INFORME

Introducción al problema

El problema al que nos enfrentamos en esta práctica, nos describe el caso de una entidad bancaria que contrata una campaña de marketing en la que se realizaron llamadas telefónicas a clientes de la entidad para determinar si están interesados o no en adquirir un nuevo depósito a plazo. Esto indica que nos encontramos ante un problema de **CLASIFICACIÓN.** El objetivo de la clasificación es predecir si el cliente suscribirá dicho depósito a plazo (variable y). Para realizar este trabajo usaremos dos conjuntos de datos extraídos de una versión modificada del problema Bank Marketing y realizaremos un proceso de aprendizaje automático que nos ayuda a determinar cuál es el mejor modelo para realizar dichas predicciones.

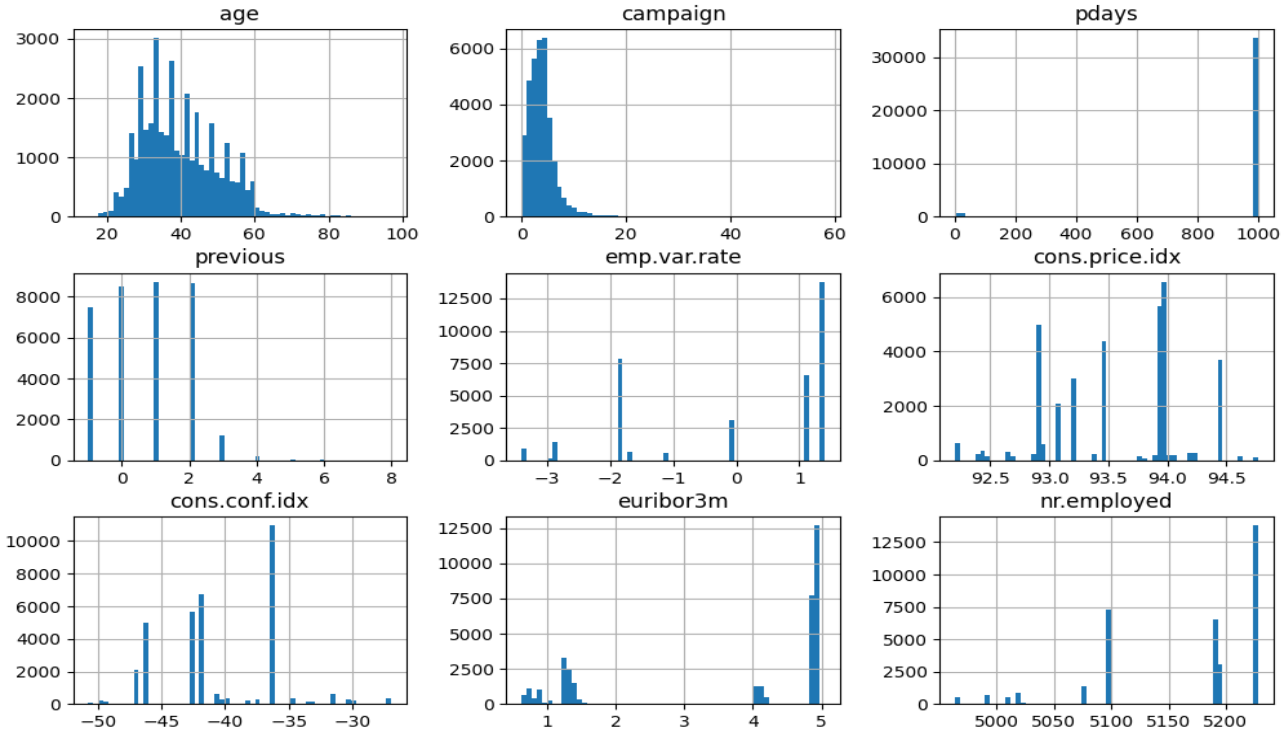
Ejercicio 2: - Realiza un análisis exploratorio de los datos e incluye los resultados en el informe. Comenta las transformaciones que consideras necesarias para adaptar los datos de cara a aplicar con éxito los modelos de aprendizaje automático.

Descripción estadística de las variables numéricas. (Ver figura 1)

- **Age**: distribución ligeramente sesgada en edades entre los 25 y 45 años.
- **Campaign**: su distribución es sesgada en valores entre 2-5, presenta outliers. Los contactos realizados durante esta campaña no aportan información sobre el cliente.
- **Pdays**: esta variable no aporta mucha información ya que todos tienen el mismo valor.
- **Previous**: los valores más altos de estas variables están entre 0 y 2, no aporta información relevante sobre la decisión del cliente.
- **Emp.var.rate**: tasa de variación del empleo, en la mayor parte de los casos alcanza sus valores más altos en 1.40. Este aumento en el empleo puede influir positivamente en la confianza del consumidor y su capacidad para gastar, lo que podría afectar las decisiones de préstamos y la demanda de productos financieros.
- **Cons.price.idx**: índice de precios al consumidor alcanza sus valores mas altos 94.0 para la mayor parte de los datos. en el momento observado, los precios de bienes y servicios para la mayoría de los clientes están en uno de los niveles más altos registrados lo que podría influir en sus decisiones de gasto y ahorro.
- **Cons.conf.idx**: índice de confianza del consumidor, este indicador en la mayor parte de clientes tiene un valor aprox -37. Lo que podría indicar falta de confianza por parte de los consumidores en la situación económica actual y menos propensión a endeudarse.
- **Euribor3m**: Este muestra un aumento considerable ya que pasa de tener un mínimo de 0.63 a un máx. de 5 la mayor parte de los clientes tienen un Euribor superior al 5%, esto indica que estos se enfrentan a tasas de interés más altas en sus préstamos o créditos vinculados a esta tasa, lo que puede tener implicaciones en sus decisiones financieras y capacidad para acceder a préstamos a tasas más bajas.

