Pontificia Universidad Católica del Perú

Facultad de Ciencias e Ingeniería

Sección Ingeniería Industrial

DATA MINING

Entrega Final

Presentado por:

CÓDIGO	*NOMBRES Y APELLIDOS*	*% PARTICIPACIÓN*
20125873	Jhersonn Jhulians Ramirez Canchos	100%
20142018	Eduardo Sebastián Villanca Rosales	100%
20143213	Bryan Hugo Quispe Ramirez	100%

Profesor: Luciano Silva
Horario: 1031
Semestre: 2020-2

San Miguel, 18 de diciembre de 2020

RESUMEN EJECUTIVO

El siguiente proyecto consiste en predecir el éxito de una campaña de marketing para un banco en particular, mediante aplicaciones de Data Mining en el programa Orange se determinará si algunas variables como la educación, el estado civil o el índice de masa corporal presentan efectos directos sobre la posibilidad de contraer esta enfermedad, se analizarán estos casos y similares como la falta de datos y se trabajará con métodos de preprocesamiento para resolverlos, automáticamente después se buscará un modelo, se especificará y evaluará respecto a otros, el fin del proyecto se dará al tener un archivo final donde la variable de predicción será realizada por todas las actividades realizadas en la computadora**.**

Introducción 4

Título 4

Objetivo del proyecto 4

Definición de las variables **4¡Error! Marcador no definido.**

Proceso ETL 5

Preprocesamiento 6

Imputación 7

Continuación 9

División de la data 10

Estandarización 11

Presentación del modelo 13

Evaluación de modelos 16

Matriz de confusión 17

ROC 18

Optimización 19

Conclusiones y recomendaciones 22

I. Introducción

Los bancos han cambiado la estrategia de captar nuevos clientes, pasando de encuentros netamente físicos a llamadas telefónicas. La comunicación de la propuesta debe ser personalizada para que el cliente se sienta interesado desde el primer segundo de contacto. Además, se deben seguir ciertas pautas para obtener una mayor efectividad: usar un tono de voz alegre y segura, ser preciso en la información, adaptarse a la situación y saber escuchar.

Considerando que cada llamada es una valiosa fuente de información, se deberá monitorear cuándo fue hecha, cuánto fue su duración, redactar un resumen y registrar a qué personas fueron hechas. Entonces, bajo esta premisa se va a utilizar una base de datos de una campaña de marketing de una institución bancaria de Portugal que recolectó información entre mayo del 2008 y noviembre de 2010 para predecir si un cliente realizará o no un depósito.

II. Título

Predicción del éxito de una campaña de marketing en un banco

III. Objetivo del proyecto

Conocer si el cliente aceptará realizar un depósito a plazo, antes de realizar la llamada de ofrecimiento de parte del área de marketing del banco.

Propuesta de valor:
Obtención de *insights* importantes sobre nuestros clientes y conocer qué variables de estos son relevantes al momento de elegir un depósito.

Definición de las variables

Tabla 1: Variables de entrada

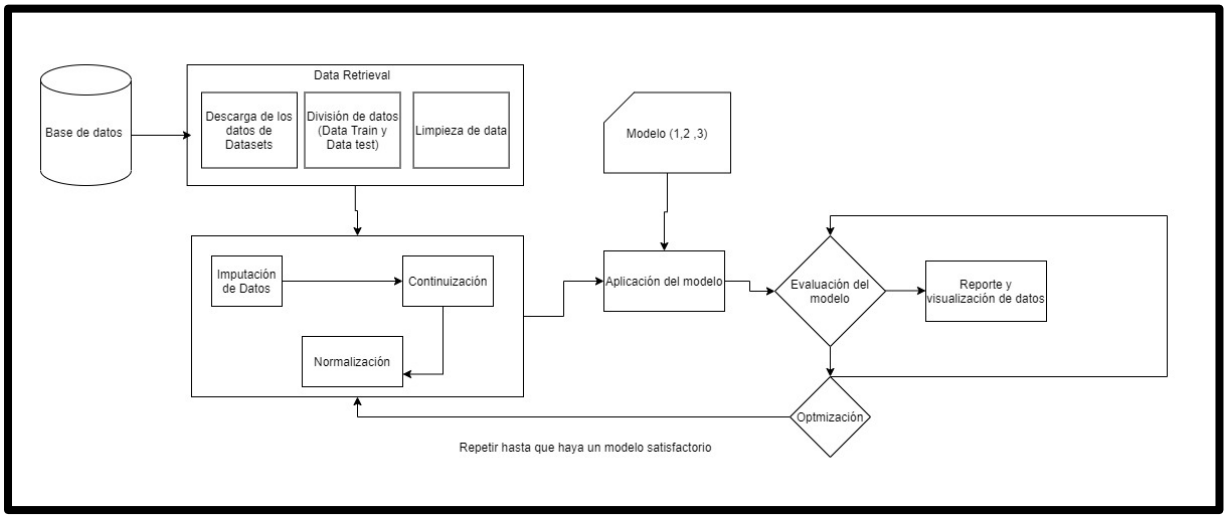
Campo	Descripción	Respuesta	Tipo
Age	**Edad del cliente**	Valor entero, en años	**Numérica**
Job	**Tipo de empleo del cliente**	"admin","unknown","unemployed", "management","housemaid","entrepreneur","student",”blue-collar","self-employed","retired","technician","services".	**Categórica**
Marital	**Estado matrimonial del cliente**	"married","divorced","single"	**Categórica**
Education	**Nivel de educación del cliente**	"unknown","secondary","primary","tertiary"	**Categórica**
Default	**Acceso a crédito por defecto del cliente**	“yes”, “no”	**Categórica**
Balance	**Consumo de alcohol**	No, a veces, frecuentemente y siempre	**Categórica**
Housing	**Si el cliente tiene préstamo de vivienda**	“yes”,”no”	**Categórica**

Loan	**Si el cliente tiene un prestamo personal**	“yes”,”no”	**Categórica**
Contact	**Tipo de contacto de comunicación**	**"unknown","telephone","cellular"**	**Categórica**
Day	**Último día de contacto en el mes**	Valor entero, en dias	**Numérica**
Month	**Último mes de contacto**	“jan”, “feb”....”dec”	**Categórica**
Duration	**Ultima duración del último contacto**	Valor entero, en segundos	**Numérica**
Otros atributos			
Campaign	**Número de contactos realizados para el cliente**	Valor entero	**Numérica**
Pdays	**Número de días desde que el cliente fue contactado**	Valor entero, “-1” si el cliente no fue contactado previamente	**Numérica**
Previous	**Número de contactos realizados previo a la campaña y para este cliente**	Valor entero	**Numérica**
Poutcome	**Resultado de campañas de marketing previas en el cliente**	"unknown","other","failure","success"	**Categórica**

Tabla 2:Variables de salida

Campo	Descripción	Respuesta	Tipo
Deposit	**Decisión si el cliente se suscribió a un depósito**	“yes” or “no”	**Categórica**

V. Proceso ETL



Gráfica 1: Proceso ETL

Para poder alcanzar los objetivos, primero a el dataset obtenido de la base de datos de la data set, se realizó un procesamiento de separación de columnas de esa manera se podía observar la dimensionalidad de la data. De esa manera se pudo observar la algunas variables tenían un gran número de celdas vacías y además se determinó que no tenían un impacto significativo en la variable objetivo de esa manera , se eliminó dicha variable.

De esa manera, se estableció tres etapas del procesamiento, dentro de las cuales, la primera era imputar por el método del promedio con la finalidad de “rellenar “ las celdas vacías. La segunda, se busca obtener una categoría estándar para todas las variables es decir se buscará transformar la variable categórica a númerica. Finalmente, se busca estandarizar todos los variables de dicha variable mediante la normalización.

Una vez ya realizado el procesamiento , se procede a evaluar qué tipo de modelo podría seguir , en una primera instancia se podría optar por bayes pero debido a la imputación este puede verse afectado , una segunda opción podría ser una regresión logística pero se conoce que este tipo modelo es preferible para agrupar y no para predecir y por último se tiene el modelo del árbol y de redes neuronales , si bien estos modelos se ajustan a los objetivos del estudio el que tiene mayor relevancia es el redes neuronales.

Por último, se evalúan los modelos para verificar cual de todos los modelos tiene una mayor precisión y cual otorga un menor error tipo II. Sin embargo, este modelo se puede ajustar o mejorar mediante aumento de las capas neuronales o también podría verse afectado por otro tipo de reprocesamiento, las cuales se deben de verificar y evaluar.

