

Realizado por:

Juan Camilo Restrepo Velez
William Leonardo Andrade Collazos
Wilder Valencia Ocampo

PRÁCTICA DE ANÁLISIS PREDICTIVO 10%

Seleccionar una base de datos en <https://www.datos.gov.co/> y realizar un informe con todos los pasos de preparación de datos, utilizar pantallazos para documentar los resultados. Después de realizar todos los pasos, responder en el informe:

Bank Marketing

Los datos están relacionados con campañas de marketing directo de una institución bancaria portuguesa. Las campañas de marketing se basaron en llamadas telefónicas. A menudo, se requería más de un contacto con el mismo cliente, para poder acceder a si el producto (depósito bancario a plazo) sería ('sí') o no ('no') suscrito.

Información de atributos

Información Bancaria de los clientes

Age - Edad

Job - Trabajo: tipo de trabajo

Marital - Estado civil: estado civil

Education - Educación: Nivel educativo

Default - Incumplimiento: ¿tiene el crédito en mora?

Housing - Vivienda: ¿tiene un préstamo de vivienda?

Loan - Préstamo: ¿tiene préstamo personal?

Relacionado con la última llamada de la actual campaña

Contact - Contacto: tipo de comunicación

Month - Mes: último mes de contacto del año

DayofWeek - Día de la semana: último día de contacto de la semana

Duration - Duración: duración del último contacto, en segundos (numérico). Nota importante: este atributo afecta en gran medida al objetivo de salida (por ejemplo, si la duración = 0, entonces y = "no"). Sin embargo, no se conoce la duración antes de una llamada se realiza. Además, después del final de la llamada se conoce obviamente y. Por

lo tanto, esta entrada sólo debe incluirse a efectos de referencia y debe descartarse si se pretende tener un modelo predictivo realista.

Otros

Campaign - Campaña: número de contactos realizados durante esta campaña y para este cliente

Pdays - pDías: número de días que pasaron después de que el cliente fue contactado por última vez en una campaña anterior. Nota, 999 significa que el cliente no fue contactado anteriormente

Previous - Anterior: número de contactos realizados antes de esta campaña y para este cliente

Poutcome: resultado de la anterior campaña de marketing

Atributos del contexto social y económico

Emp.var.rate - Tasa de variación del empleo - indicador trimestral

Cons.price.idx: Índice de Precios al Consumidor - Indicador mensual; el Índice de Precios al Consumidor o IPC mide los cambios en los precios pagados por los consumidores por una cesta de bienes y servicios cada mes.

Cons.conf.idx: Índice de confianza del consumidor - Indicador mensual; En Portugal, el índice de confianza del consumidor se basa en entrevistas con los consumidores sobre sus percepciones de la situación económica actual y futura del país y sus tendencias de compra. Se estima utilizando la diferencia entre la proporción de respuestas de evaluación positivas y las respuestas de evaluación negativas, pero no incluye la proporción de respuestas neutras

Euribor3m: euribor 3 meses - Euribor es la abreviatura de Euro Interbank Offered Rate. es un índice de referencia publicado diariamente que indica el tipo de interés promedio al que un gran número de bancos europeos dicen concederse préstamos a corto plazo entre ellos para prestárselo a terceros.

Nr.employed - Número de empleados: Número de empleados - Indicador trimestral; Número de personas empleadas para el trimestre.

Variable objetivo

y - ¿el cliente ha suscrito un depósito a plazo?

*Tomado de <https://www.kaggle.com/henriqueyamahata/bank-marketing>

PREPARACIÓN DE DATOS

1. Cuáles son las variable predictoras y la variable objetivo?

Predictoras

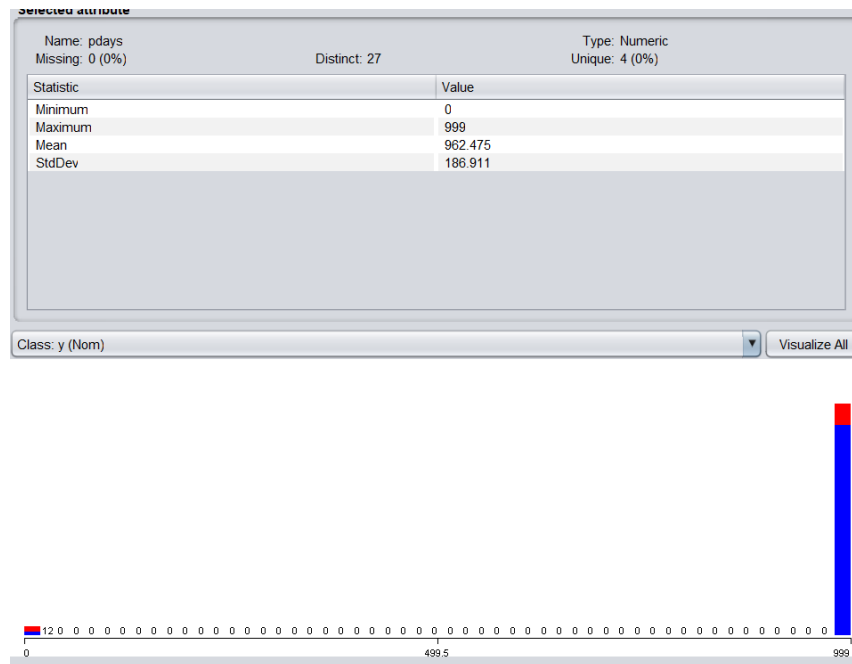
- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan
- Contact
- Month
- DayofWeek
- Duration
- Campaign
- Pdays
- Previous
- Poutcome
- Emp.var.rate
- Cons.price.idx
- Cons.conf.idx
- Euribor3m
- Nr.employed