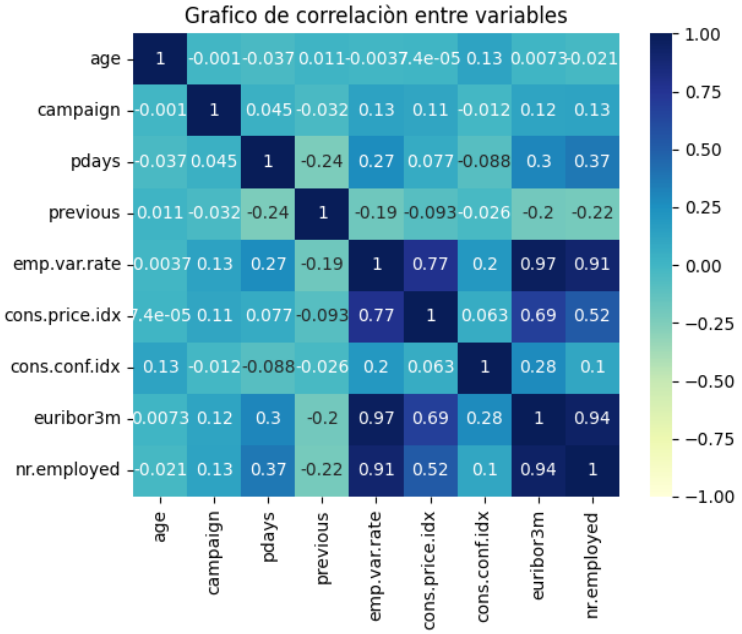
- **Nr.employed**: puede proporcionar información sobre la salud del mercado laboral, ya que muestra un notable aumento. puede tener efectos positivos en la capacidad de pago y en la demanda de préstamos.

FIGURA 1.Ejer.2



Trazamos la correlación entre las características numéricas y graficamos la matriz de correlación. Ver figura 2

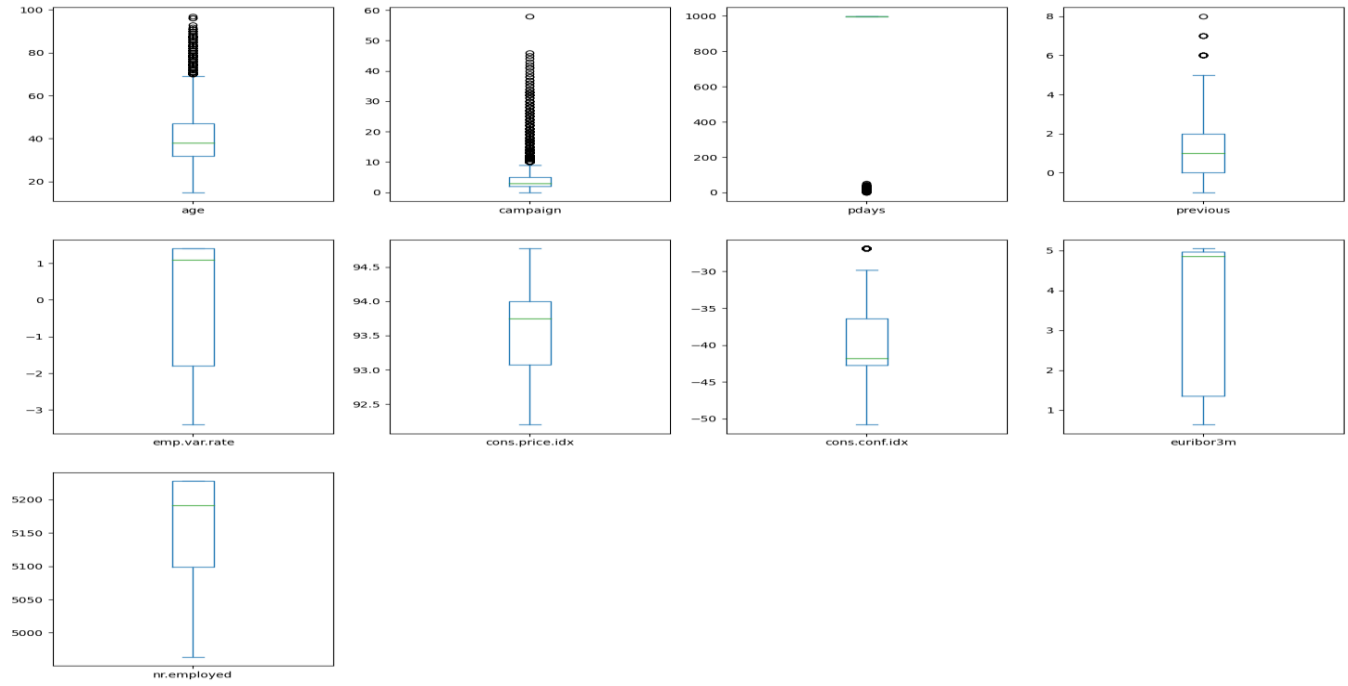
FIGURA 2. Ejer.2



De la figura 2 Gráfico de correlación entre variables numéricas se puede inferir que existe un alto grado de correlación entre las variables de indicadores económicos: emp.var.rate y nr.employed con el euribor3m.

Detectamos los outliers en las variables numéricas, ver figura 3.

FIGURA 3. Ejer.2



En la figura 3 podemos detectar los outliers en las variables edad, campaign y previous, los valores atípicos más relevantes se encuentran en campaign y edad, esto sugiere la presencia de datos que son excepcionales o diferentes del resto de la muestra, lo que puede requerir un análisis más detallado para comprender su impacto en el conjunto de datos y en la toma de decisiones basadas en estos.

Análisis Multivariado De Datos Numéricos

Con estas variables numéricas hemos realizado un análisis multivariado de los datos. Ver figura 4.

En el primer gráfico de la figura 4 tenemos la comparación entre nr.employed y euribor3m este gráfico de dispersión nos muestra una relación lineal entre estas dos variables a medida que aumenta el número de empleos aumenta el euribor3m.

En el segundo gráfico de la figura 4 la comparación entre las variables emp.var.rate y cons.price.idx: la tasa de variación del empleo presenta dispersión entre los datos, entre estas dos variables parece no haber una relación lineal, hay que normalizarlos.

En el tercer gráfico de la figura 4 la comparación entre índice de precios al consumidor y el índice de confianza del consumidor (cons.price.idx y cons.conf.idx) nos indica que no existe una correlación clara donde el cambio en una variable se asocie consistentemente con un cambio en la otra variable. De lo anterior podemos inferir que un índice de Precios al Consumidor (IPC) alto, como 94.0, indica un nivel alto de precios para bienes y servicios, lo que generalmente sugiere que los consumidores están experimentando una inflación relativamente alta.