VI. Preprocesamiento

Una vez ya establecida la estrategia con la que se va abordar el problema, se procede al preprocesamiento de datos.

Antes de realizar cualquier proceso, se verificará la correlación que se tiene entre las variables independientes y la variable objetivo. La finalidad de esta verificación es observar qué variables afectan significativamente y evaluar la posibilidad de eliminar variables independientes.

	#	Inf...ain	Gai...tio	Gini	ANOVA	χ^2	ReliefF	FCBF
N duration		0.081	0.041	0.023	NA	3785.439	0.055	0.069
N previous		0.018	0.018	0.006	NA	2255.965	0.001	0.000
N pdays		0.021	0.021	0.007	NA	1184.946	0.038	0.029
C housing	2	0.014	0.014	0.004	NA	388.950	0.026	0.000
N balance		0.007	0.004	0.002	NA	367.031	0.003	0.000
N co...qn		0.006	0.003	0.001	NA	323.613	0.004	0.000
C job	11	0.012	0.004	0.004	NA	183.869	0.125	0.000
C loan	2	0.004	0.006	0.001	NA	176.516	0.028	0.007
C edu...ion	3	0.004	0.003	0.001	NA	80.117	0.033	0.000
C month	12	0.035	0.012	0.014	NA	44.322	0.100	0.000
C marital	3	0.003	0.002	0.001	NA	29.766	0.010	0.003
C default	2	0.000	0.003	0.000	NA	22.314	0.006	0.000
N day		0.003	0.001	0.001	NA	17.319	0.088	0.000
N age		0.003	0.002	0.001	NA	7.722	0.025	0.000
C contact	2	0.000	0.000	0.000	NA	3.722	0.049	0.000

Gráfico 2: Ranking de variables de entrada

Por otro lado, se verificó si las variables siguen algún tipo de distribución en específica. Esto se pudo evidenciar mediante la aplicación del widget y se pudo observar que ninguna variable seguía una distribución en específica. Por lo que más adelante se optaría por normalizar los datos

Gráfico 3: Distribución de variables sin preprocesamiento

Imputación

Entonces, se optó por imputar los datos, ya que se cuenta con el 6.8% de valores faltantes. Se completarán los valores según el promedio para variables continuas y la moda para variables categóricas. Sin embargo, antes de realizar esta imputación, se decidió eliminar la variable “poutcome” debido a que no aportaba mucho al modelo y además la gran parte de sus celdas estaban completamente vacías.

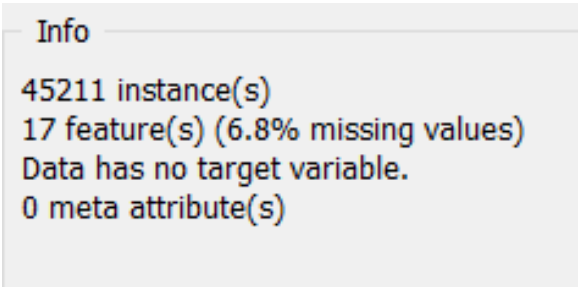


Gráfico 4: Datos faltantes

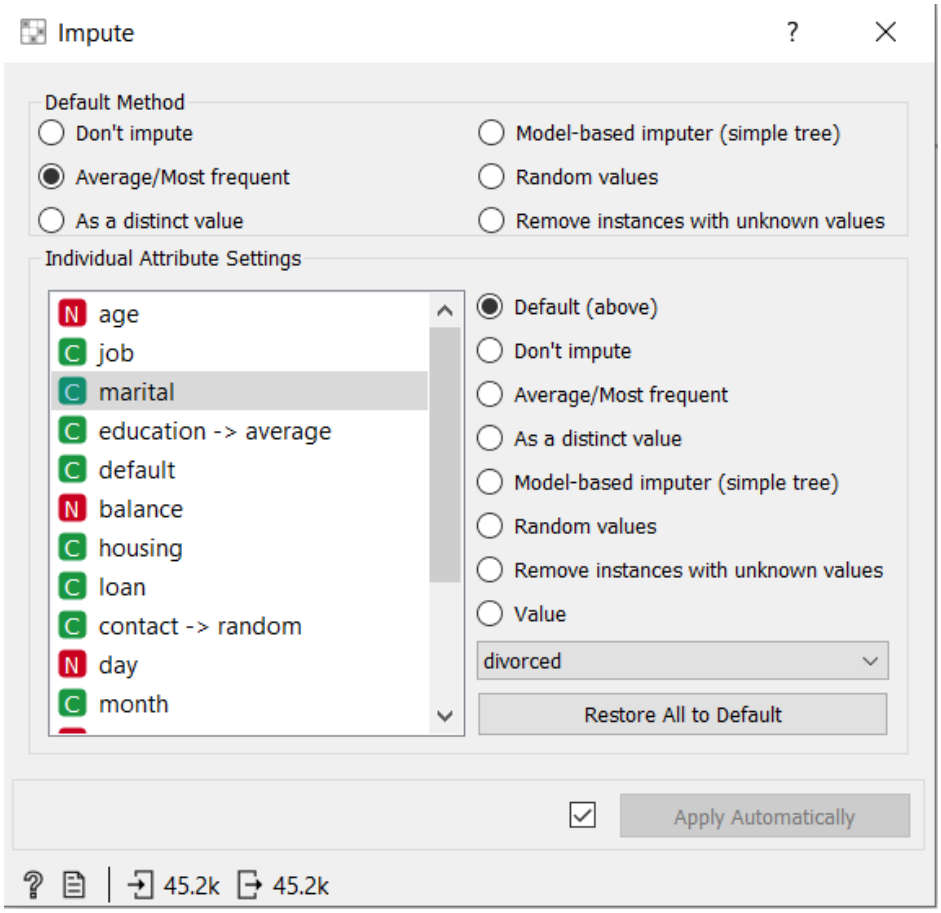


Gráfico 5: Proceso de imputación

Se puede observar que la imputación no influyó mucho en la correlación de las variables independientes con la variable objetivo.

	#	Inf...ain	Gai...tio	Gini	ANOVA	χ^2	ReliefF	FCBF
duration		0.081	0.041	0.023	NA	2785.439	0.044	0.069
previous		0.018	0.018	0.006	NA	2255.965	0.001	0.000
pdays		0.021	0.021	0.007	NA	1184.946	0.017	0.029
housing	2	0.014	0.014	0.004	NA	388.950	0.030	0.000
balance		0.007	0.004	0.002	NA	367.031	0.003	0.000
co...qn		0.006	0.003	0.001	NA	323.613	0.001	0.000
job	11	0.012	0.004	0.004	NA	184.577	0.148	0.000
loan	2	0.004	0.006	0.001	NA	176.516	0.028	0.007
edu...ion	3	0.004	0.002	0.001	NA	79.280	0.046	0.000
month	12	0.035	0.012	0.014	NA	44.322	0.070	0.000
marital	4	0.003	0.002	NA	NA	29.766	0.000	NA
default	2	0.000	0.003	0.000	NA	22.314	0.000	0.000
day		0.003	0.001	0.001	NA	17.319	0.050	0.000
age		0.003	0.002	0.001	NA	7.722	0.021	0.002
contact	2	0.000	0.000	0.000	NA	4.095	0.000	0.000

Gráfico 6: Ranking de variables imputadas

Continuación

Con este procesamiento se quiere transformar las variables categóricas y numéricas para luego estandarizarlas. En este caso se tiene 8 campos categóricos. Se utilizará una continuación a variables numéricas ordinales.

	Name	Type	Role	Values
2	job	C categorical	feature	admin., blue-collar, entrepreneur, housemaid, management, retired, self-...
3	marital	C categorical	feature	divorced, married, single
4	education	C categorical	feature	primary, secondary, tertiary
5	default	C categorical	feature	no, yes
6	balance	N numeric	feature	
7	housing	C categorical	feature	no, yes
8	loan	C categorical	feature	no, yes
9	contact	C categorical	feature	cellular, telephone
10	day	N numeric	feature	
11	month	C categorical	feature	apr, aug, dec, feb, jan, jul, jun, mar, may, nov, oct, sep

Gráfico 7: Variables categóricas

Gráfico 8: Distribuciones de variables después de la continuación

División de la data

Se dividirá el dataset principal en dos partes: el entrenamiento (train) y el testeo (test), para evitar un sobreajuste en el modelo y que no se pueda generalizar en futuras predicciones. El primero tendrá el 70% de los 45 211 registros iniciales y el segundo, solo el 30%. En ambos casos, se tendrá una misma proporción (11.6%) de la característica que se quiere evaluar (term_deposit==1).