Variable Objetivo

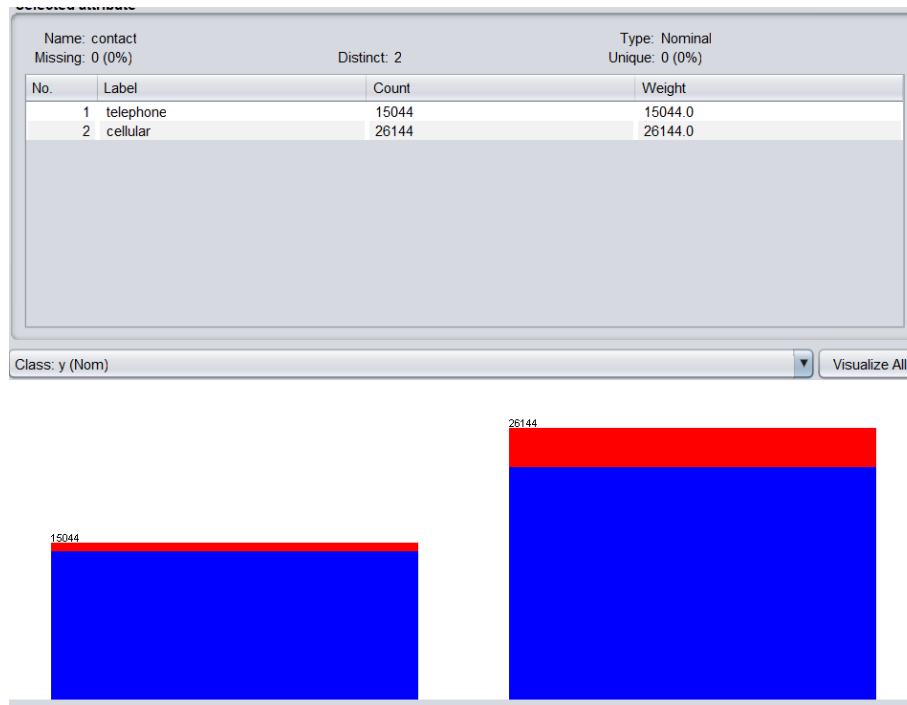
- y

2. Cuáles variables son irrelevantes y/o redundantes?

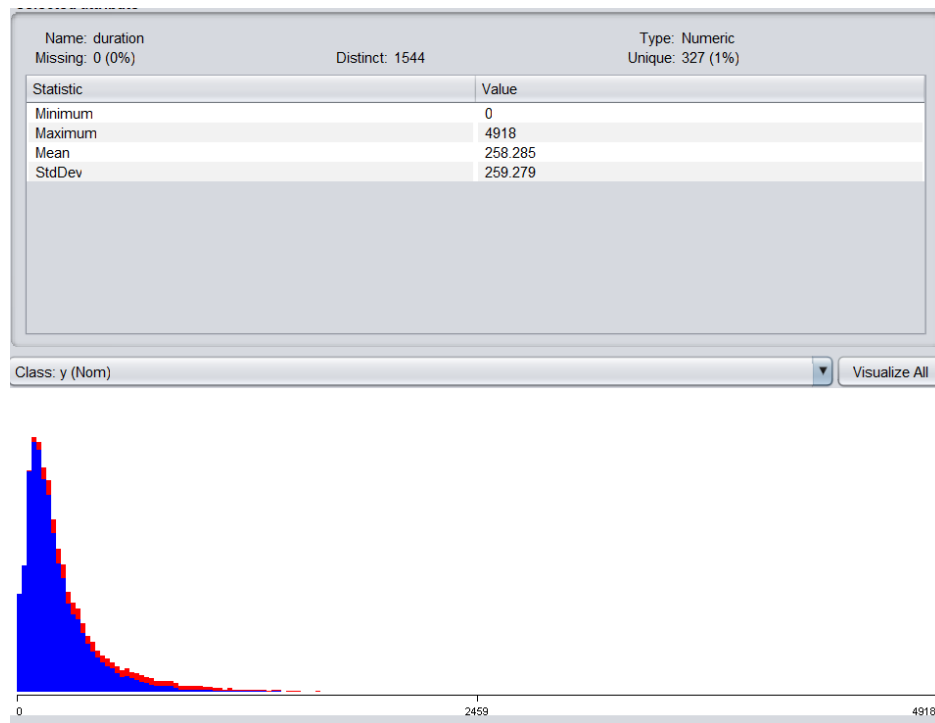
La variable **pdays** resulta irrelevante desde reglas del negocio,



La variable **contact** pues no representa información que se contacte por teléfono o celular



La variable **duration** resulta redundante desde reglas del negocio, porque si la duración es igual 0, entonces la variable objetivo es "no".



Paso extra -> Estadística descriptiva (Paso 3)



3. Cuáles son los datos atípicos?

No hay presencia de atípicos. Esto se debe en gran medida a que los datos ya tenían una alta calidad, por su procedencia de una institución como lo es un banco.

4. Cuáles son los datos nulos?

Varias variables contaban con la categoría “unknown”, por lo cual desde la sabana de datos se procedió a pasar esta categoría a nula.

Variables con Missing							
Name: job Missing: 330 (1%) Distinct: 11 Type: Nominal Unique: 0 (0%)				Name: education Missing: 1731 (4%) Distinct: 7 Type: Nominal Unique: 0 (0%)			
No.	Label	Count	Weight	No.	Label	Count	Weight
1	housemaid	1060	1060.0	1	basic.4y	4176	4176.0
2	services	3969	3969.0	2	high.school	9515	9515.0
3	admin.	10422	10422.0	3	basic.6y	2292	2292.0
4	blue-collar	9254	9254.0	4	basic.9y	6045	6045.0
5	technician	6743	6743.0	5	professional.c...	5243	5243.0
6	retired	1720	1720.0	6	university.degr...	12168	12168.0
7	management	2924	2924.0	7	illiterate	18	18.0
8	unemployed	1014	1014.0				
9	self-employed	1421	1421.0				
10	entrepreneur	1456	1456.0				
11	student	875	875.0				

Name: marital Missing: 80 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)				Name: default Missing: 8597 (21%) Distinct: 2 Type: Nominal Unique: 0 (0%)			
No.	Label	Count	Weight	No.	Label	Count	Weight
1	married	24928	24928.0	1	no	32588	32588.0
2	single	11568	11568.0	2	yes	3	3.0
3	divorced	4612	4612.0				

Name: loan Missing: 990 (2%) Distinct: 2 Type: Nominal Unique: 0 (0%)							
No.	Label	Count	Weight				
1	no	33950	33950.0				
2	yes	6248	6248.0				

El valor de la imputación

Name: job Missing: 0 (0%) Distinct: 11 Type: Nominal Unique: 0 (0%)				Name: education Missing: 0 (0%) Distinct: 7 Type: Nominal Unique: 0 (0%)			
No.	Label	Count	Weight	No.	Label	Count	Weight
1	housemaid	1060	1060.0	1	basic.4y	4176	4176.0
2	services	3969	3969.0	2	high.school	9515	9515.0
3	admin.	10752	10752.0	3	basic.6y	2292	2292.0
4	blue-collar	9254	9254.0	4	basic.9y	6045	6045.0
5	technician	6743	6743.0	5	professional.c...	5243	5243.0
6	retired	1720	1720.0	6	university.degr...	13899	13899.0
7	management	2924	2924.0	7	illiterate	18	18.0
8	unemployed	1014	1014.0				
9	self-employed	1421	1421.0				
10	entrepreneur	1456	1456.0				
11	student	875	875.0				

admin

University

Name: marital Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)				Name: default Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 0 (0%)			
No.	Label	Count	Weight	No.	Label	Count	Weight
1	married	25008	25008.0	1	no	41185	41185.0
2	single	11568	11568.0	2	yes	3	3.0
3	divorced	4612	4612.0				
married				no			
Name: loan Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 0 (0%)							
No.	Label	Count	Weight				
1	no	34940	34940.0				
2	yes	6248	6248.0				
no							

5. Cuáles variables tienen una alta correlación?

marital=married con **marital=single** de $-0.78 \rightarrow$ Al ser variables dummies de una misma variable que tiene más categorías no se puede eliminar ninguna de las dos, además de que no se puede evidenciar una colinealidad entre ellas.