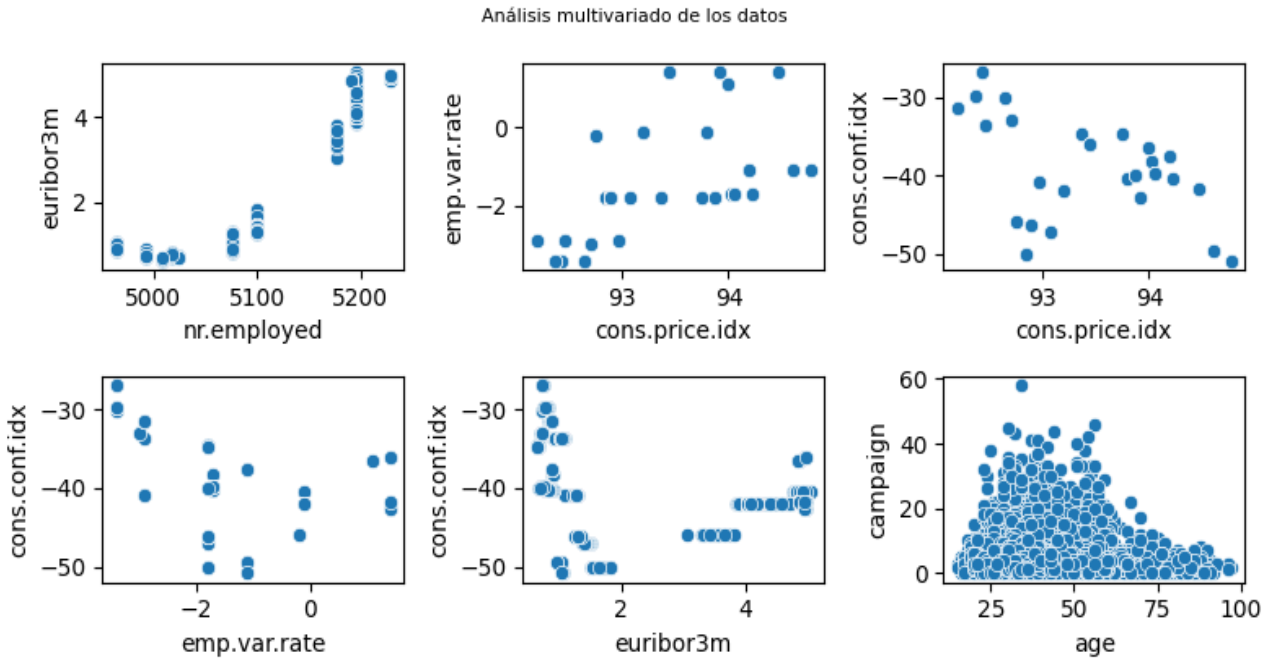
Por otro lado, los valores negativos en el Índice de Confianza del Consumidor (ICC), como -50 o -30, indicarían un pesimismo generalizado entre los consumidores con respecto a la economía, sus finanzas personales y una reticencia por parte de los consumidores para gastar o invertir.

En el cuarto gráfico de la figura 4 la comparación entre la tasa de variación del empleo y el índice de confianza del consumidor (emp.var.rate y cons.conf.idx) se ve una diferencia de valor muy alta y están descompensados. podría implicar que, en términos generales, los cambios en la tasa de variación del empleo no se corresponden directamente o no influyen significativamente en los niveles de confianza del consumidor, al menos de manera lineal o predecible.

En el quinto gráfico de la figura 4 el euribor3m y el índice de confianza del consumidor (euribor3m y cons.conf.idx) no presentan ningún tipo de relación lineal, no podemos anticipar un cambio específico en el índice de confianza del consumidor basado únicamente en las fluctuaciones y aumento del euríbor a 3 meses es decir, los cambios en las tasas de interés a corto plazo (euribor3m) no tienen un impacto lineal o directo en la confianza del consumidor, al menos en términos de su relación observable en los datos.

En el sexto gráfico de la figura 4 edad y campaña (age y campaign) muestra que No existe una tendencia lineal entre estos. el número de contactos realizados durante esta campaña presenta valores atípicos.

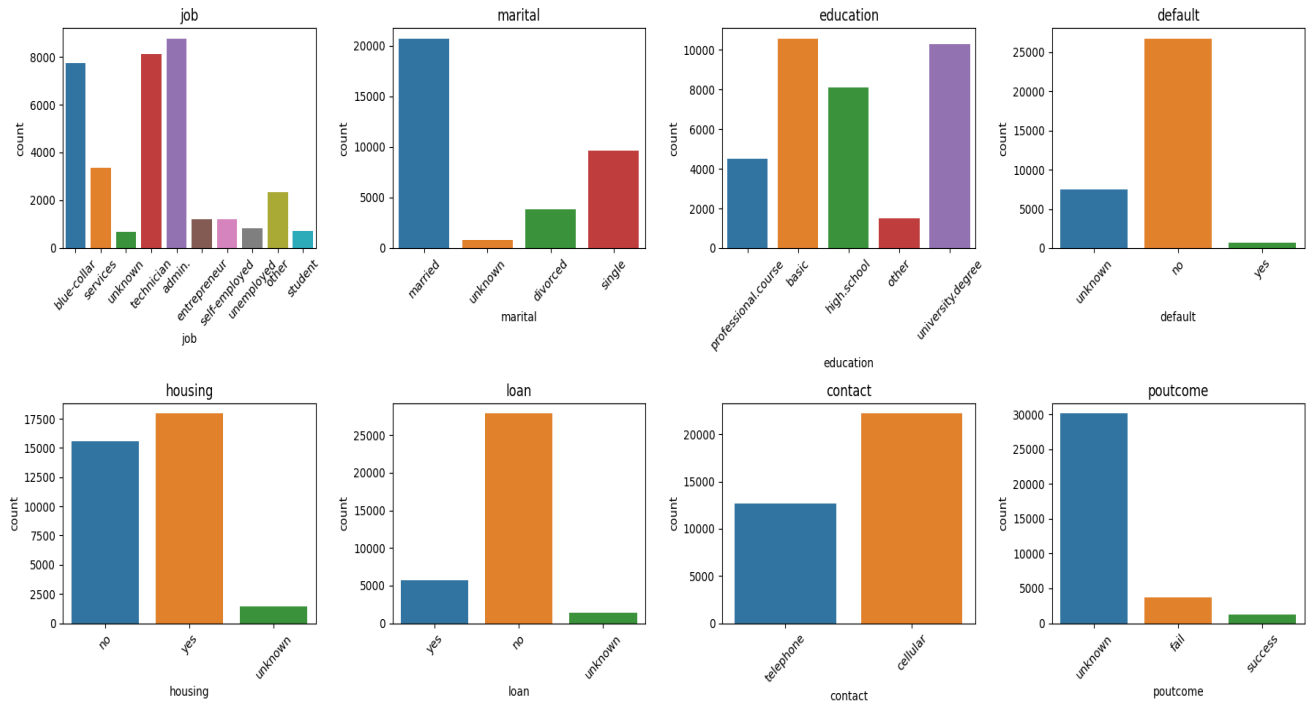
FIGURA 4. Ejer.2



Descripción Y Distribución De Las Variables Categóricas. Ver Figura 5.