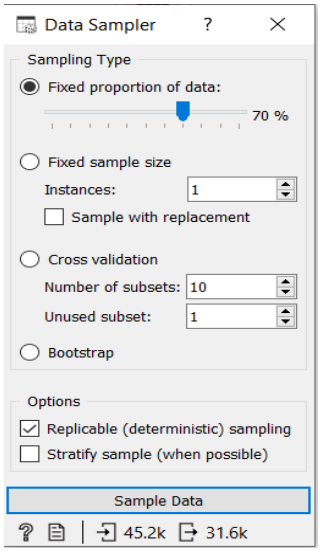


Gráfico 9: Proceso de Oversampling

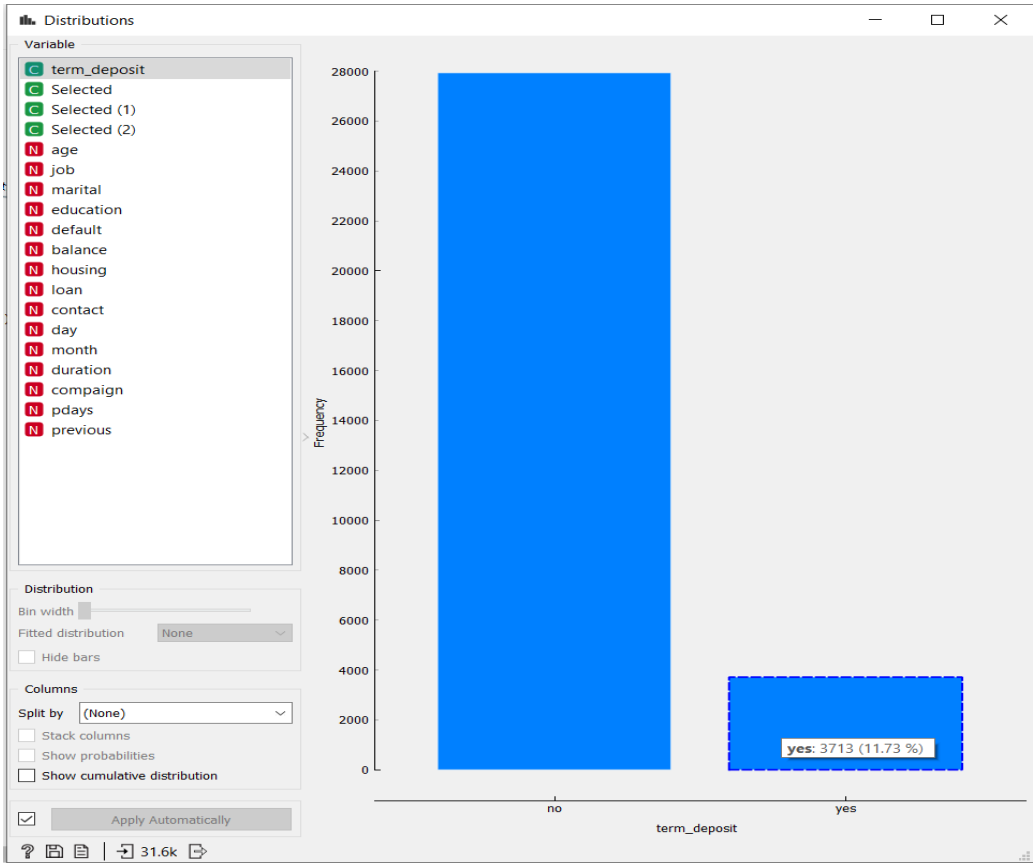


Gráfico 10: Proporciones en la data de entrenamiento

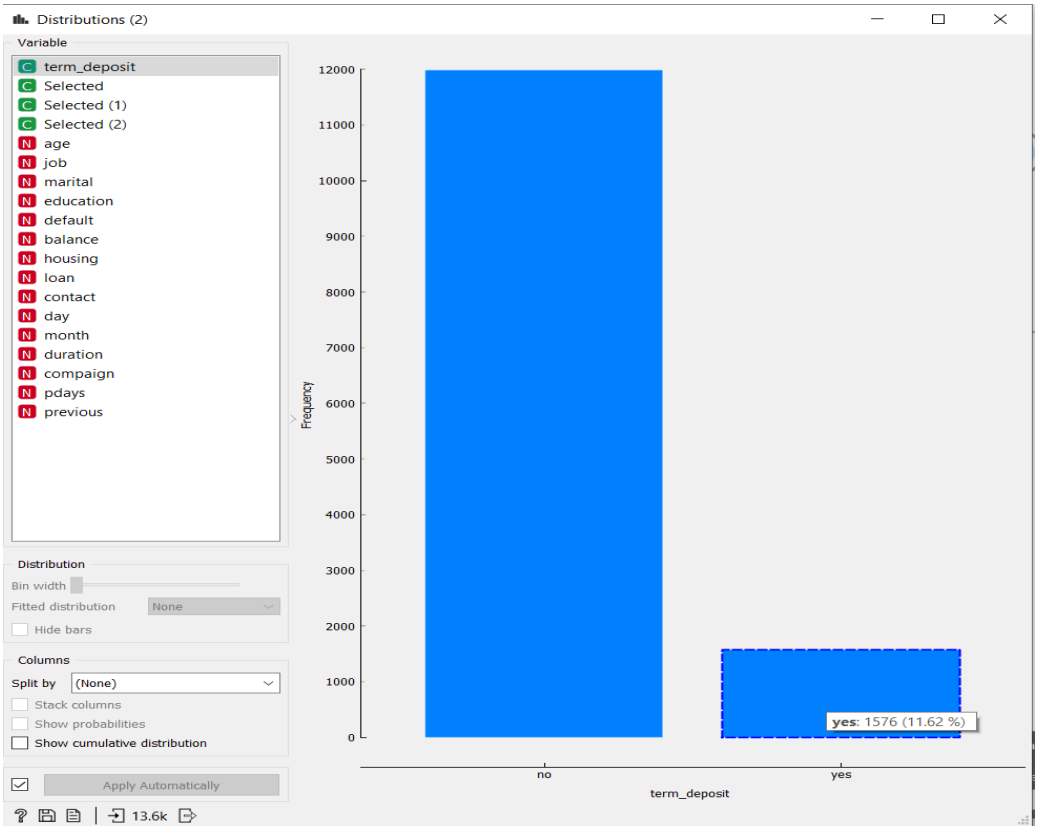


Gráfico 11: Proporciones en la data de testeo

Estandarización

Las variables están muy compactas, pero también están superpuestas por lo que al momento de predecir, una variable puede influir más en el target que otra, lo cual podría perjudicar la predicción.

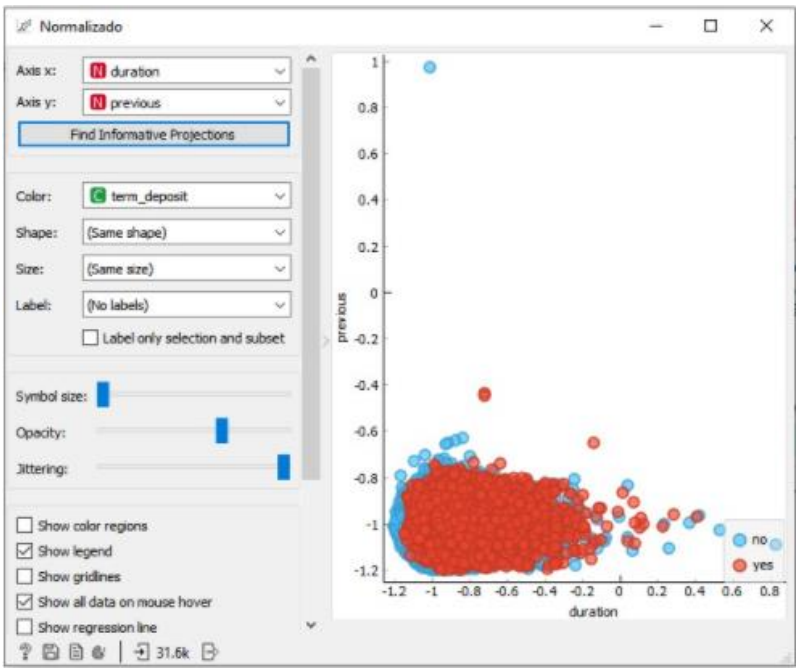


Gráfico 12: Correlación duration vs previous

En el mapa de calor, se ve que la variable Balance resalta más debido a que tiene un rango de valores más grande; lo cual puede reducir la generalización del modelo.

