Clasificación supervisada de la comercialización bancaria

INTRODUCCIÓN A CIENCIA DE DATOS

01 de Octubre de 2025

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Iván García Mestiza ivan.garcia@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

Objetivo

Aplicar técnicas de clasificación supervisada para analizar las características de los clientes y determinar cuáles factores influyen en la contratación de un servicio bancario tras ser contactados por un campaña de marketing. Para ello, usaremos los datos del **Bank Market Dataset** disponible en el repositorio **UCI Machine Learning**, comparando el desempeño de distintos métodos de clasificación.

Introducción

La página del sitio web UC Irvine Machine Learning Repository ofrece una amplia colección de bases de datos, teorías de dominios y generadores de datos que son utilizados en la investigación y práctica del aprendizaje automático. En este trabajo, se emplean los datos de una campaña de marketing realizada por una institución bancaria en Portugal, cuyo objetivo fue identificar clientes con mayor probabilidad de contratar un depósito a plazo tras ser contactados por teléfono.

Con este propósito, retomaremos el problema de clasificación supervisada, donde aplicaremos los métodos **Naive Bayes, LDA, QDA, Fisher y k-NN**, con el fin de analizar y comparar su desempeño en la detección de patrones entre los clientes, y haremos un análisis para identificar el mejor k en el método k-NN. La base de datos utilizada se encuentra disponible públicamente en el repositorio de UCI en el siguiente enlace: `https://archive.ics.uci.edu/dataset/222/bank+marketing`.

Podremos observar que la proporción de personas que aceptaron el producto es muy pequeña en comparación con el total de datos, lo cual hace que los clasificadores en general aprendan más sobre las personas que dijeron que no en comparación con las que dijeron que sí. Además, veremos que las distribuciones de las variables de las personas que aceptaron el producto en general son muy similares a las de la población en general, por lo que los métodos de clasificación considerando todas las variables en general no tienen un buen rendimiento, pero introduciremos una técnica para identificar las variables más relevantes y veremos cómo es que cambian los resultados.

Exploración inicial

La campaña de marketing del banco consiste en ofrecer depósitos a plazo por medio de llamadas telefónicas a sus clientes, quienes fueron contactados una o más veces con el objetivo de venderles este producto. Los archivos contienen información sobre 45211 llamadas, en las cuales se realizó una encuesta a los clientes sobre 17 variables, incluyendo la respuesta final sobre si contratan o no el producto. La información la podemos distinguir en dos grupos:

1. **Datos del cliente**

   I. *Age*: la edad del cliente.

   II. *Job*: A qué se dedica (las respuestas son una de: trabajo manual, gerencia, técnico, administración, servicios, retirado, auto empleado, empresario, desempleado, labores del hogar, estudiante y desconocido).

   III. *Marital*: Su estado civil (casado, divorciado o soltero).

   IV. *Education*: Su nivel máximo de estudios (codificado en educación primaria, secundaria y terciaria, con algunos desconocidos).

   V. *Default*: Si alguna vez ha dejado de pagar un crédito.

   VI. *Balance*: El saldo promedio anual que tiene en su cuenta.

   VII. *Housing*: Si tiene un crédito hipotecario.

   VIII. *Loan*: Si tiene otro préstamo personal.

2. **Datos de la interacción del cliente con la campaña**

   I. *Contact*: Se refiere a si se le contactó por teléfono fijo o celular.

   II. *Day and month*: Día y mes en que se realizó la llamada.

   III. *Duration*: La duración en segundos de la llamada.

   IV. *Campaign*: El número total de veces que se ha contactado a este cliente durante la campaña actual.

   V. *Pdays*: Número de días que han pasado desde que se contactó al cliente en una campaña anterior.

   VI. *Previous*: Número de veces que se contactó al cliente antes de esta campaña (numérico, -1 indica que no se había contactado al cliente previamente).

   VII. *Poutcome*: Resultado de la campaña de marketing anterior (desconocido, éxito, fracaso, otro).

   VIII. *Y*: Indica con Yes si el cliente contrató el depósito a plazo y con No si no lo hizo.

De las variables (columnas) anteriores, las numéricas son *age*, *balance*, *day*, *duration*, *campaign*, *pdays*, y *previous*, siendo el resto categóricas. La base de datos presente codifica los datos faltantes con "unknown", y solamente 4 columnas están incompletas, que son las que se muestran en el Cuadro 1.