- **Marital**: se observa que en esta categoría predominan los casados en segunda posición se encuentran los solteros.
- **Job**: la distribución de este atributo muestra tres clases predominantes, la primera es admin con el mayor número de empleos. La segunda son los empleados técnicos y la tercera es la clase obrera.
- **Education**: la distribución de la variable educación muestra que hay un mayor numero de clientes con educación básica, seguido por los que tienen un título universitario, la categoría que ocupa el tercer puesto es para los que tienen escuela secundaria.
- **Housing**: La distribución de esta variable evidencia que cerca 17.500 clientes tienen un préstamo de vivienda, este atributo podría ser relevante para la decisión de endeudamiento del cliente.
- **Loan**: En la distribución de este atributo vemos que la gran mayoría de los clientes no tienen un préstamo personal al momento de la encuesta.
- **Contact**: Esta variable no aporta mucha información, visto que su distribución indica que la mayor parte de clientes fueron contactados vía celular, pero no nos aporta información relevante sobre el cliente.
- **Poutcome**: la distribución de resultado de la campaña de marketing anterior, nos indica que tuvo muy poco suceso y esta muy desbalanceada visto el gran número de unknown.
- **Default:** La distribución de este atributo sugiere que la mayoría de los clientes analizados no han incumplido con los pagos de sus créditos en el momento en que se recopilaron los datos.

FIGURA 5. Ejer.2



Relación Entre Variables Categóricas Con Grafico Boxplot. Ver figura 6

Usamos el gràfico boxplot para determinar y comparar de manera visual las distribuciones, medidas de tendencia central, dispersión y la presencia de valores atípicos entre las variables “age”, “loan”, “housing” y “campaign”. Además de considerar la edad como un atributo relevante en la toma de decisiones y la demanda de productos financieros.

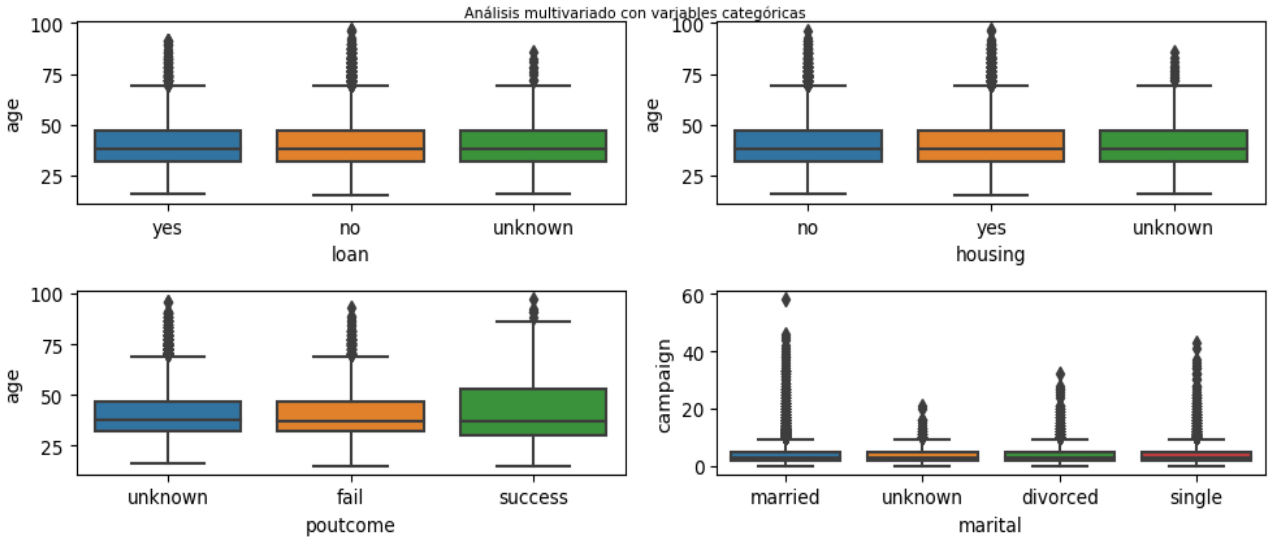
El primer gráfico de la figura 6 nos indica que los clientes que tienen un préstamo personal o de consumo y no lo tienen se encuentran en la fase de edad media de 40 años presentando valores atípicos para edades superiores a 75 años. Estos valores atípicos podrían indicar una presencia inusual de personas de mayor edad en ambos grupos.

El segundo gráfico de la figura 6 las 3 clases de la variable housing (si, no y desconocido) nos indica que los que tienen hipotecas sus edades tienden a concentrarse entre el primer y tercer cuartil (25% al 75%). La presencia de valores atípicos (outliers) en la distribución de edades para ambos grupos ('yes' y 'no housing') a partir de los 75 años sugiere que hay individuos más mayores dentro de la muestra que se alejan considerablemente de la mediana y el rango intercuartílico.

El tercer gráfico de la figura 6 muestra el resultado de la campaña de marketing anterior y la edad. El grafico indica que tuvo más éxito en los rangos de edad entre el primer y tercer cuartil (25% al 75%) para clientes con edades comprendidas entre 30 y 50 años.

El cuarto gráfico de la figura 6 marital y campaign, El boxplot muestra que, en general, para todas las clases de estado civil ('marital'), la cantidad de contactos tiene una mediana de alrededor de 3. Esto sugiere que, en promedio, se han realizado alrededor de 0 y 10 contactos para cada tipo de cliente, independientemente de su estado civil. La presencia de outliers en la variable married sugiere que algunos clientes casados han recibido un número inusualmente alto de contactos durante esta campaña en comparación con la mayoría de los clientes en esa categoría.