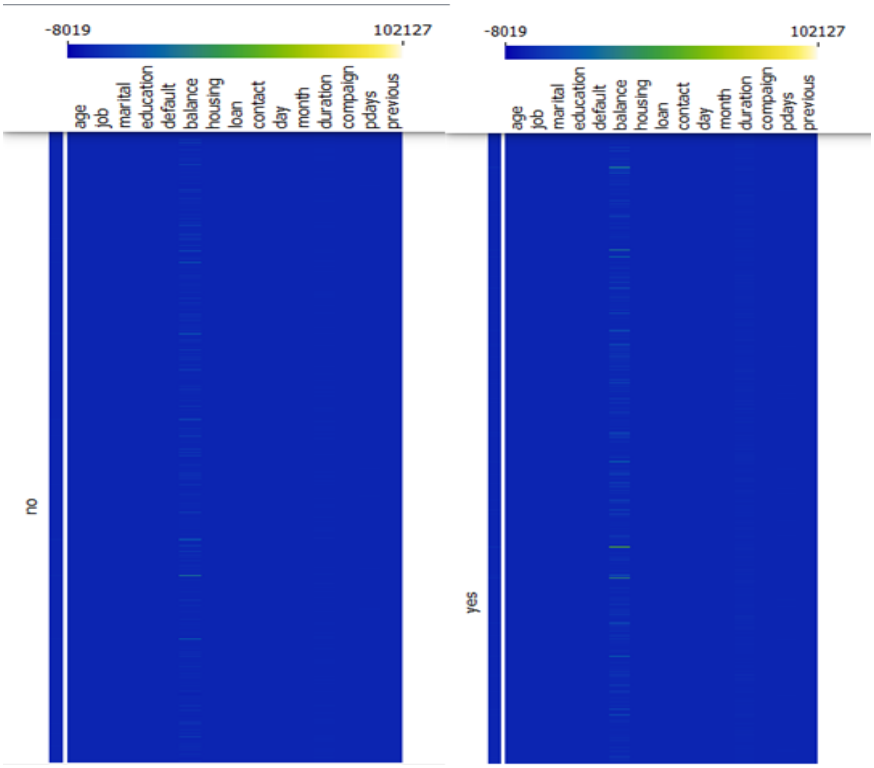


Gráfico 13: Mapa de calor de datos de entrenamiento antes de la estandarización

Ante ello, se debe reducir la dispersión de valores y tener un rango fijo para todas las variables continuas con una normalización de intervalo [-1,1]. Esto nos da como resultado la mejor participación de las otras variables en la predicción.

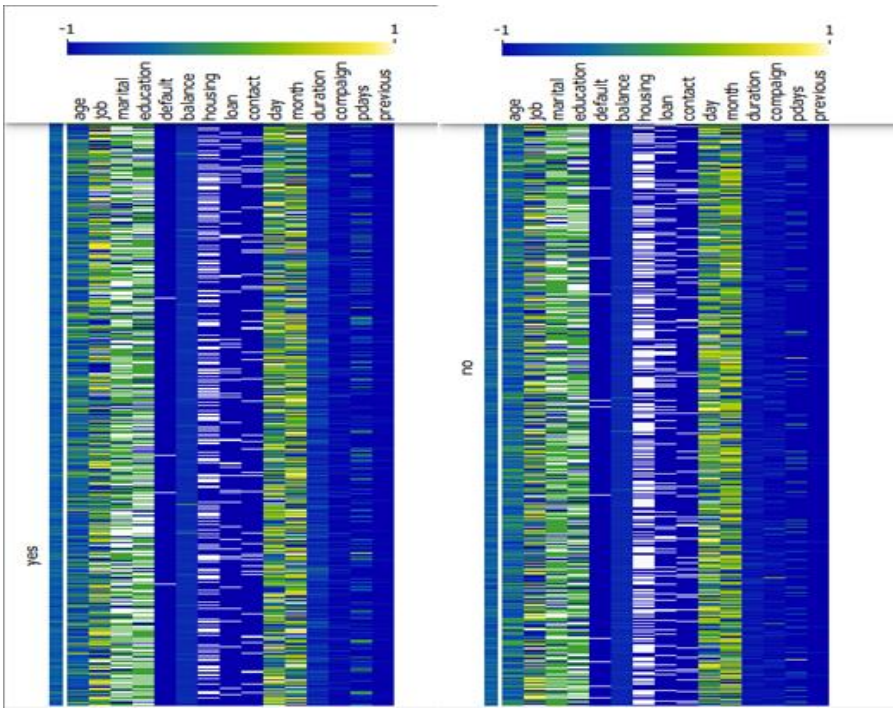


Gráfico 14: Mapa de calor de datos de entrenamiento después de la estandarización

	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	previous		-0.995756	-0.0180195	-1	1	0 (0%)
N	pdays		-0.905579	-0.252293	-1	1	0 (0%)
N	compaign		-0.943389	-0.105435	-1	1	0 (0%)
N	duration		-0.895097	-0.116586	-1	1	0 (0%)
N	month		0.00333211	164.076	-1	1	0 (0%)
N	day		-0.0149457	-37.1182	-1	1	0 (0%)
N	contact		-0.818188	-0.702713	-1	1	0 (0%)
N	loan		-0.6784	-1.08298	-1	1	0 (0%)
N	housing		0.10977	9.05491	-1	1	0 (0%)
N	balance		-0.829775	-0.0672011	-1	1	0 (0%)
N	default		-0.964358	-0.274381	-1	1	0 (0%)
N	education		0.142884	4.5784	-1	1	0 (0%)
N	marital		0.167499	3.63917	-1	1	0 (0%)
N	job		-0.146556	-4.42757	-1	1	0 (0%)
N	age		-0.405533	-0.68143	-1	1	0 (0%)

