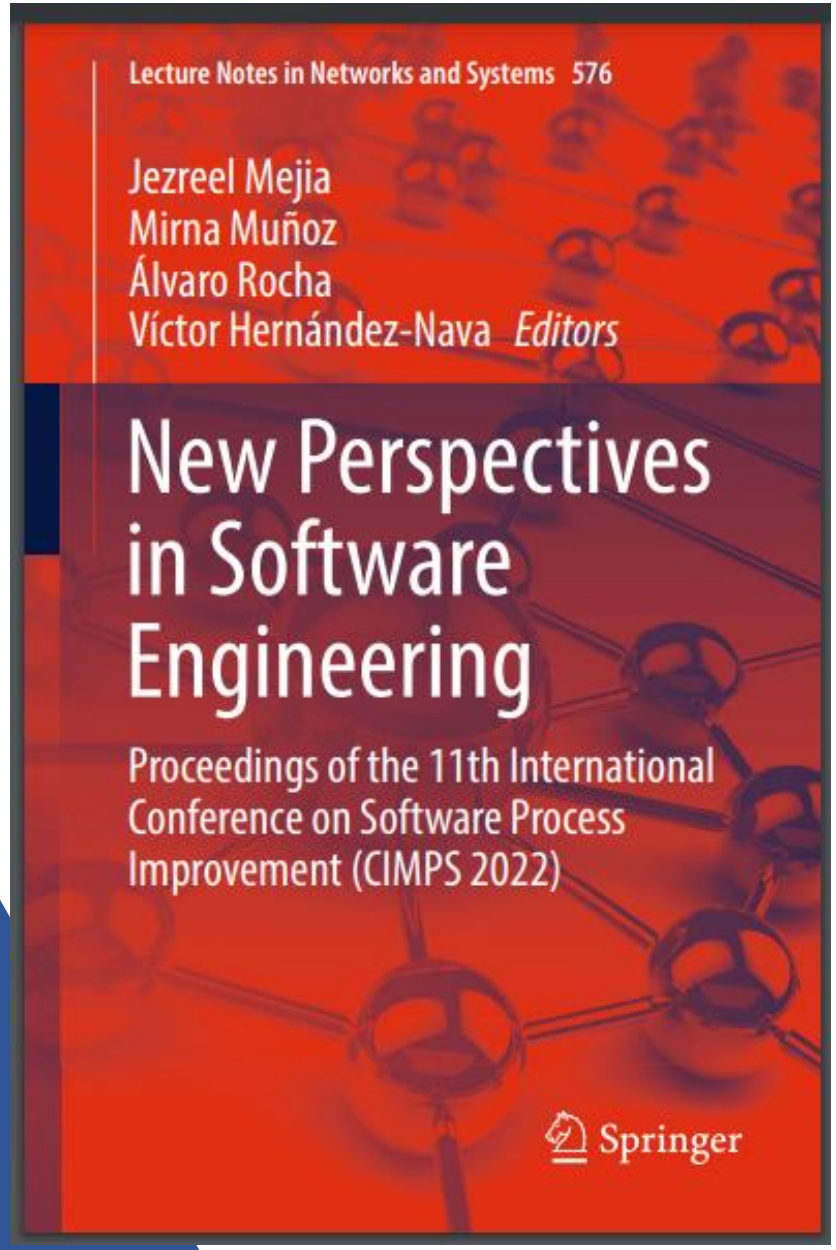
We would like to extend our best wishes for the success of your new publication and hope you enjoyed working with us.

New Perspectives in Software Engineering

eBook ISBN
978-3-031-20322-0

Print ISBN
978-3-031-20321-3





Data Mining Prospective Associated with the Purchase of Life Insurance Through Predictive Models

José Quintana Cruz and Freddy Tapia(✉)

Department of Computer Science, Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
{jaquintana,fmtapia}@espe.edu.ec

Abstract. This work proposes the creation of an analytical model which allows improving the effectiveness of sales through the use of business intelligence and data mining methodologies which allow analyzing the historical information of corporate clients and determining the probability of buying a product. In this research, methodologies, and tools are used that allow structuring the steps to define the task, collect and analyze data, choose and configure the model, format data, evaluate results and report them to decision-makers. This will allow testing various analytical models that train them and compare them with historical data, provide new data, which eventually will help increase sales effectiveness, highlighting that the data used for this analysis are demographic data, socio-economic aspects, and any information that contributes to having a framework to be reused in future sales campaigns.

Keywords: Life insurance · Data mining · Data warehouse · Data mart · CRISP-DM · SEMMA · KDD · Data science · Machine learning platforms · Random forest · Decision tree · Neural networks · Bayesian

1 Introduction

All companies currently want to analyze information to take advantage of making more timely business decisions, but not having the right tools at hand means opportunities are lost. In this context what you want is to take advantage of the information and create an analytical model that indicates the probability of purchase. The first step is to define a methodology to structure a reusable model. Then what must be done is to identify the data that the company has to run analytical models and take advantage of that information for efficient decision making.

In their article, Thuring F. Nielsen J. P., Guillén M., and Bolancé C. mention that "insurance policies or credit instruments are financial products that involve a long-term relationship between the customer and the company." For many companies a possible way to expand their business is to sell more products to preferred customers in their portfolio. Data on the customers' past behavior is stored in the company's database, and this data can be used to assess whether or not more products should be offered to a

J. Mejia et al. (Eds.): CIMPS 2022, LNNS 576, pp. 165–179, 2023.
https://doi.org/10.1007/978-3-031-20322-0_12


EL CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS A.C.
LA UNIVERSIDAD HIPÓCRATES & EL COORPORATIVO CAMSA
CIMPS
CONGRESO INTERNACIONAL DE MEJORA DE PROCESOS DE SOFTWARE
2022
In recognition and appreciation to
José Quintana Cruz and Freddy Tapia.
At the international Conference CIMPS 2022 with their article's presentation:
Data Mining Prospective Associated with the Purchase of Life Insurance through Predictive Models.
CIMPS was helt at the Hippocrates University of Acapulco, Guerrero, México, October 19-21, 2022.
SpringerLink
Dr. Jezreel Mejía Miranda
CIMPS Chair
CIMAT A.C. Unidad Zacatecas, México
Dr. Jair de Jesús Cambrón Navarrete
CEO Corporativo CAMSA
CIMPS 2022 ACAPULCO
maily
CAMSA
Conta LISTO
CIMPS
UH HIPÓCRATES
IEEE
canieti
Springer

Gracias