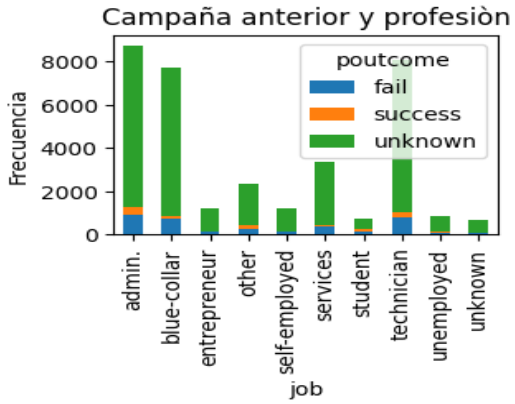
FIGURA 6. Ejer.2



Ahora queremos visualizar la relación entre Poutcome: resultado de la campaña de marketing anterior ('fracaso', 'inexistente', 'éxito') y las variables job, marital y education. Ver figuras 7-14

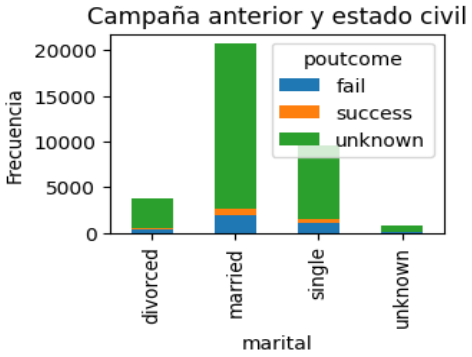
Relación Entre Variables Categóricas- Análisis multivariado

FIGURA 7. Ejer.2



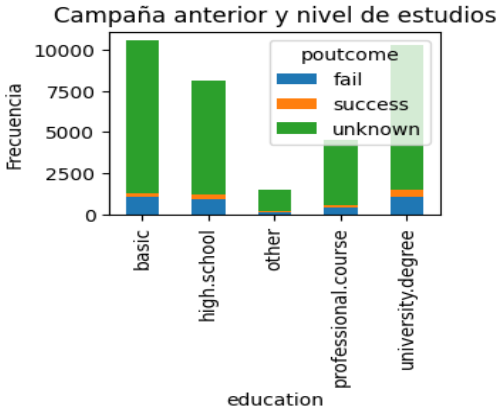
La figura 7 nos muestra la relación entre poutcome y el tipo de profesión, esta indica que la campaña de marketing anterior tuvo éxito en los clientes que tienen profesiones de administradores, técnicos y la clase obrera.

FIGURA 8. Ejer.2



La figura 8 nos muestra la relación entre poutcome y la categoría del estado civil del cliente. Se puede inferir que la campaña anterior tuvo un mayor impacto o éxito en los clientes casados en comparación con otros estados civiles.

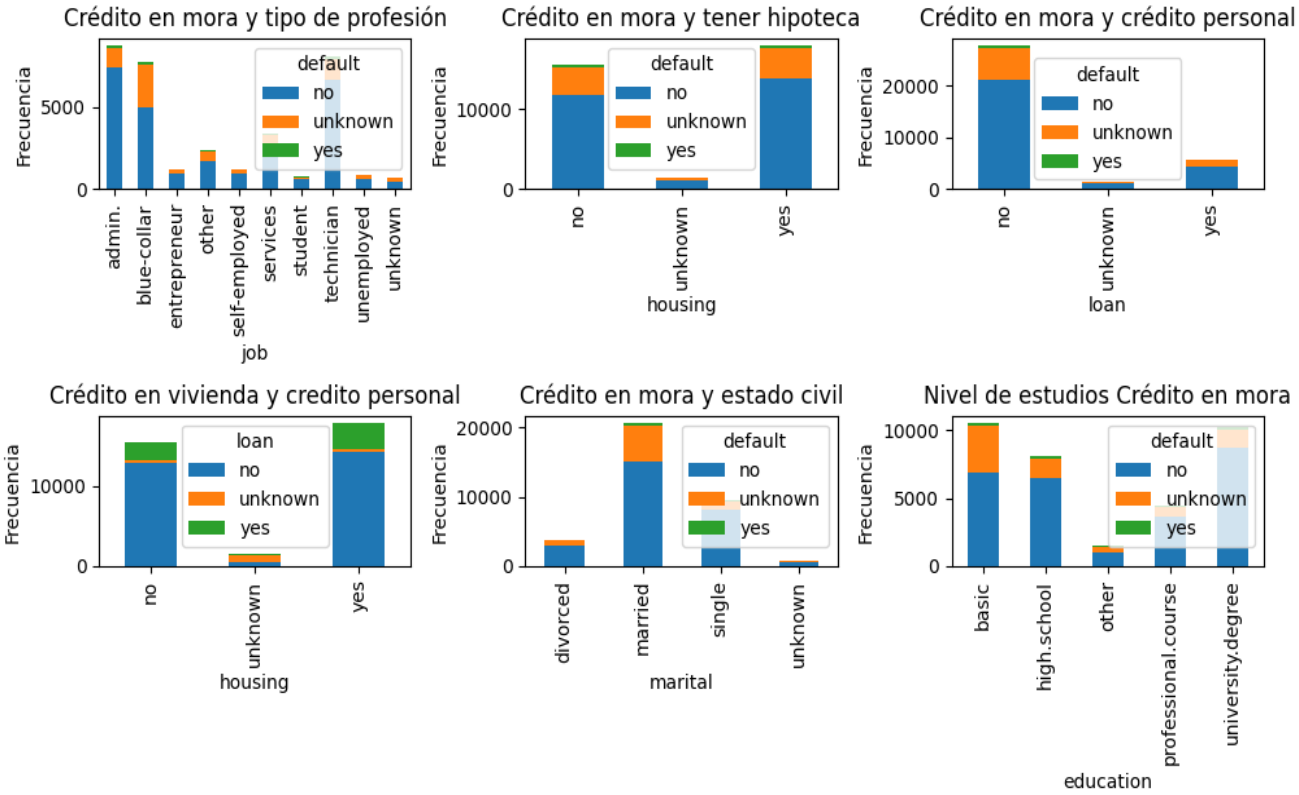
FIGURA 9. Ejer.2



En la figura 9 el análisis se centra en la relación entre el resultado de la campaña anterior ('poutcome') y el nivel de educación de los clientes(‘education’). La campaña anterior fue más exitosa entre los clientes con un título universitario en comparación con otros niveles de educación, incluidos aquellos con educación básica, esto sugiere que el nivel educativo puede ser un factor importante que influye en la efectividad de la campaña.