Gráfico 15: Distribución de variables después de la estandarización

VII. Presentación del modelo

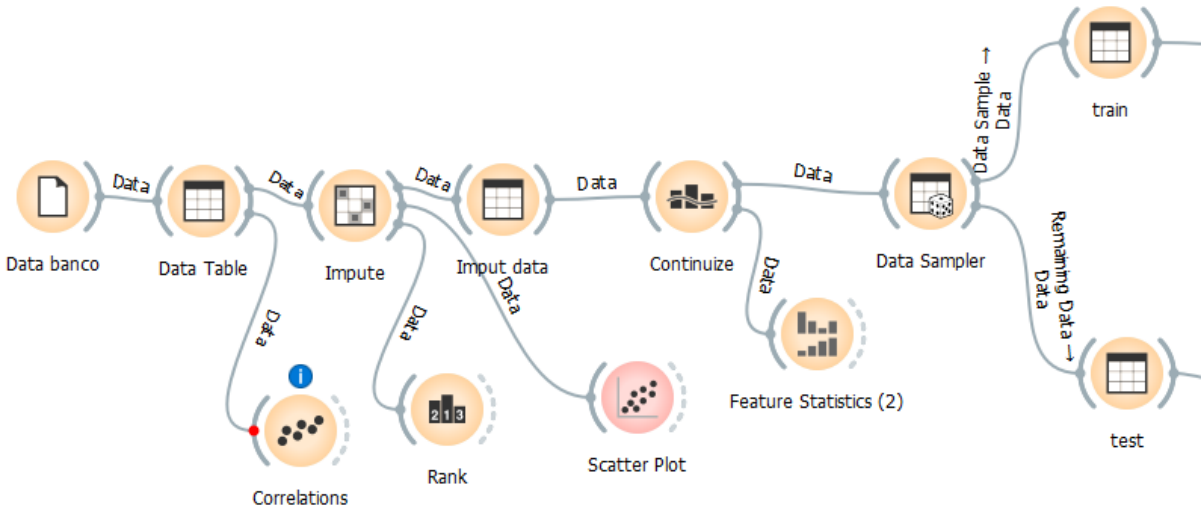


Gráfico 16: Modelo de predicción parte 1

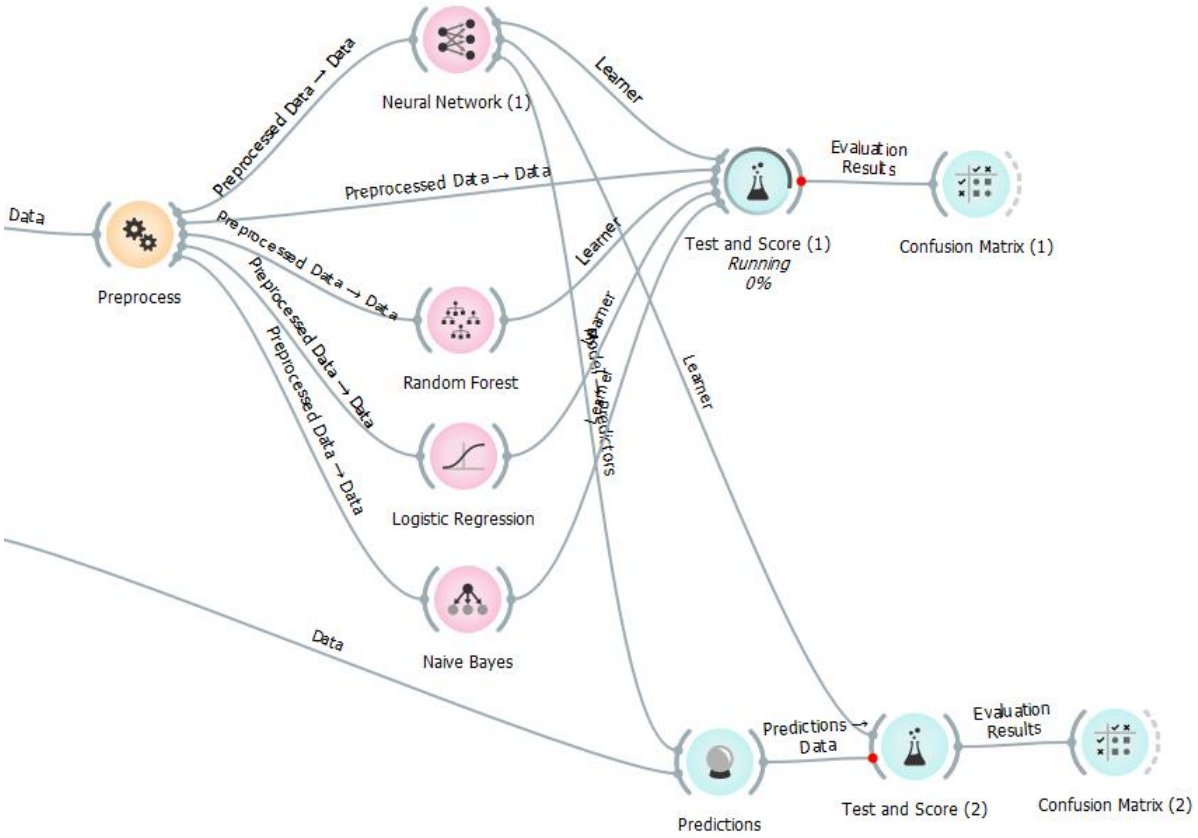


Gráfico 17: Modelo de predicción parte 2

Como se puede observar, se utilizó el mismo preprocesamiento para los 4 algoritmos, los cuales son el Naive Bayes, Redes neuronales, Random Forest y Regresión logística.

Naive Bayes:

En primer lugar se utilizó el Naive Bayes ya que se considera que no existen variables que no afecten a la respuestas buscada y que existe una dependencia entre las variables y el resultado. Finalmente este es el más rápido de los 4, ya que no consume muchos recursos computacionales.

Regresión Logística:

Este es el segundo algoritmo que se comparó ya que la variable de respuesta es dicotómica, sí o no, por lo que este modelo se ajusta para poder encontrar una relación entre las variables y la respuesta. Para el primer intento se utilizó la siguiente configuración:

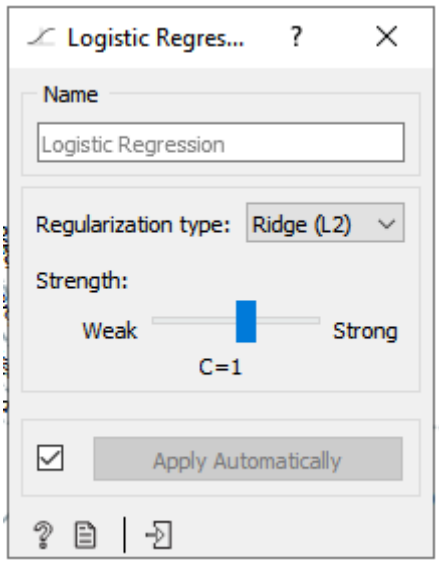


Gráfico 18: Parámetros de Regresión Logística

Random Forest:

De igual manera, al estar buscando una respuesta dicotómica, también se procedió a utilizar el random forest, se consideró pertinente usa los siguientes parámetros para el primer análisis:

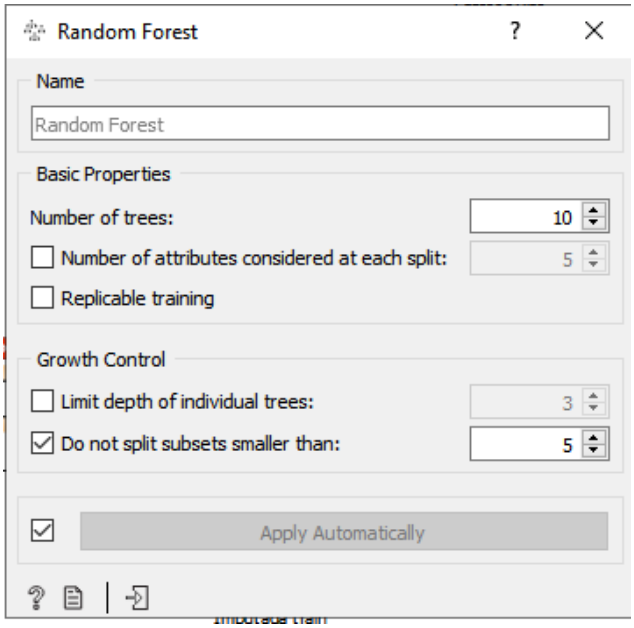


Gráfico 19: Parámetros de random Forest

Redes Neuronales:

Finalmente se usó el algoritmo de redes neuronales al ser el más completo de los 4, aun así existe la desventaja de que es el que más tiempo demora, causando que su análisis tome más tiempo. Se utilizó los siguientes parámetros iniciales:

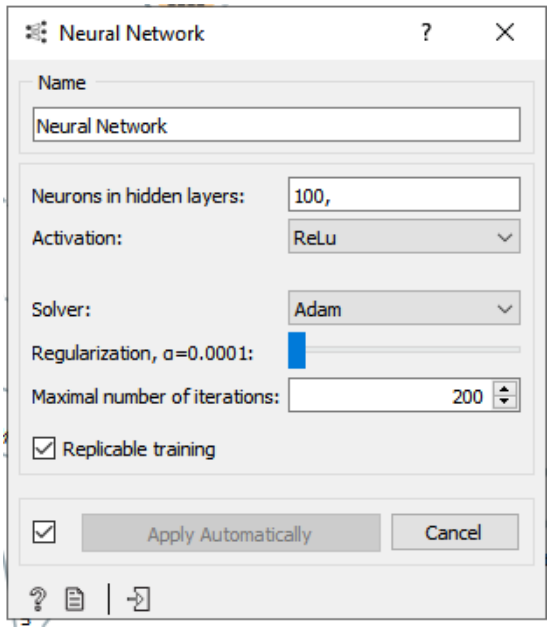


Gráfico 20:Parámetros de Redes Neuronales

VIII. Evaluación de modelos

Como parte de la evaluación, se realizará el análisis de los indicadores de los modelos.