poutcome=nonexistent con **previous** de $-0.88 \rightarrow$ Al existir una alta correlación entre estas variables es posible una redundancia, por lo que se debe eliminar una de las dos variables. En este sentido se analiza la correlación de la variable **previous** con las demás categorías de **poutcome**, en este modo, las correlaciones son de: 0.68 con failure y 0.52 con success, llegando a la conclusión de que se debe eliminar una de las dos variables, **lo cual se decidirá según la que tenga menor correlación con la variable objetivo, que es poutcome con 0.18091**

poutcome=failure con **poutcome=nonexistent** de $-0.85 \rightarrow$ Al ser variables dummies de una misma variable que tiene más categorías no se puede eliminar ninguna de las dos, además de que no se puede evidenciar una colinealidad entre ellas.

cons.price.idx con **emp.var.rate** de 0.78 \rightarrow Al ser variables económicas y de contexto social representan grupos de personas, por lo tanto, presentan redundancia entre ellas. Al aumentar la tasa de empleo, los precios de la canasta básica pueden tender a aumentar debido a la oferta y demanda.

euribor3m con **emp.var.rate** de 0.97 \rightarrow Al ser variables económicas y de contexto social representan grupos de personas, por lo tanto, presentan redundancia entre ellas. Por razones de especulación y de movimientos de libre mercado los intereses que manejan los bancos tienen relación directa en las empresas y en como estas pueden o no contratar.

nr.employed con **emp.var.rate** de 0.91 \rightarrow Al ser variables económicas y de contexto social representan grupos de personas, por lo tanto, presentan redundancia entre ellas. La cantidad de personas contratadas por trimestre influye directamente en la tasa de empleo.

Además, como la misma variable **emp.var.rate** presenta tres correlaciones altas con otras esta es la que se decide eliminar. **También se tiene en cuenta su correlación con la variable objetivo que es de 0.29833**

nr.employed con **euribor3m** de 0.95 \rightarrow Al ser variables económicas y de contexto social representan grupos de personas, por lo tanto, presentan redundancia entre ellas. En este sentido, se relacionan entre la capacidad empleabilidad (**nr.employed**) y la tasa de interés que reciben de otros bancos (**euribor3m**), entre mayor empleabilidad mayor tasa de interés. **Se decidirá cuál eliminar según la correlación que tengan con la variable objetivo, por lo que se decide eliminar esta última variable al presentar una correlación de 0.30777**