Ahora el análisis se centra en otra variable relevante “Default”(¿Tiene crédito en mora?) queremos saber qué relación tiene, tener un crédito en mora entre los que tienen un crédito personal o una hipoteca y su estado civil y nivel de estudios. Ver figuras 10

FIGURA 10. Ejer.2



En la figura 10 el gráfico muestra la relación entre ‘default’ y la profesión de los clientes ‘job’, tener hipoteca ‘housing’, tener crédito de consumo ‘loan’, estado civil ‘marital’ y nivel de estudios ‘education’.

Los gráficos evidencian una clara diferencia entre las clases(‘no’, ‘yes’) en default, podemos inferir que un alto número de clases 'no' en relación con 'default' en el gráfico podría indicar una tendencia general de ausencia de morosidad en las distintas profesiones, estado civil, nivel de educación y tipos de crédito de los clientes, destacando así la influencia potencial de la ocupación y nivel de estudio en la capacidad de pago de los clientes y una estabilidad financiera, sin embargo es mas probable que las clases estén muy desbalanceadas vista la diferencia significativa entre las clases

Conclusiones finales del análisis exploratorio de los datos

Del análisis exploratorio de los datos se considera que las variables con potencial para predecir o inferir si un cliente puede contratar un nuevo depósito a plazo podrían ser 'poutcome', 'default', 'marital' y 'age 'pero antes de hacer esta consideración es vital realizar las siguientes transformaciones en el set de datos:

1. Corregir valores atípicos outliers, normalizar todas las variables numéricas y eliminar las variables que no aportan información sobre el cliente.
2. Se considera eliminar las siguientes columnas visto que no aportan mucha información para el presente estudio: Pdays, contact y previous.
3. Transformas las siguientes variables categóricas: Education, Default, Housing y Loan.
4. Transformar en variables dummies las siguientes por no tener una relación entre el orden de estas: Job, Marital y Poutcome.
5. Normalizar todo el nuevo dataset.

Ejercicio 3: considerando los cambios indicados en el ejercicio 2, procedemos a implementar la clase DataTransformer (véase fichero pml.py) para transformar los datos del problema, con esta hemos conseguido realizar todas las transformaciones necesarias para poder utilizar los datos en el análisis además hemos conseguido dos dataframes con los atributos y los targets respectivamente para datos de entrenamiento y un dataframes con los atributos de prueba.

Ejercicio 4: hemos Implementado la función PCA (véase fichero pml.py) que calcula las componentes principales de los datos contenidos en el dataframe data. Esta función nos permite reducir la dimensionalidad de conjuntos de datos. Los componentes principales se ordenan por la cantidad de varianza que explican.

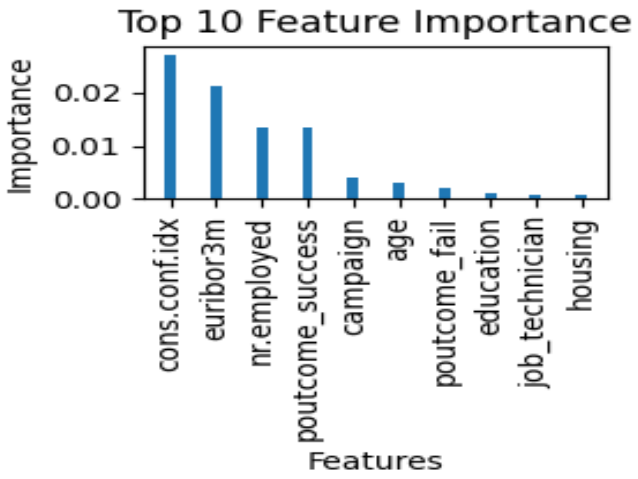
Ejercio 5: hemos construido la función feature_importance (véase fichero pml.py) la cual nos permite calcular la importancia por permutación de cada variable en el dataframe de entrenamiento. Los resultados se pueden visualizar en la siguiente tabla.

Tabla 1 Feature importance

index	Feature	Importance
8	cons.conf.idx	0.02734744136765843
9	euribor3m	0.02127659574468077
10	nr.employed	0.013602130523180822
26	poutcome_success	0.013373042009106273
5	campaign	0.00398041293204654
0	age	0.0030354228114887194
25	poutcome_fail	0.002119068755190301
1	education	0.0011168065061137877
18	job_technician	0.0006586294779644675
3	housing	0.0005727212851864838

El resultado anterior nos muestra que después de permutar 10 veces cada uno de los atributos, usando como estimador un árbol de decisicion (DecisionTreeClassifier) con 7 de profundidad, el atributo que tiene mayor importancia es el índice de confianza del consumidor 'cons.conf.idx' con una importancia media de 0.028, el segundo atributo más importante es 'euribor3m' con 0.02 teniendo menos importancia el atributo 'housing' como indicado en el la figura 11.

FIGURA 11. Ejercicio 5



Ejercicio 6: Implementamos la función train_classifier_single_test (véase fichero pml.py) con la cual entrenamos y evaluamos un clasificador utilizando una única partición.

A continuación, los resultados del score en datos de entrenamiento 0.9073392243495336, y el *score* obtenido sobre los datos de test y 0.8980624224491743 clf es el clasificador entrenado en este caso usamos clf= DecisionTreeClassifier(max_depth=7) aquí no se está teniendo en cuenta que las clases están desequilibradas

Ejercicio 7: Implementa la función train_classifier_nfold_val (véase fichero pml.py) con esta hemos entrenado y evaluado un clasificador utilizando validación cruzada que intenta mantener la misma proporción de etiquetas de clase en cada pliegue que en el conjunto de datos original con nfolds = 5. Es útil cuando tienes clases desequilibradas como nuestro caso, ya que busca preservar esa distribución. Hemos usado como clasificador clf = LogisticRegression(max_iter=1000). Los resultados se pueden visualizar en la tabla 2

Tabla 2 Validaciòn cruzada

score_train	Validacion	Clasificador
0.8992339633447881	0.9000715819613457	LogisticRegression(max_iter=1000
0.8993449547195476	0.900057273768614	LogisticRegression(max_iter=1000
0.8992375702473422	0.8999140893470791	LogisticRegression(max_iter=1000
0.8994165443676845	0.8997709049255441	LogisticRegression(max_iter=1000
0.9000966460249847	0.8969072164948454	LogisticRegression(max_iter=1000

Ejercicio 8: Implementamos la función fit_hyperparams (véase fichero pml.py) que entrena y evalúa un conjunto de clasificadores utilizando validación cruzada, hemos entrenado el modelo con un clasificador LogisticRegression y nos devolvió una lista con el mejor modelo el score promedio obtenido por ese clasificador sobre todos los conjuntos de validación.