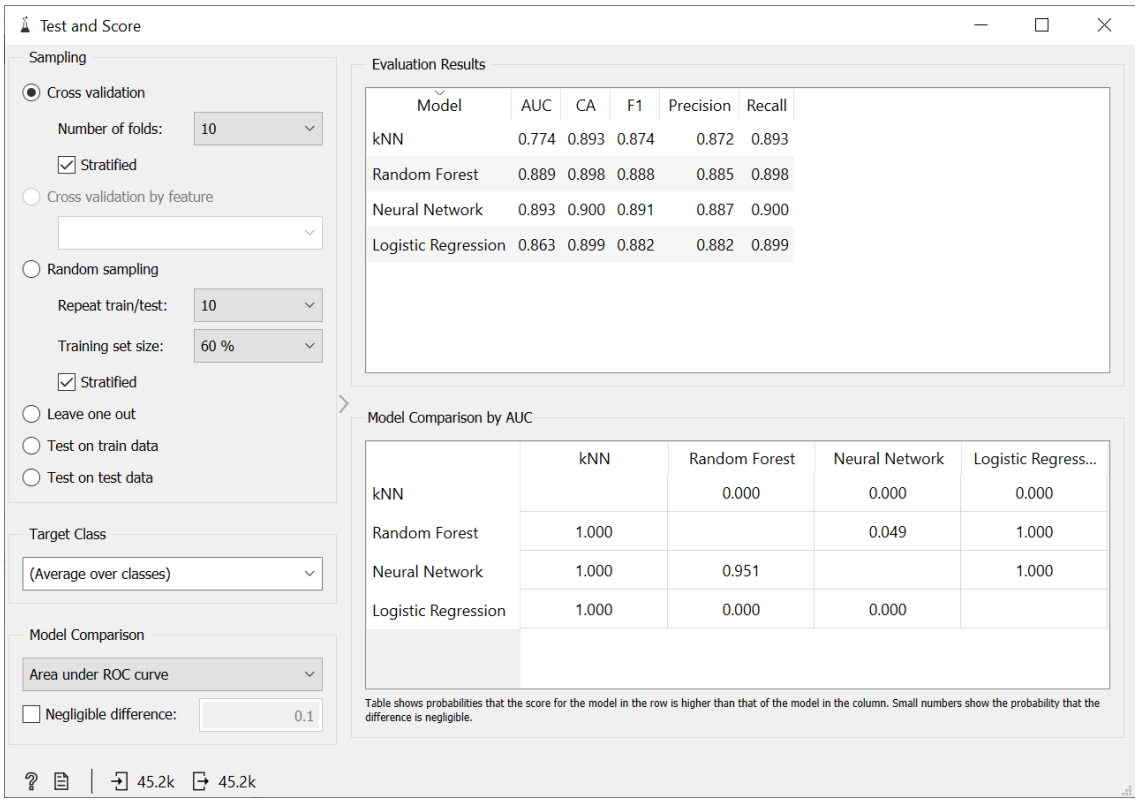


Gráfico 21:Indicadores de precisión por modelo

De los indicadores de precisión, se aprecia que el modelo que emplea Redes Neuronales y Random Forest, posee una mayor precisión de predicción, siendo el de redes ligeramente superior. Seguidos de estos se encuentra el modelo de regresión logística y como el modelo de menor precisión está el de kNN.

Matriz de confusión

El indicador de precisión de las redes neuronales (0.899) es el mayor entre los cuatro modelos empleados.

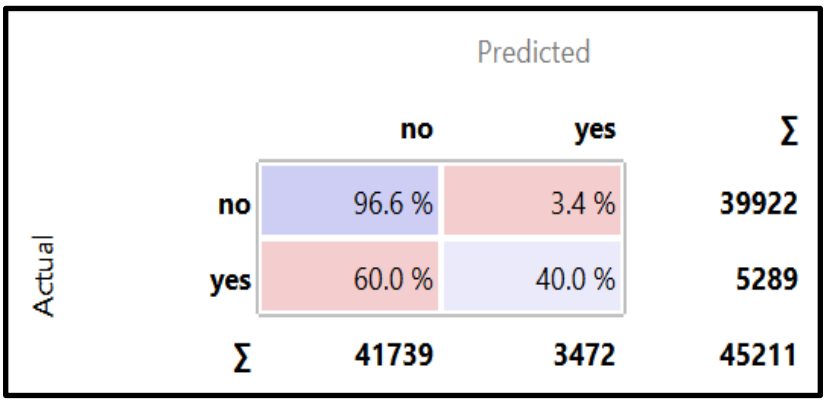


Gráfico 22: Matriz de confusión - Redes neuronales

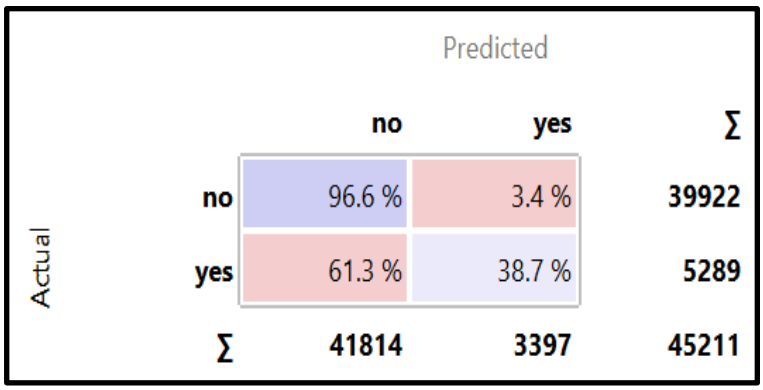


Gráfico 23: Matriz de confusión -Random Forest

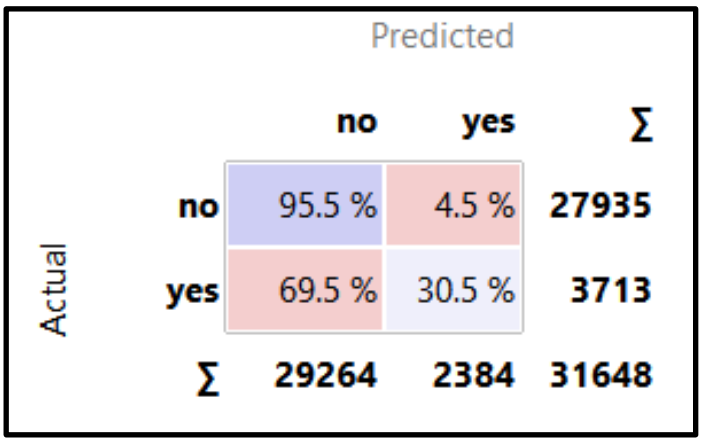


Gráfico 24: Matriz de confusión -Naive Bayes

ROC

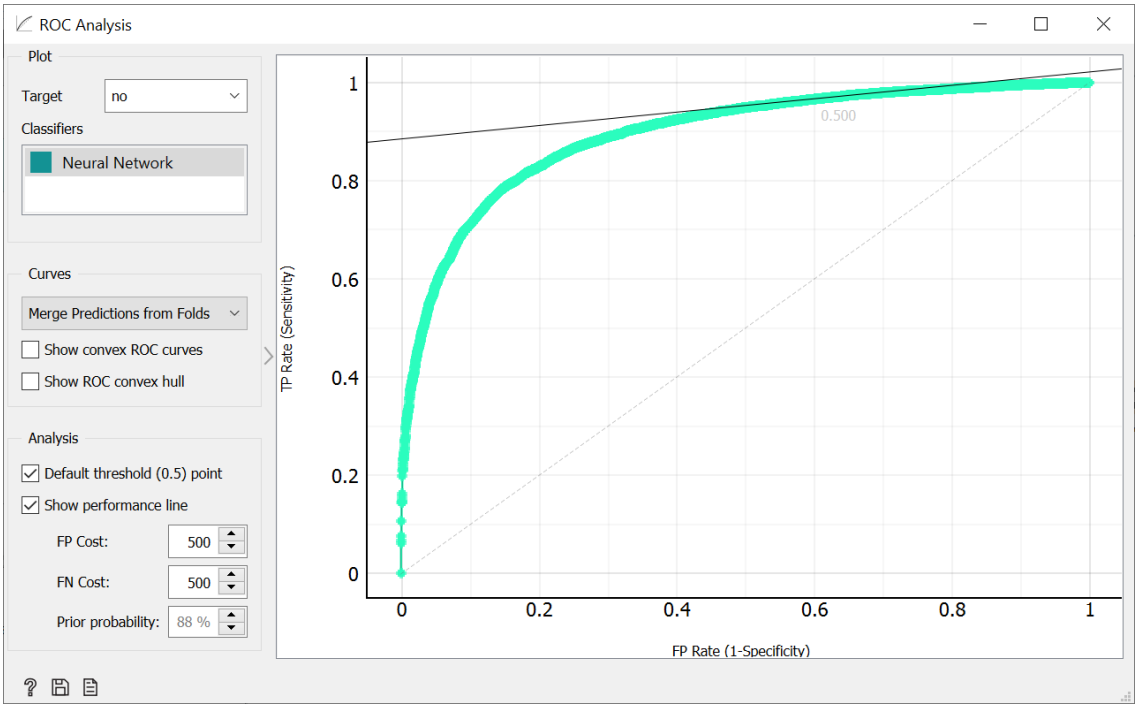


Gráfico 25: Análisis ROC Redes Neuronales

Del gráfico ROC, se aprecia que el modelo presenta una elevada precisión de predicción.