	age	job=housemaid	job=services	job=admin	job=blue-collar	job=technician	job=retired	job=management	job=unemployed	job=self-employed	job=entrepreneur	job=student	marital=married	marital=single	marital=divorced	education=basic,4y	education=high_school	education=basic,6y	education=basic,9y	education=professional,course	education=university,degree	education=illiterate	default=yes	housing=yes	loan=yes	month=may	month=jun	month=jul	month=aug	month=oct	month=nov	month=dec	month=mar	month=apr	month=sep	day_of_week=mon	day_of_week=tue	day_of_week=wed	day_of_week=thu	day_of_week=fri	campaign	previous	poutcome=nonexistent	poutcome=failure	poutcome=success	emp,var,rate	cons,price,idx	cons,conf,idx	eurbor3m	nr,employed	
age		0.09	-0.07	-0.09	-0.02	-0.06	0.44	0.06	0	0	0.03	-0.2	0.27	-0.41	0.17	0.24	-0.11	0.01	-0.04	0	-0.04	0.02	0	0	-0.01	-0.07	-0.01	-0.04	0.07	0.05	0.03	0.05	0.01	0.01	0.04	0.02	0.02	0.02	-0.02	-0.02	0.01	0	0.02	0	0.02	0	0.13	0.01	-0.02		
job=housemaid	0.09		-0.05	-0.1	-0.09	-0.07	-0.03	-0.04	-0.03	-0.03	-0.03	-0.02	0.04	-0.06	0.02	0.04	-0.19	-0.03	-0.03	0	-0.03	0	0	0	0	-0.02	0	0.02	0	-0.01	0.01	-0.01	0	0	0	0	0.01	0	-0.01	0	-0.01	0	0.04	0.03	0.04	0.04	0.03				
job=services	-0.07	-0.05		-0.19	-0.18	-0.14	-0.07	-0.09	-0.05	-0.06	-0.06	-0.05	-0.02	0	0.02	-0.07	0.34	0	-0.05	-0.07	-0.18	-0.01	0	0	0	0.06	0.01	0.02	-0.07	-0.02	-0.02	-0.02	-0.02	0.01	-0.02	0	0	0	0	0	0	-0.01	0.01	-0.03	0.02	0.03	-0.06	0.01	0.02		
job=admin	-0.09	-0.1	-0.19		-0.32	-0.26	-0.12	-0.16	-0.09	-0.11	-0.11	-0.09	-0.12	0.11	0.02	-0.16	-0.05	-0.16	-0.18	0.3	-0.01	-0.01	0.01	0.02	-0.05	-0.01	0	0.08	0.01	0	0	0.01	-0.01	-0.01	0.01	0	0	0	0.01	0.01	0.02	0.02	-0.04	0.04	-0.02	-0.02					
job=blue-collar	-0.02	-0.09	-0.18	-0.32		-0.24	-0.1	-0.15	-0.09	-0.1	-0.1	-0.08	0.13	-0.1	-0.06	0.27	-0.17	0.23	0.37	-0.13	-0.32	0.01	0	-0.01	-0.01	0.14	0.03	0.03	-0.13	-0.05	-0.06	-0.03	-0.04	-0.01	-0.05	-0.01	-0.01	0.02	-0.01	0	-0.05	0.04	-0.01	0.05	0.06						
job=technician	-0.06	-0.07	-0.14	-0.26	-0.24		-0.11	-0.12	-0.07	-0.08	-0.08	-0.07	-0.06	0.06	0	-0.14	-0.11	-0.08	-0.11	0.48	-0.04	-0.01	0.01	0.01	-0.01	-0.06	-0.03	0.02	0.14	0	-0.01	-0.01	-0.02	0	0	0	-0.01	0	-0.02	0.02	-0.02	-0.01	0.05	-0.01	0.05	0.05					
job=retired	0.44	-0.03	-0.07	-0.12	-0.11	-0.09		-0.06	-0.03	-0.04	-0.04	-0.03	0.06	-0.11	0.06	0.17	-0.03	-0.01	-0.04	0.01	-0.05	0.01	0	0	-0.01	-0.06	-0.02	-0.01	0.03	0.09	-0.01	0.05	0.04	0.02	0.06	0	0.01	0	-0.01	0	-0.01	0.07	-0.05	0.02	0.07	-0.1	-0.05	0.09	-0.1	-0.13	
job=management	0.06	-0.04	-0.09	-0.16	-0.15	-0.12	-0.06		-0.04	-0.05	-0.04	-0.05	-0.07	0	-0.06	-0.08	-0.03	0.24	-0.01	-0.08	-0.01	-0.01	0	-0.01	0	-0.01	-0.01	-0.03	-0.02	0	0.09	0	0	0	0	0	-0.01	0	-0.01	-0.01	0.01	0	-0.02	-0.03	0	0	0	0			
job=unemployed	0	-0.03	-0.05	-0.09	-0.09	-0.07	-0.03	-0.04		-0.03	-0.04	-0.03	0.01	-0.01	0.01	0	0.01	-0.02	0.02	0	-0.02	0	0.02	0.01	0	-0.02	0.02	0	-0.01	0.01	0.03	0.01	-0.01	-0.02	-0.01	-0.01	0	-0.01	-0.01	-0.01	-0.01	0.02	-0.02	0	0	0.02	0.01	-0.01	-0.01	-0.13	
job=self-employed	0	-0.03	-0.06	-0.11	-0.1	-0.08	-0.04	-0.05	-0.03		-0.04	-0.03	0.01	-0.01	-0.01	-0.02	-0.07	-0.03	0	-0.01	0.09	0.02	0	-0.01	-0.02	0.01	-0.01	0	0	0.03	0	0.01	0	0	0.01	-0.01	-0.01	-0.01	0	0.01	-0.01	0.01	0	-0.01	0	0.01	0.01	0.01			
job=entrepreneur	0.03	-0.03	-0.06	-0.11	-0.1	-0.08	-0.04	-0.05	-0.03	-0.04		-0.03	0.05	-0.06	0.01	0	-0.03	-0.01	0	-0.02	0.05	0.01	0	-0.01	0	-0.01	0	0.01	-0.05	-0.01	0.05	-0.01	-0.02	0.01	-0.01	0.01	-0.01	0	0.01	0	-0.01	0	-0.01	0	-0.02	0.01	0.01	0.01	0.02	0.02	0.02
job=student	-0.2	-0.02	-0.05	-0.09	-0.08	-0.07	-0.03	-0.04	-0.02	-0.03	-0.03		-0.17	-0.22	-0.05	-0.03	0.06	-0.03	-0.01	0	0.01	0	0	0	-0.02	-0.01	-0.02	-0.01	-0.05	-0.02	0.04	0.04	0.03	0.05	0	0	0	0.01	0	-0.02	0.1	-0.08	0.04	0.08	-0.14	-0.06	0.01	-0.15	-0.17		
marital=married	0.27	0.04	-0.02	-0.12	0.13	-0.06	0.06	0.06	0.01	0.01	0.05	-0.17		-0.78	-0.44	0.12	-0.11	0.08	0.07	0	-0.1	0.01	0.01	-0.01	0	0.02	0.02	0.05	0.04	-0.01	0.01	0	-0.04	-0.02	-0.01	0.01	0.01	0	-0.01	0.01	0	-0.04	0.09	0.04	0.06	0.09	0.08				
marital=single	-0.41	-0.06	0	0.11	-0.1	0.06	-0.11	-0.06	-0.07	-0.01	-0.06	0.22	-0.78		-0.22	-0.13	0.06	-0.01	-0.06	0	-0.02	-0.01	0.11	-0.01	0	-0.02	-0.02	0.04	-0.02	0.02	0.05	0.01	0.01	-0.01	-0.01	-0.01	-0.01	0	-0.01	0.01	0	-0.04	0.05	-0.01	-0.06	-0.11	-0.16				
marital=divorced	0.17	0.02	0.02	0.02	-0.06	0	0.06	0	0.01	-0.01	0.01	-0.05	-0.44	-0.22		0.01	0.02	-0.03	-0.02	0.02	-0.01	0	0	-0.01	-0.01	-0.01	0.01	0.02	-0.03	-0.01	0.02	0	0	0.01	-0.01	0.01	0	-0.01	0.01	0	-0.01	0.02	0.02	-0.02	0.02	0.02					
education=basic,4y	0.24	0.19	-0.07	-0.18	0.27	-0.14	0.17	-0.06	0	-0.02	0	-0.03	0.12	-0.13	0.01	0.1	-0.18	-0.08	-0.14	-0.13	-0.24	-0.01	0	-0.01	0	0.02	0.02	0	-0.03	0.01	-0.03	0	0	0	0	0	0.02	0	-0.01	0.01	0.03	0.05	0.02	0.03	0.01						
education=high_school	-0.11	-0.03	0.34	0.12	-0.17	-0.11	-0.03	-0.08	0.01	-0.07	-0.03	0.06	-0.07	0.06	0.02	-0.18	0.1	-0.13	-0.23	-0.21	-0.39	-0.01	0	-0.01	0	0.04	0	0.03	-0.07	0	-0.01	0	-0.02	0	-0.01	0.01	0	-0.02	0.01	-0.02	0.03	-0.01	-0.02	-0.02	-0.02						
education=basic,6y	0.01	0.01	0	-0.1	0.23	-0.08	-0.01	-0.03	-0.02	-0.03	-0.01	-0.03	0.08	-0.07	-0.03	-0.08	-0.13	0.1	-0.1	-0.09	-0.17	-0.01	0	-0.01	-0.01	0.06	0.01	0.01	-0.06	-0.01	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	0	0.01	0	-0.02	0.02	-0.01	-0.02	0.02	0.03						
education=basic,9y	-0.04	-0.03	-0.05	-0.16	0.37	-0.11	-0.04	-0.07	0.02	0	0	-0.01	0.07	-0.06	-0.02	-0.14	-0.23	-0.1	-0.16	-0.3	-0.01	0	0	-0.01	0.09	0.01	0.02	-0.1	-0.03	-0.02	-0.01	-0.02	0	-0.03	0	-0.01	0.01	0	-0.03	0.02	-0.03	-0.07	0.02	0.03							
education=professional,course	0	-0.03	-0.07	-0.16	-0.13	0.48	0.01	-0.08	0.01	-0.01	0.02	-0.03	0	-0.01	0.02	-0.13	-0.21	-0.09	-0.16	-0.7	-0.01	0.01	0	-0.03	-0.01	-0.02	0.07	0.01	0	0	0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0	-0.01	0	-0.03	0.02	0.03	0.02	0.03						
education=university,degree	-0.04	-0.06	-0.18	0.3	-0.32	-0.04	-0.05	0.24	-0.02	0.09	0.05	0.04	-0.1	0.11	-0.01	-0.24	-0.39	-0.17	-0.3	-0.27	-0.01	-0.01	0.01	0.01	-0.03	-0.04	0.02	0.05	0.04	0.01	0.04	0	-0.01	0.01	0	-0.01	0.01	0	-0.01	0.03	-0.04	-0.05	-0.08	-0.04							
education=illiterate	0.02	0	-0.01	-0.01	0.01	-0.01	0.01	-0.01	0	0.02	0.01	0	0.01	-0.01	0	-0.01	-0.01	-0.01	-0.01	-0.01	0.1	-0.01	0	0	-0.01	-0.01	0.01	0.01	0	0	0	0	0	0	-0.01	0	0.01	0	0	0	0	-0.01	0	0	0	0					
default=yes	0	0	0	-0.01	0	0.01	0	0	0.02	0	0	0	0.01	-0.01	0	0	0	0	0	0.01	-0.01	0	0	0	0	-0.01	0	0.01	0	0.01	0	0.01	0	0	0	0	0.02	0	0	0.01	0	0	0	0	0.01	0.01	0.01				
housing=yes	0	0	0	0.01	-0.01	0.01	0	-0.01	0.01	0	0	0	-0.01	0.01	0	-0.01	-0.01	-0.01	0	0.01	-0.01	0	0	0	0	0.04	0	0.02	-0.05	0	0.03	0.01	0.03	0.01	0.01	-0.01	0.02	-0.03	0.02	0.01	-0.06	-0.08	-0.03	-0.06	-0.05						
loan=yes	-0.01	0	0	0.02	-0.01	-0.01	-0.01	0	-0.01	0	-0.01	0	0	0	-0.01	0	0	-0.01	-0.01	0	0.01	0	0	0	0.04	0	-0.01	0.02	0	-0.01	0	0	0	0	0	0.01	-0.01	0	0	0.01	-0.01	0	0	-0.01	0	0	0	0			
month=may	-0.07	-0.02	0.06	-0.05	0.14	-0.06	-0.06	-0.01	-0.02	0	-0.02	0.02	-0.02	-0.02	-0.01	0.02	-0.03	-0.33	-0.33	-0.09	-0.24	-0.05	-0.08	-0.19	-0.08	-0.03	0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03					
month=jun	-0.01	0	0.01	-0.01	0.03	-0.03	-0.02	0	-0.01	0.02	0.01	-0.01	0.02	0.01	-0.01	0.02	-0.02	-0.01	0	-0.05	-0.01	-0.01	0	-0.05	-0.01	-0.27	-0.18	-0.16	-0.05	-0.13	-0.03	-0.04	-0.1	-0.05	0.03	0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07							
month=jul	-0.04	0.02	0.02	0	0.03	-0.02	-0.01	-0.03	0	-0.01	0.01	-0.02	-0.05	0.04	0.02	0	0.03	0.01	0.02	-0.02	-0.04	0.01	0	0	0.02	-0.33	-0.18	0.1	-0.19	-0.06	-0.15	-0.03	-0.05	-0.12	-0.05	0.01	0.02	0.01	0.03	-0.06	0.1	-0.12	0.14	-0.13	-0.05	0.31	0.25	-0.19	0.28	0.3	
month=aug	0.07	0.02	-0.07	0.08	-0.13	0.14	0.03	-0.02	-0.01	0	-0.05	-0.01	0.04	-0.02	-0.03	-0.03	-0.07	-0.06	-0.1	0.07	0.14	0.01	0.01	0.02	0	-0.3	-0.16	-0.19	0.1	-0.06	-0.14	-0.03	-0.05	-0.11	-0.05	-0.01	0.01	0	0.01	-0.02	0.02	-0.05	0.08	-0.09	0	0.18	-0.19	0.45	0.16	0.19	
month=oct	0.05	0	-0.02	0.01	-0.05	0	0.09	0	0.01	0	-0.01	0.05	-0.01	0.02	-0.01	0.01	0.01	-0.03	0.02	0	0	0	0	-0.01	-0.09																										