Tabla 3 Validaciòn cruzada en conjunto de clasificadores

LogisticRegression	C=10.0, max_iter=1000	Score
LogisticRegression	C=0.0001,max_iter=1000	0.8882334424691759
LogisticRegression	C=0.001,max_iter=1000	0.889951589931296
LogisticRegression	C=0.01,max_iter=1000	0.8992583723426325
LogisticRegression	C=0.1,max_iter=1000	0.899544745285471
LogisticRegression	max_iter=1000	0.899544745285471
LogisticRegression	C=10.0, max_iter=1000	0.899573382169778

Hemos entrenado un segundo modelo con un clasificador RandomForestClassifier(max_depth=5, n_estimators=501) a continuación los resultados de la validación, por lo visto no difieren mucho del modelo de regresión logística.

Tabla 4 Validaciòn cruzada en conjunto de clasificadores Rforest

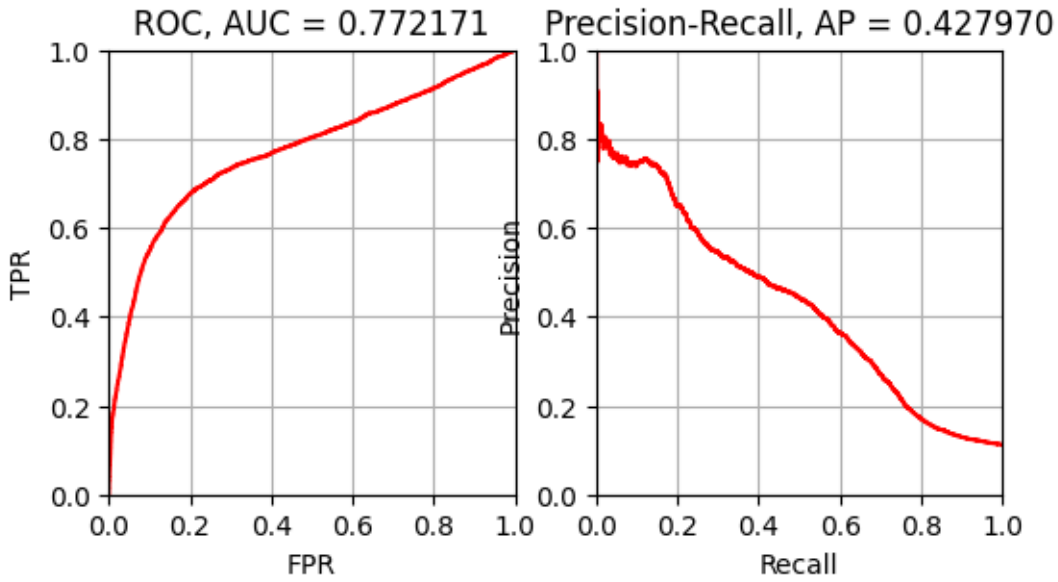
RandomForestClassifier	max_depth=5,n_estimators=501	Score
RandomForestClassifier	max_depth=5,n_estimators=501	0.899544732986165

Ejercicio 9: Implementamos la función get_metrics (véase fichero pml.py) que devuelve un conjunto de métricas de evaluación para el clasificador clf sobre los datos x, t. usamos como clasificador LogisticRegression, a continuación, los resultados de las métricas:

Confusion matrix:
[[30717 301]
[3204 699]]

score = 0.8996305947710547

Figura 12 Curva ROC, Precision y Recall - Training clasificador LogisticRegression



Con los datos de entrenamiento el clasificador LogisticRegression(maxi ter= 1000) tiene estas métricas de evaluación: ROC AUC = 0.7721 y el Precisión- Recall, Ap= 0.4279, y un score (0.8996305947710547) son bajos porque las clases están desbalanceadas.

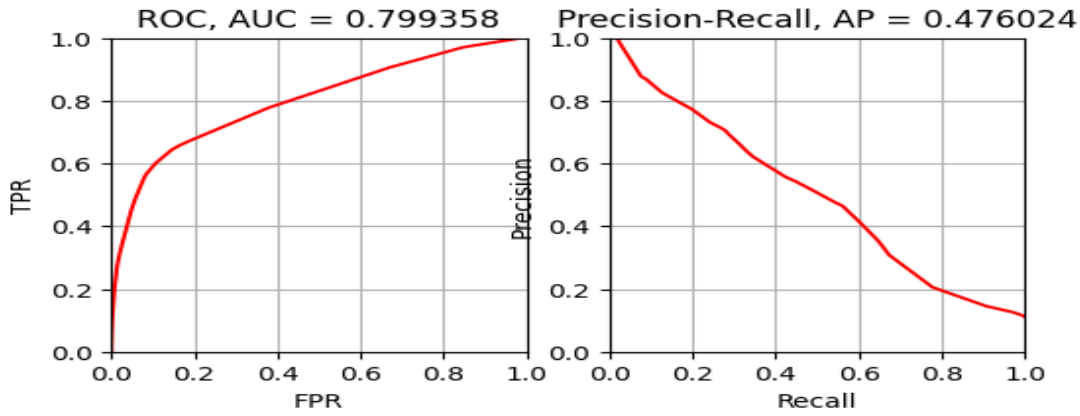
Mientras que con los datos de entrenamiento para el clasificador DecisionTreeClassifier(max_dept=7) los resultados de las métricas:

ROC AUC = 0.799 y el precisión- Recall, Ap= 0.476

score = 0.9063314338077375

Confusion matrix:
[[30565 453]
[2818 1085]]