Comparados al resto de gráficas ROC, se cumple lo mencionado anteriormente.

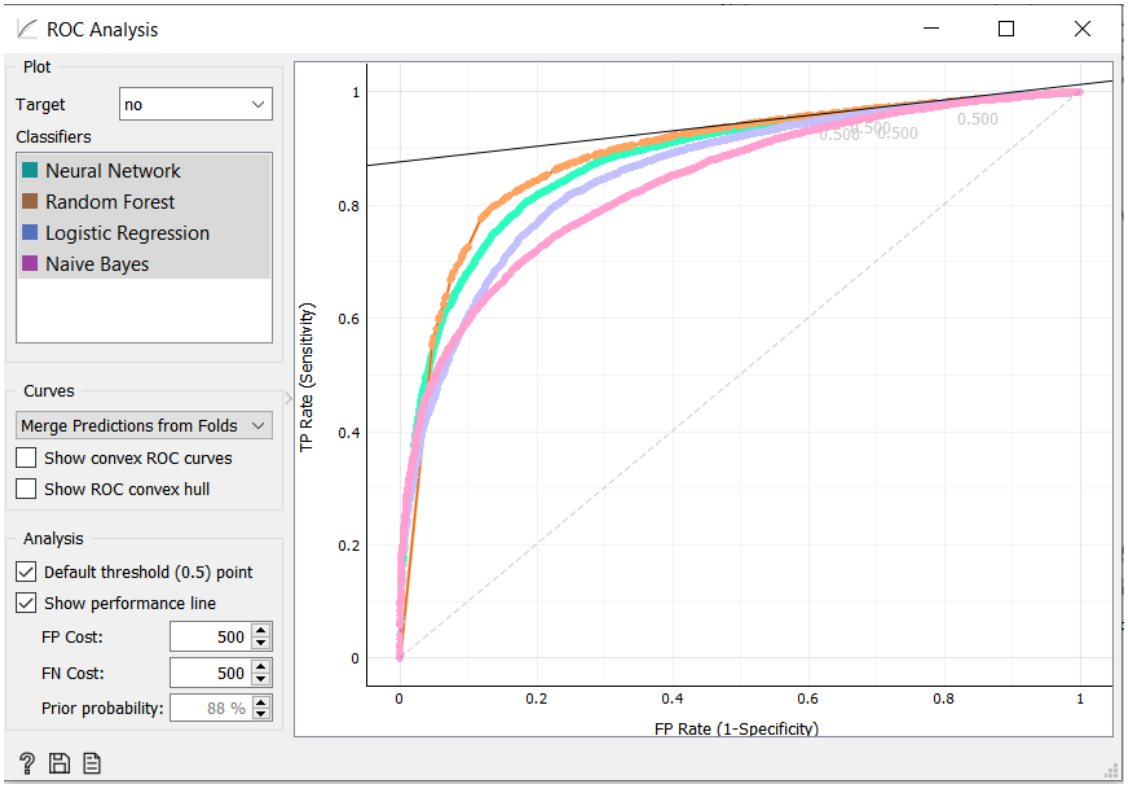


Gráfico 26: Análisis ROC 4 modelos

IX. Optimización

La optimización se hará del modelo con mejores resultados anteriormente, Redes neuronales, se cambiarán los parámetros y después se evaluará el impacto que genera en Test and Score y la Matriz de Confusión

Modelo con 2 capas
Este modelo presenta dos capas de 50 neuronas cada una y un número máximo de iteraciones de 300.

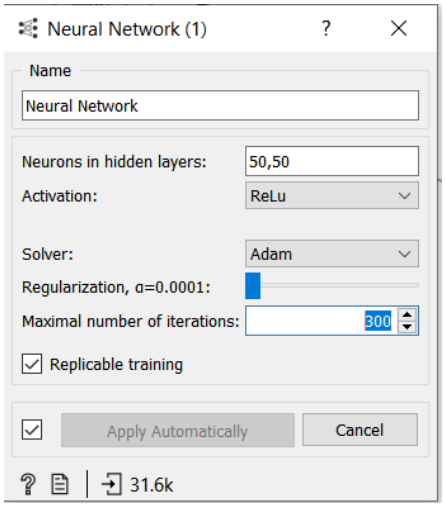


Gráfico 27: Estrategia de optimización 1 Redes Neuronales

Se observa que el valor del F1 para Neural Network es de 0.878

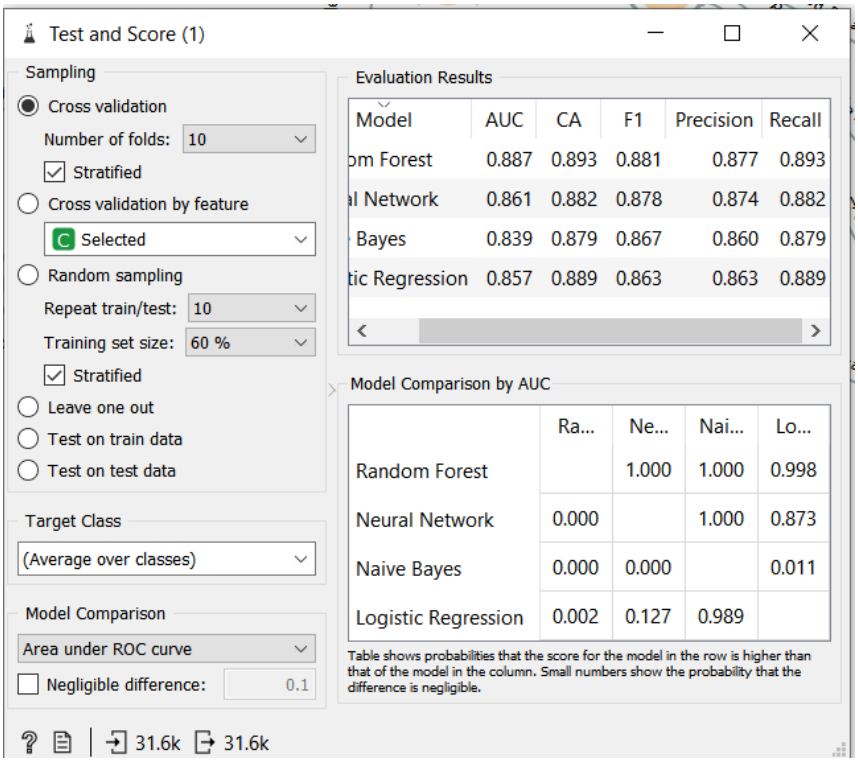


Gráfico 28: Indicadores después de Optimización 1

Y en la matriz de confusión el error tipo I es de 5.5%

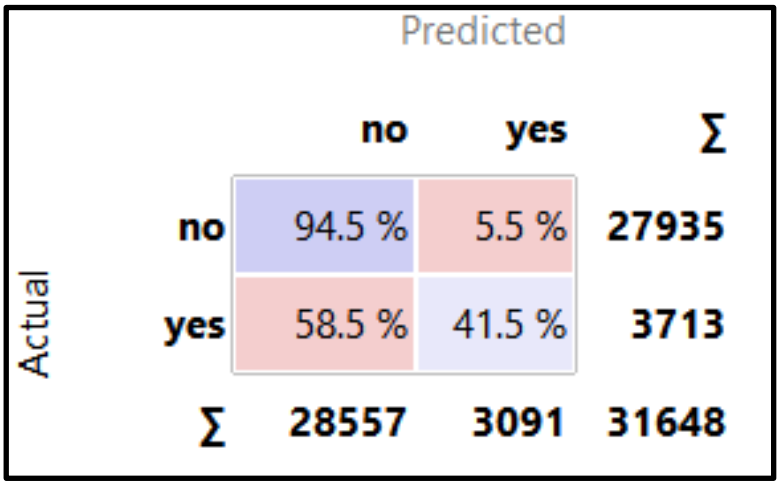


Gráfico 29: Matriz de confusión después de Optimización 1

Modelo con 2 capas y máximo de 2000 iteraciones
Este modelo presenta dos capas de 100 neuronas cada una y un número máximo de iteraciones de 2000.