6. Cuál variable tiene la correlación más alta con la variable objetivo?

nr.employed de 0.35468

7. Cuál variable tiene la correlación más baja con la variable objetivo?

default de 0.00304

```
Ranked attributes:
0.35468  17 nr.employed
0.30777  16 euribor3m
0.29833  13 emp.var.rate
0.23018  11 previous
0.18091  12 poutcome
0.13621  14 cons.price.idx
0.06636  10 campaign
0.05675   8 month
0.05488  15 cons.conf.idx
0.04248   3 marital
0.03611   2 job
0.0304    1 age
0.03035   4 education
0.01146   9 day_of_week
0.01109   6 housing
0.00447   7 loan
0.00304   5 default
```

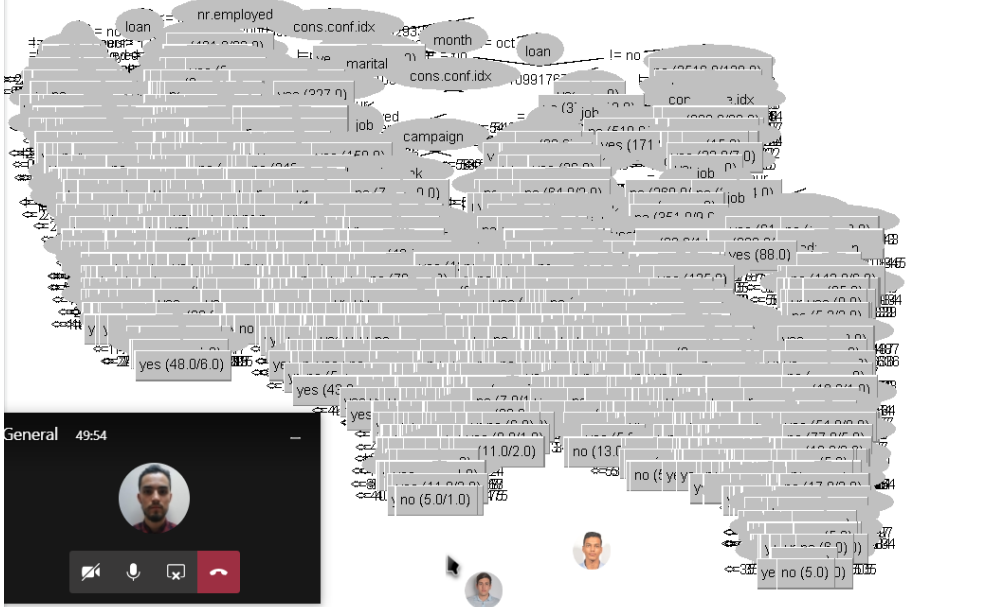
Aunque se considera que las variables con una correlación menor a 0.3 se consideran irrelevantes, solo se decide eliminar las dos últimas variables ya que estas no ayudarían en prácticamente nada en un análisis y no se eliminan todas las que están por debajo del 0.3 para no quedar con tan pocas variables predictoras.