Columna	Porcentaje de datos faltantes
Job	$0.637\,\%$
Education	$4.107\,\%$
Contact	$28.798\,\%$
Poutcome	$81.747\,\%$

Cuadro 1: Porcentaje de datos faltantes por columna.

Por otro lado, la columna *previous* no tiene ningún dato igual a -1, y solamente 8257 de las 45211 filas tienen un valor mayor a 0, representando el $18.263\,\%$ de las llamadas. Esto y el cuadro anterior parece indicar que las columnas *previous* y *poutcome* no deberían tener mucho peso en la clasificación, porque tienen demasiados datos faltantes, por lo que podemos decidir eliminarlas. Sin embargo, un análisis más formal de selección de variables se presenta más adelante.

Por otra parte, las columnas de *job* y *education* tienen pocos datos faltantes, y corresponden a datos categóricos, por lo que, para un mejor análisis, fueron imputadas con "hot-deck", permitiendo mantener de manera aproximada la proporción de las clases existentes, y al ser en porcentajes pequeños, el sesgo que se introduce es intrascendente.

Con el objetivo de verificar si las columnas numéricas presentan información repetida en algún sentido, podemos graficar la matriz de correlación entre ellas, la cual se muestra en la Figura 1. En ella podemos observar que las únicas que presentan una correlación relativamente alta son *pdays* y *previous*, lo cual tiene sentido, pues ambas miden cosas similares. Sin embargo, más adelante se hace análisis un poco más profundo para decidir si son representativas o se pueden eliminar.

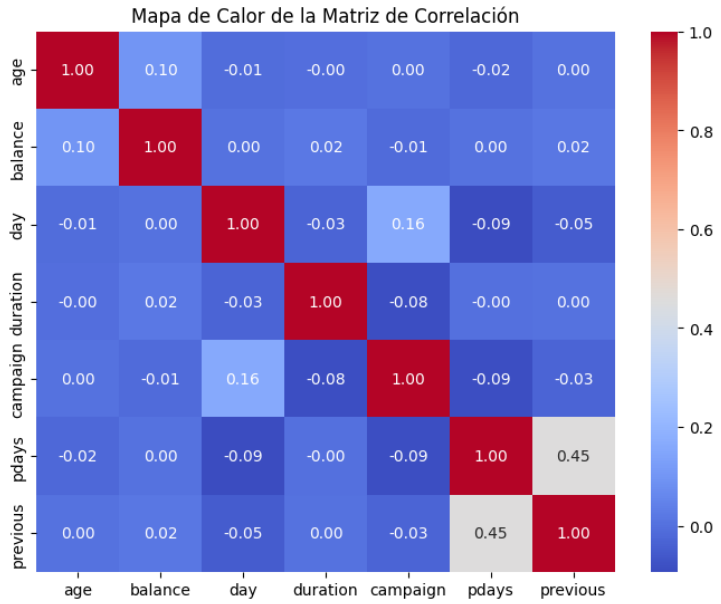

Figura 1: Mapa de calor de la matriz de correlación de variables numéricas.

Dado el número de registros telefónicos disponibles, nos interesa analizar la distribución de las variables y detectar posibles valores atípicos. En primer lugar, se examinó la variable objetivo Y, observando que únicamente $5,289$ de los $45,211$ clientes aceptaron el depósito a plazo, lo que representa aproximadamente el $11.7\,\%$, mientras que el $88.3\,\%$ restante lo rechazó. Este hallazgo es relevante, ya que plantea una de las preguntas principales del estudio: ¿qué características comparten los clientes que aceptaron el depósito? A continuación se mostramos histogramas y gráficos de barras de algunas variables de interés.