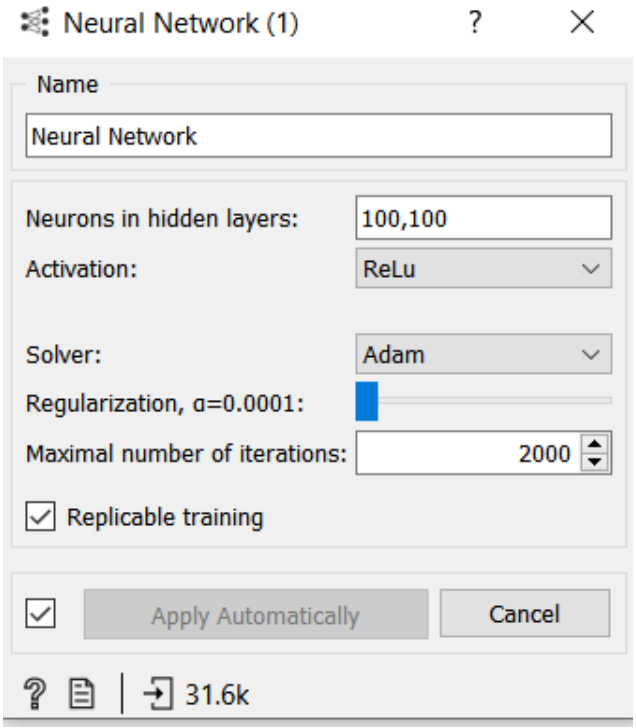


Gráfico 30: Estrategia de optimización 2 Redes Neuronales

Se observa que el valor del F1 para Neural Network es de 0.867

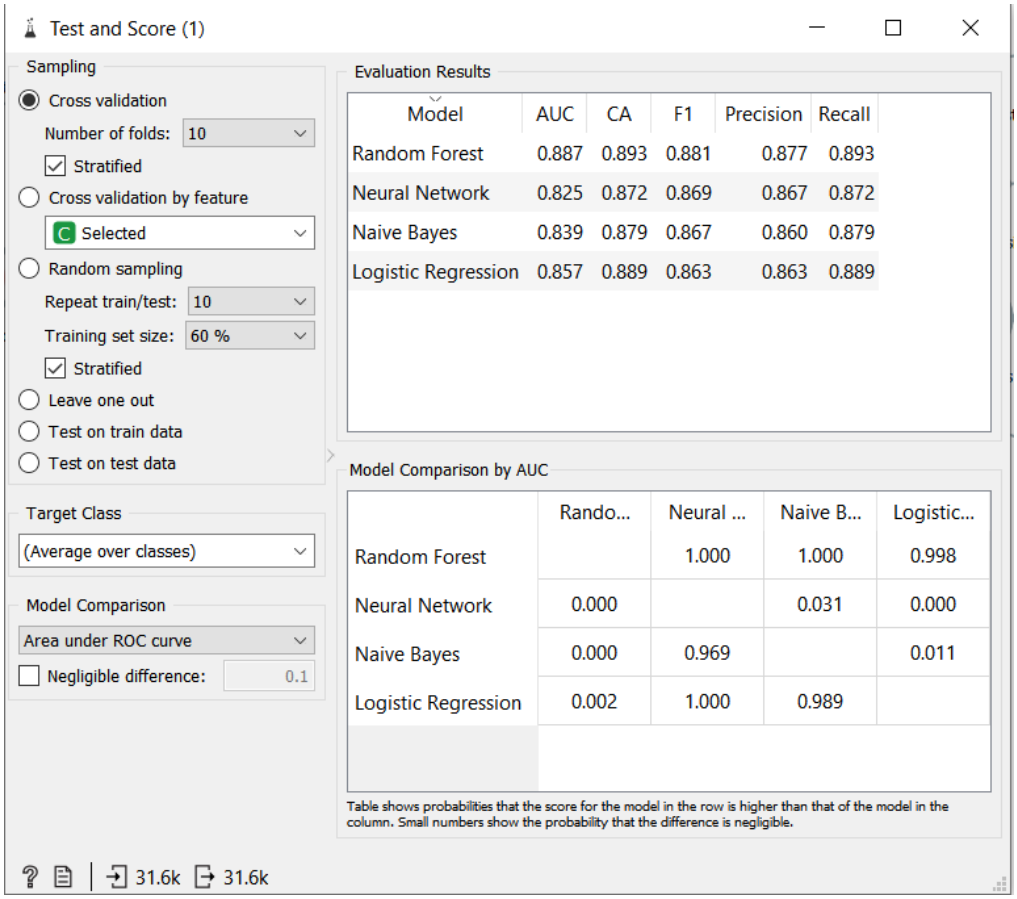


Gráfico 31: Indicadores después de Optimización 2

Y en la matriz de confusión el error tipo I es de 6.7%

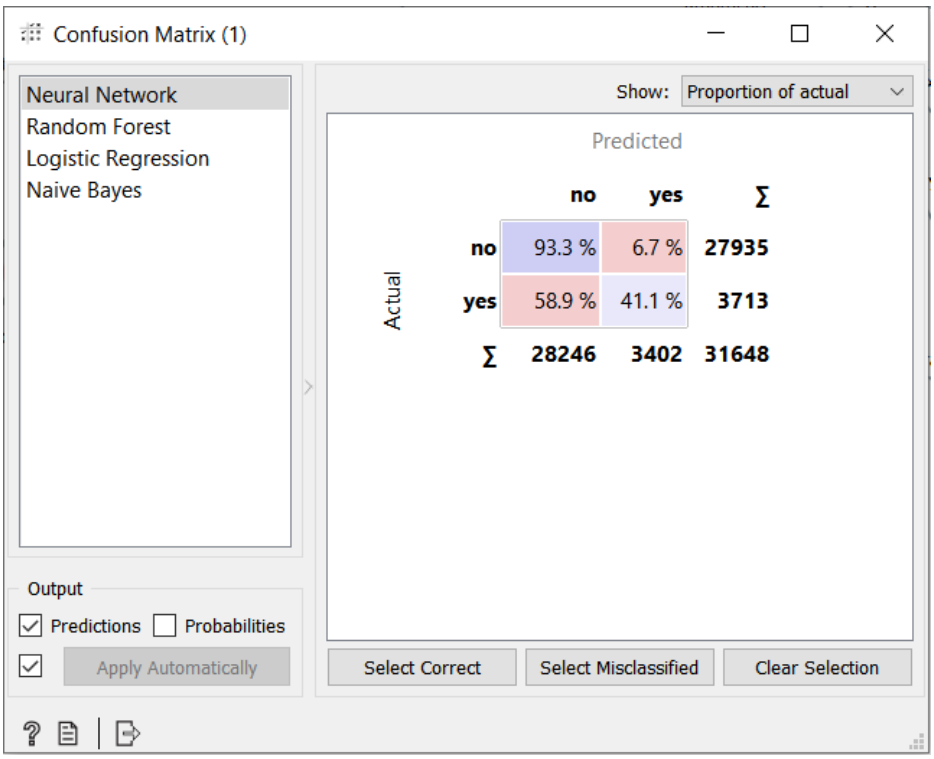


Gráfico 32: Matriz de confusión después de Optimización 2

Se denota que para el caso con menores capas y menor número de iteraciones el resultado que se quiere , el menor error tipo I, es mejor.

X. Conclusiones y recomendaciones

- Se concluye que el mejor algoritmo de predicción son las Redes Neuronales, ya que se obtiene un alto AUC y Recall, y por ende un bajo Error tipo I.
- Si se utilizan menos capas y pocas iteraciones, se mejoran los resultados.

❖ Se recomienda, hacer un análisis más profundo de outliers .

❖ Se recomienda utilizar PCA en las variables numéricas que tengan mucha correlación.

Bibliografía

Hopkins, T. (2018, 27 junio). *Aprende a usar el teléfono para vender*. Entrepreneur. https://www.entrepreneur.com/article/261288