Paso Extra → Balanceo de datos

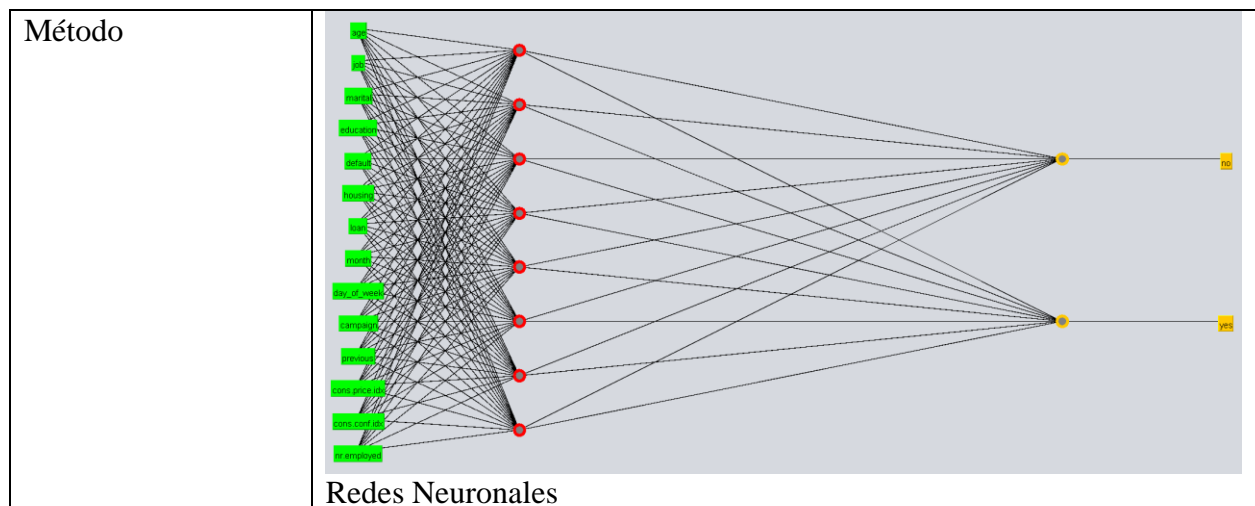
% Balanceo	% Aumento	Categoría	Evidencia
			Nota: el balanceo se realiza con 5 vecinos

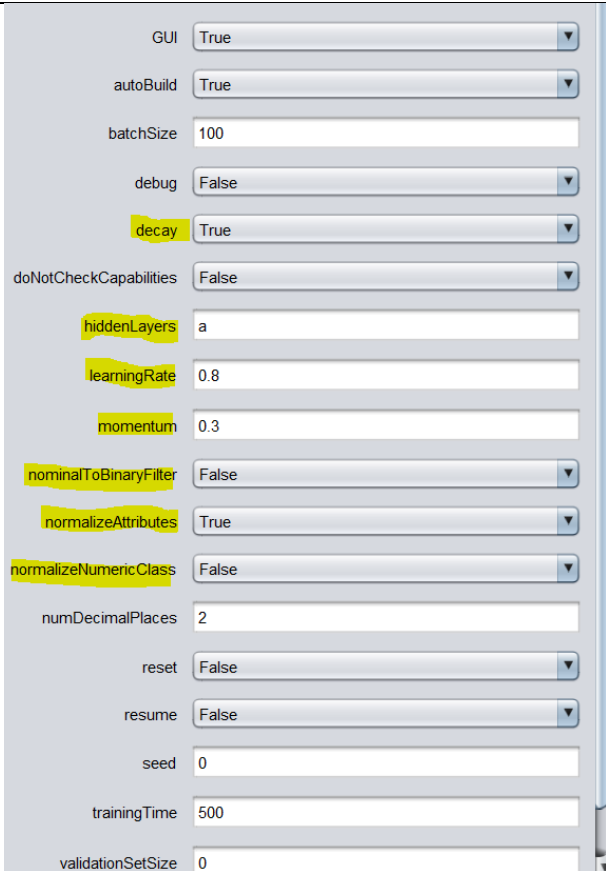
DIVISIÓN DE DATOS

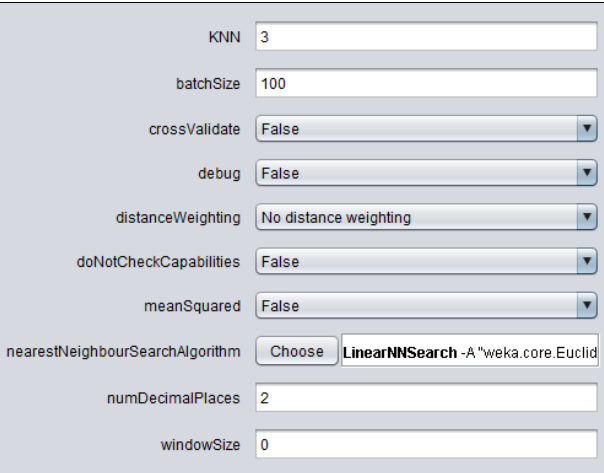
8. Cómo te da un mejor resultado, división 70-30 o validación cruzada? ¿cuál de las dos opciones seleccionas?

Método																																			
Configuración	<table border="1"><tr><td>batchSize</td><td>100</td></tr><tr><td>binarySplits</td><td>True</td></tr><tr><td>collapseTree</td><td>True</td></tr><tr><td>denceFactor</td><td>0.25</td></tr><tr><td>debug</td><td>False</td></tr><tr><td>doNotCheckCapabilities</td><td>False</td></tr><tr><td>doNotMakeSplitPointActualValue</td><td>False</td></tr><tr><td>minNumObj</td><td>5</td></tr><tr><td>numDecimalPlaces</td><td>2</td></tr><tr><td>numFolds</td><td>3</td></tr><tr><td>reducedErrorPruning</td><td>False</td></tr><tr><td>saveInstanceData</td><td>False</td></tr><tr><td>seed</td><td>1</td></tr><tr><td>subtreeRaising</td><td>True</td></tr><tr><td>unpruned</td><td>False</td></tr><tr><td>useLaplace</td><td>False</td></tr><tr><td>useMDLcorrection</td><td>True</td></tr></table>	batchSize	100	binarySplits	True	collapseTree	True	denceFactor	0.25	debug	False	doNotCheckCapabilities	False	doNotMakeSplitPointActualValue	False	minNumObj	5	numDecimalPlaces	2	numFolds	3	reducedErrorPruning	False	saveInstanceData	False	seed	1	subtreeRaising	True	unpruned	False	useLaplace	False	useMDLcorrection	True
batchSize	100																																		
binarySplits	True																																		
collapseTree	True																																		
denceFactor	0.25																																		
debug	False																																		
doNotCheckCapabilities	False																																		
doNotMakeSplitPointActualValue	False																																		
minNumObj	5																																		
numDecimalPlaces	2																																		
numFolds	3																																		
reducedErrorPruning	False																																		
saveInstanceData	False																																		
seed	1																																		
subtreeRaising	True																																		
unpruned	False																																		
useLaplace	False																																		
useMDLcorrection	True																																		

División 70 - 30	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.868	0.109	0.886	0.868	0.877	0.759	0.923	0.903	no
	0.891	0.132	0.873	0.891	0.882	0.759	0.923	0.915	yes
	0.879	0.121	0.879	0.879	0.879	0.759	0.923	0.909	
Validación cruzada	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.875	0.096	0.901	0.875	0.888	0.779	0.929	0.911	no
	0.904	0.125	0.879	0.904	0.891	0.779	0.929	0.921	yes
	0.890	0.110	0.890	0.890	0.890	0.779	0.929	0.916	
Decisión	Se elige la validacion cruzada ya que el area ROC da mejor resultado en esta, ademas por diseño de experimentos la validacion cruzada evita una serie de inconveniente que puede tener la division 70-30.								