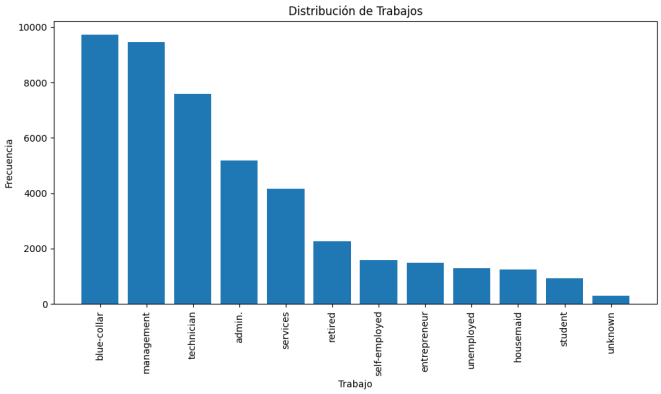

Figura 2: Gráfico de barras del trabajo del cliente.

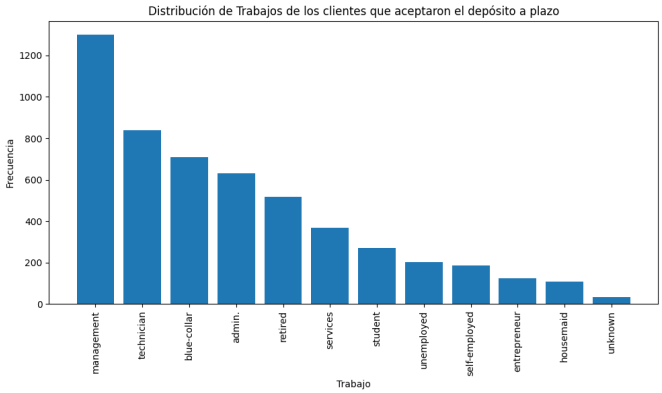

Figura 3: Gráfico de barras del trabajo de los clientes que aceptaron el depósito a plazo.

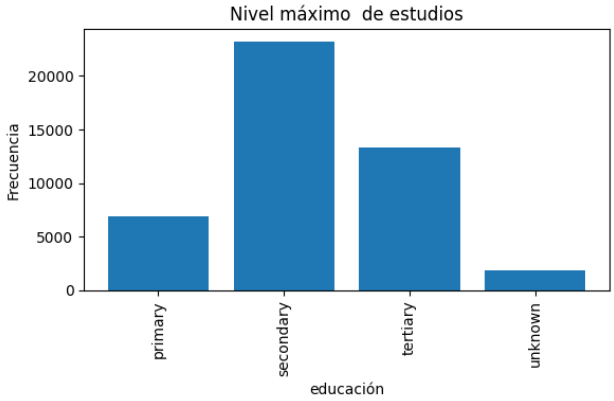

Figura 4: Gráfico de barras del máximo nivel de estudios del cliente.

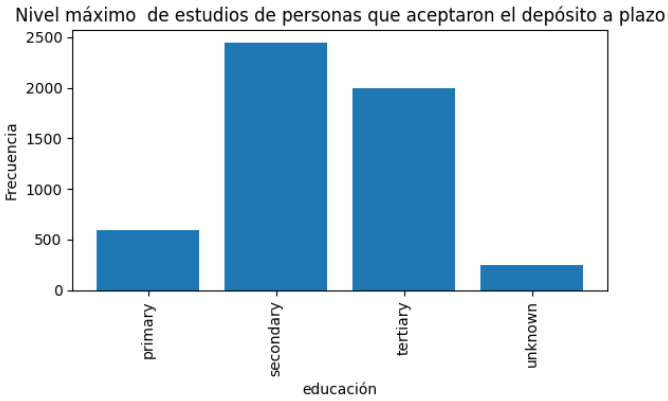

Figura 5: Gráfico de barras del máximo nivel de estudios de los clientes que aceptaron el depósito a plazo.