Figura 13 Curva ROC, Precision y Recall Training clasificador DecisionTreeClassifier



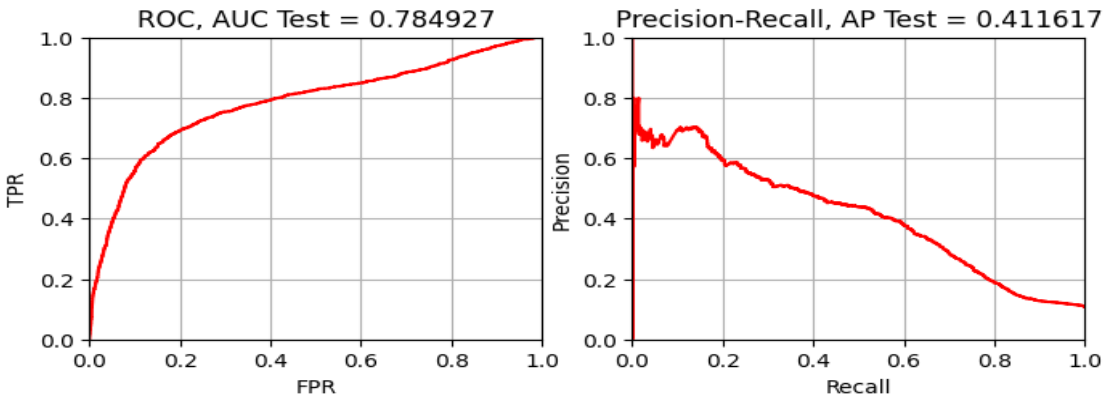
Hemos separado de estos datos un conjunto de test para obtener las métricas sobre este último

Ahora realizamos el mismo procedimiento en los datos de prueba usamos el mismo clasificador clasificador LogisticRegression(maxi ter= 1000) obtenemos las siguientes métricas: LogisticRegression = ROC AUC = 0.784 y el precisión- Recall, Ap= 0.4116

score = 0.8993032356590627

Confusion matrix:
[[9222 115]
[940 200]]

Figura 14 Curva ROC, Precision y Recall Prueba- LogisticRegression

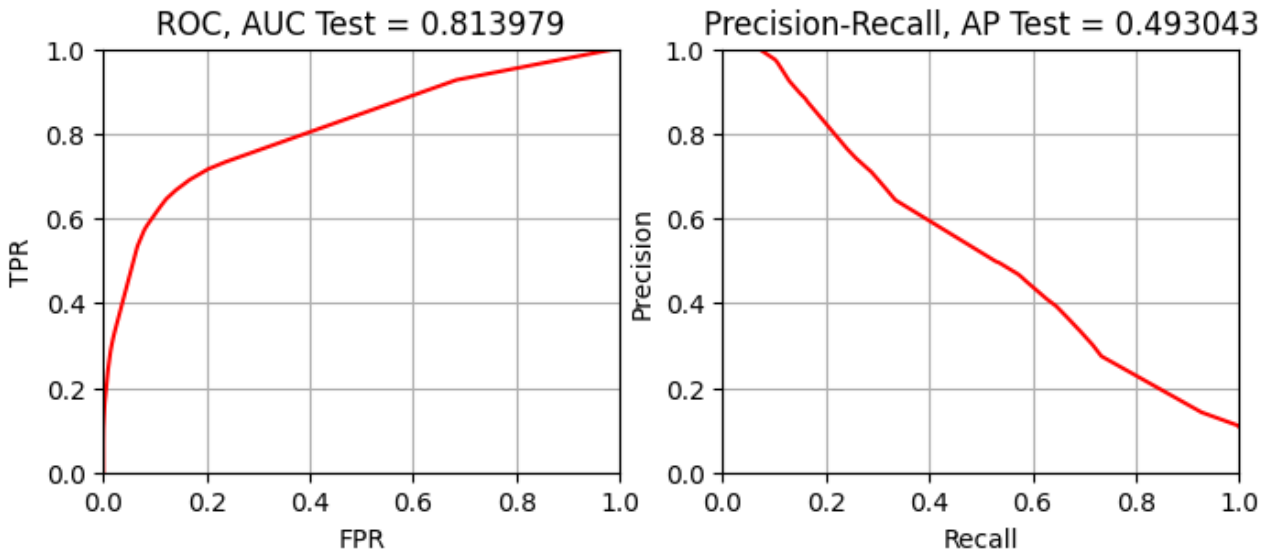


Con los datos de prueba usamos el mismo clasificador y repetimos el procedimiento realizado con los datos de entrenamiento.

DecisionTreeClassifier(max_depth =7) ROC AUC = 0.8139 y el precisión- Recall, Ap= 0.4930

score = 0.9097069771881263
Confusion matrix:
[[9212 125]
[821 319]]

Figura 15 Curva ROC, Precision y Recall Prueba- DecisionTreeClassifier



Ejercicio 10: Implementamos la función predict_test (véase fichero pml.py) que aplica el clasificador clf a los datos de test x y escribe en el fichero fname las probabilidades asignadas a la clase 1 (yes) hemos conseguido un fichero con tales probabilidades.

Ejercicio 11: Utilizando las funciones anteriores hemos estimado las probabilidades de clase 1 (yes) de todos los puntos contenidos en el fichero bank-transformed-test-no-labels.csv. ver archivo adjunto (predictions.csv)

Conclusiones

Del proceso de entrenamiento y validación que hemos realizado podemos concluir que, el modelo de Regresión Logística tiene un ROC AUC más bajo y una precisión-recall y score razonablemente buenos en comparación con el Decision Tree Classifier. No obstante, lo anterior nos hemos decantado por el modelo de Regresión Logística por el siguiente motivo vista la naturaleza y distribución de las clases. La Regresión Logística tiende a funcionar mejor con clases altamente desbalanceadas debido a la naturaleza de su función de pérdida (entropía cruzada). Esta técnica se adapta bien a conjuntos de datos desequilibrados y puede manejarlos mejor que algunos otros algoritmos de clasificación.

Los árboles de decisión, incluido el DecisionTreeClassifier, podrían ser más propensos al sobreajuste en situaciones de desequilibrio de clases, especialmente si el árbol no se controla adecuadamente o no damos una profundidad adecuada. En este caso en concreto se podría inferir que este modelo está sobre ajustado.