Configuración																																					
División 70 - 30	<table><tr><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>0.801</td><td>0.252</td><td>0.756</td><td>0.801</td><td>0.778</td><td>0.549</td><td>0.854</td><td>0.838</td><td>no</td></tr><tr><td>0.748</td><td>0.199</td><td>0.794</td><td>0.748</td><td>0.770</td><td>0.549</td><td>0.854</td><td>0.848</td><td>yes</td></tr><tr><td>0.774</td><td>0.225</td><td>0.775</td><td>0.774</td><td>0.774</td><td>0.549</td><td>0.854</td><td>0.843</td><td></td></tr></table>	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	0.801	0.252	0.756	0.801	0.778	0.549	0.854	0.838	no	0.748	0.199	0.794	0.748	0.770	0.549	0.854	0.848	yes	0.774	0.225	0.775	0.774	0.774	0.549	0.854	0.843	
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																													
0.801	0.252	0.756	0.801	0.778	0.549	0.854	0.838	no																													
0.748	0.199	0.794	0.748	0.770	0.549	0.854	0.848	yes																													
0.774	0.225	0.775	0.774	0.774	0.549	0.854	0.843																														
Validación cruzada	<table><tr><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>0.797</td><td>0.264</td><td>0.751</td><td>0.797</td><td>0.773</td><td>0.533</td><td>0.847</td><td>0.837</td><td>no</td></tr><tr><td>0.736</td><td>0.203</td><td>0.783</td><td>0.736</td><td>0.759</td><td>0.533</td><td>0.847</td><td>0.840</td><td>yes</td></tr><tr><td>0.766</td><td>0.234</td><td>0.767</td><td>0.766</td><td>0.766</td><td>0.533</td><td>0.847</td><td>0.839</td><td></td></tr></table>	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	0.797	0.264	0.751	0.797	0.773	0.533	0.847	0.837	no	0.736	0.203	0.783	0.736	0.759	0.533	0.847	0.840	yes	0.766	0.234	0.767	0.766	0.766	0.533	0.847	0.839	
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																													
0.797	0.264	0.751	0.797	0.773	0.533	0.847	0.837	no																													
0.736	0.203	0.783	0.736	0.759	0.533	0.847	0.840	yes																													
0.766	0.234	0.767	0.766	0.766	0.533	0.847	0.839																														
Decision	Se elige la 70-30 ya que el area ROC da mejor resultado en esta.																																				

Método	KNN
Configuración	

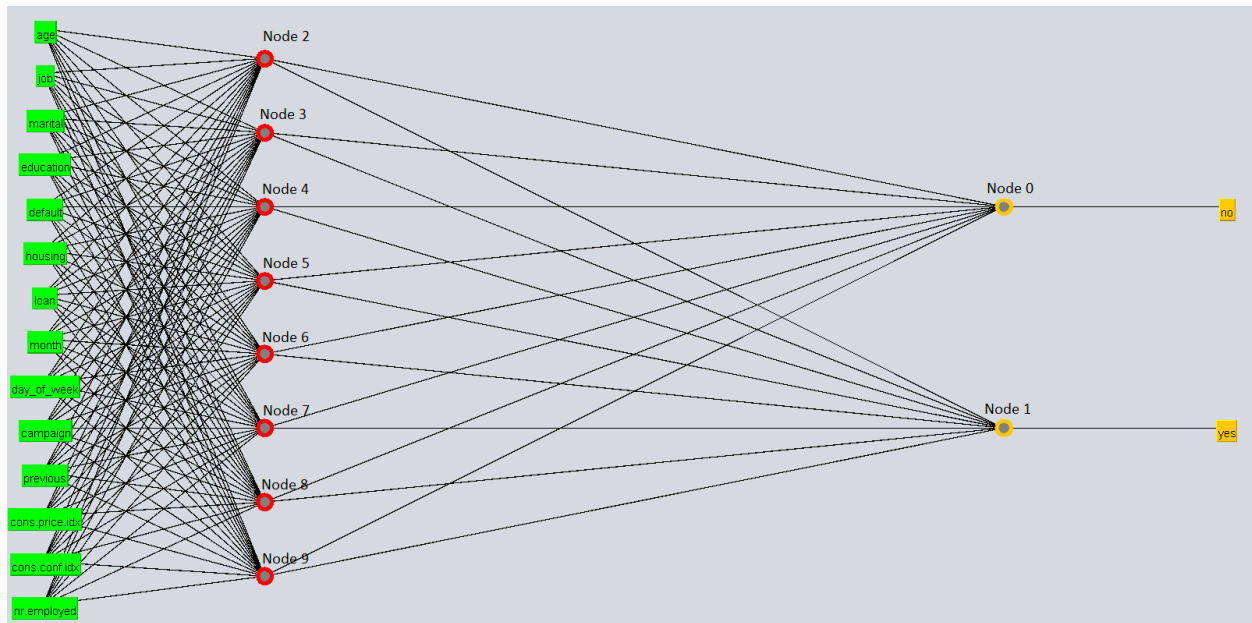
[illegible]

APRENDIZAJE

9. Según el árbol, ¿cuál es la variable más relevante?

nr.employed

10. Cuál es la arquitectura de la red neuronal con los pesos asignados? Puedes usar el gráfico de weka y asigna los pesos después de realizar el aprendizaje



Nodo	Entradas - Pesos
Sigmoid Node 0	Threshold -0.22491626585235294 Node 2 1.8422347436974864 Node 3 3.1771026473760804 Node 4 0.9828442976125559 Node 5 -2.3304972927497287 Node 6 2.632171109641684 Node 7 -2.404456101680537 Node 8 -2.068415830401729 Node 9 2.20282173222468
Sigmoid Node 1	Threshold 0.22491626585235286 Node 2 -1.8422347436974864 Node 3 -3.1771026473760804 Node 4 -0.982844297612556 Node 5 2.330497292749728 Node 6 -2.6321711096416838 Node 7 2.4044561016805375 Node 8 2.068415830401729 Node 9 -2.20282173222468
Sigmoid Node 2	Threshold 0.2708612351446873 Attrib age -1.0911298952601236 Attrib job 7.459082047943109 Attrib marital -0.9263265976884699 Attrib education 0.3452320225485071

	Attrib default 0.7585193276174776 Attrib housing -0.46920468710373847 Attrib loan -0.6112109569549065 Attrib month -6.292998633945483 Attrib day_of_week -0.17881123930145776 Attrib campaign 4.715173687606699 Attrib previous 0.14066445360517352 Attrib cons.price.idx 0.8184443513855352 Attrib cons.conf.idx 2.4664485532377265 Attrib nr.employed -1.1308026718312685
Sigmoid Node 3	Threshold 2.6561485418982578 Attrib age -1.9567296084823376 Attrib job 1.2156216550525931 Attrib marital -0.9359033213915614 Attrib education -0.9768964128055345 Attrib default -2.163349628622533 Attrib housing -0.17320427963063315 Attrib loan 0.08017948575129892 Attrib month 10.344296259682984 Attrib day_of_week -1.1529562924977304 Attrib campaign 9.331364504677325 Attrib previous 2.933509083181121 Attrib cons.price.idx -0.7265603583150595 Attrib cons.conf.idx 3.392360240920952 Attrib nr.employed 4.928550961437286
Sigmoid Node 4	Threshold -3.003027361233724 Attrib age 3.7707195644967113 Attrib job 4.00295792342088 Attrib marital 2.067645753360329 Attrib education 1.1828629808786972 Attrib default 3.3259933860411044 Attrib housing 3.9942866128141863 Attrib loan -0.9384312635931528 Attrib month 0.033473417654837904 Attrib day_of_week 7.616232538417621 Attrib campaign 0.16903462702570993 Attrib previous 1.5610071244645554 Attrib cons.price.idx 6.300793483835978 Attrib cons.conf.idx -1.8567422541870506 Attrib nr.employed 7.692843143589405
Sigmoid Node 5	Threshold -1.3487525518948784 Attrib age -0.7124771243376068 Attrib job 7.51137157044065 Attrib marital 3.4391045958609663 Attrib education -1.6073670865688285

	Attrib default 1.250099280526859 Attrib housing 0.054589685489276014 Attrib loan 3.4424573015853266 Attrib month 0.35018770955346595 Attrib day_of_week 0.31246990752201864 Attrib campaign 0.9745020983635744 Attrib previous 1.9920987074527583 Attrib cons.price.idx -3.2637862189318465 Attrib cons.conf.idx -2.693903570570824 Attrib nr.employed -1.0616293283029907
Sigmoid Node 6	Threshold 1.6351227833713384 Attrib age -5.717510764039165 Attrib job -2.7731941288268103 Attrib marital -0.5838892778235024 Attrib education -0.10233084268527282 Attrib default -1.5912706004119963 Attrib housing 1.387218559334451 Attrib loan 10.547905157500905 Attrib month -3.621052363442202 Attrib day_of_week -1.4467228746597376 Attrib campaign 0.6522334300058131 Attrib previous 5.770826029727162 Attrib cons.price.idx 0.276521636146194 Attrib cons.conf.idx 0.25541203317504596 Attrib nr.employed 2.569433956066817
Sigmoid Node 7	Threshold -1.0268700896441065 Attrib age 0.5233495192029207 Attrib job -2.1946643138127047 Attrib marital -0.6673245914440195 Attrib education 1.0070824444277864 Attrib default 1.0281099816600296 Attrib housing -0.10580302248439435 Attrib loan -1.1671486674706018 Attrib month 10.765086985893504 Attrib day_of_week -11.832783241854807 Attrib campaign -0.21747568842499576 Attrib previous 1.9335029378998827 Attrib cons.price.idx 10.929112066977646 Attrib cons.conf.idx -8.777113291521204 Attrib nr.employed -18.20167205453048
Sigmoid Node 8	Threshold -1.2672212265476723 Attrib age -0.6333831521005691 Attrib job 0.5692251608744784

	Attrib marital 1.0740066035380338 Attrib education -0.46553755309546313 Attrib default 1.2876609604085745 Attrib housing -0.7198978148995636 Attrib loan 0.5952394451671806 Attrib month -2.101404066327142 Attrib day_of_week -1.3033901924555809 Attrib campaign -2.313996536013069 Attrib previous -0.3501663237420886 Attrib cons.price.idx -6.511889000866612 Attrib cons.conf.idx 11.978317966799477 Attrib nr.employed -12.224920568872825
Sigmoid Node 9	Threshold 5.239469051858437 Attrib age 14.698599050357922 Attrib job 2.1529530962507106 Attrib marital 20.41742163877893 Attrib education -0.35937497308953376 Attrib default -5.157945177298546 Attrib housing 0.25700442856759764 Attrib loan 0.21086565906222832 Attrib month -0.3924320952749085 Attrib day_of_week -0.9681934744620337 Attrib campaign 5.99546481675796 Attrib previous -4.37128177385953 Attrib cons.price.idx 0.5298041878557174 Attrib cons.conf.idx 0.5235735518739753 Attrib nr.employed -1.0010652471041153

11. Con cuántos vecinos te da un mejor aprendizaje Knn?

Con 3 vecinos da un mejor aprendizaje

EVALUACIÓN

12. Con cuál método te da mejor resultado?

Método	Arboles								
Resultado	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.875	0.096	0.901	0.875	0.888	0.779	0.929	0.911	no
	0.904	0.125	0.879	0.904	0.891	0.779	0.929	0.921	yes
	0.890	0.110	0.890	0.890	0.890	0.779	0.929	0.916	
Validación Cruzada									

Método	Red neuronal								
Resultado	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.801	0.252	0.756	0.801	0.778	0.549	0.854	0.838	no
	0.748	0.199	0.794	0.748	0.770	0.549	0.854	0.848	yes
	0.774	0.225	0.775	0.774	0.774	0.549	0.854	0.843	
División 70-30									

Método	KNN								
Resultado	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,867	0,075	0,920	0,867	0,893	0,793	0,934	0,915	no
	0,925	0,133	0,874	0,925	0,899	0,793	0,934	0,918	yes
	0,896	0,104	0,897	0,896	0,896	0,793	0,934	0,917	
Validación Cruzada									

Según los resultados del área ROC se obtienen mejores resultados con el método de Arboles de decisiones y KNN, además de que estos fueron realizados por medio de la división de validación cruzada que cumple con el diseño de experimentos serían los dos métodos con mayor relevancia para la predicción futura.

PREDICCIÓN FUTURA

13. Crea un conjunto de 5 datos futuros y compara las predicciones con los 3 métodos.

Método	Arboles de decisión
Predicción	<pre>=== Predictions on test set === inst# actual predicted error prediction 1 1:? 1:no 0.965 2 1:? 1:no 0.965 3 1:? 1:no 1 4 1:? 1:no 0.937 5 1:? 1:no 0.986 6 1:? 1:no 1</pre>

Método	Red neuronal
Predicción	<pre>=== Predictions on test set === inst# actual predicted error prediction 1 1:? 1:no 0.56 2 1:? 2:yes 0.532 3 1:? 1:no 0.848 4 1:? 1:no 0.592 5 1:? 1:no 0.919 6 1:? 2:yes 0.502</pre>

Método	KNN
Predicción	<pre>=== Predictions on test set === inst# actual predicted error prediction 1 1:? 1:no 0.667 2 1:? 1:no 1 3 1:? 1:no 1 4 1:? 1:no 1 5 1:? 2:yes 0.667 6 1:? 2:yes 0.667</pre>

Como se observa en punto anterior los métodos con mejores resultados respecto al área ROC son Arboles de decisión y KNN. En la evaluación se observa que son estos los que mejores resultados dan a la hora de predecir respecto al índice de confianza que estos arrojan, sin embargo, el método que mejores resultados arroja es el método de Arboles de decisión, pues su índice de confianza que es cerca de 1, e incluso 1.