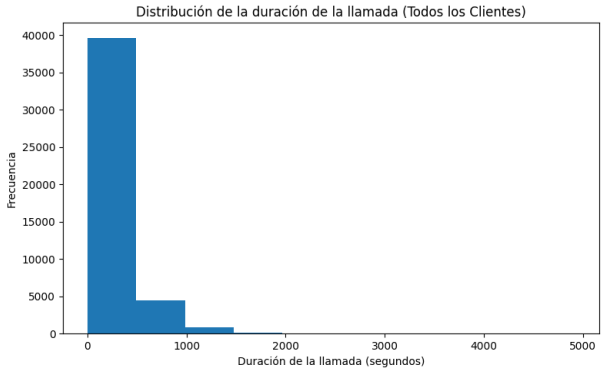

Figura 6: Histograma de la duración en la llamada con el cliente.

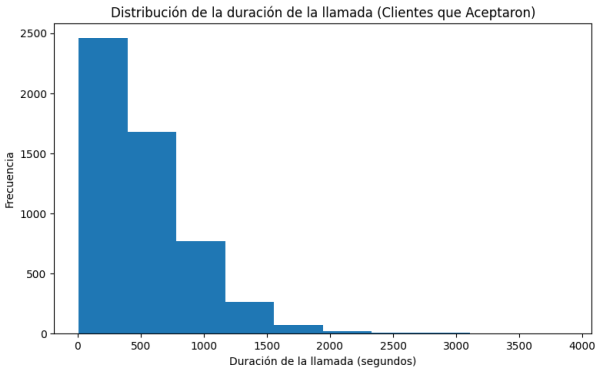

Figura 7: Histograma de la duración en la llamada con clientes que aceptaron el depósito a plazo.

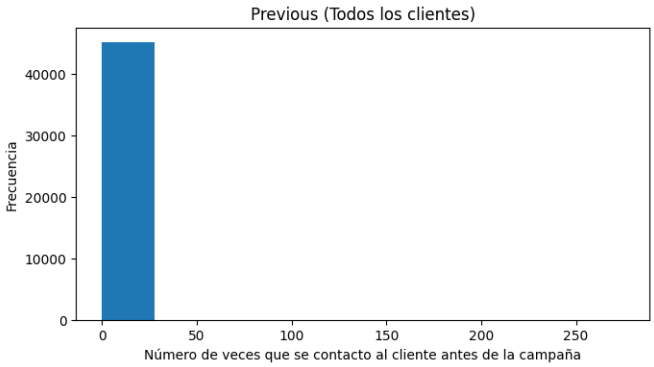

Figura 8: Histograma del número de veces que se contacto al cliente antes de la campaña.

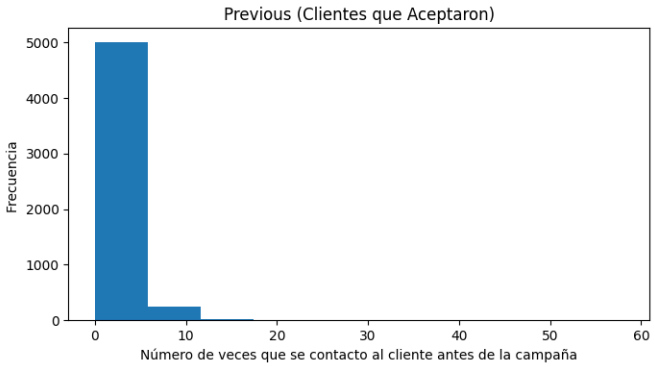

Figura 9: Histograma del número de veces que se contacto al cliente que acepto antes de la campaña.

Estos gráficos permiten identificar patrones de distribución y posibles diferencias significativas entre los clientes que aceptaron y los que rechazaron el depósito. Comparando las gráficas de barra de los trabajos de los clientes, vemos que la principal diferencia es que, aunque los trabajadores blue collar son el grupo más grande de clientes, no son los que más aceptaron la oferta. Sin embargo, los estudiantes son un grupo minoritario en general, pero tienen mayor representación en la gráfica de los que aceptaron, lo cual indica que son un segmento muy propenso a aceptar la oferta. También podemos observar que, los clientes con educación terciaria y aquellos con los que se logra una llamada más larga son mucho más propensos a aceptar la oferta. El número de contactos previos parece tener un impacto menor, ya que la mayoría de los éxitos provienen de clientes nuevos.

Clasificación supervisada

Usaremos los datos para construir modelos de predicción que permitan identificar las características que comparten los clientes que tuvieron mejores interacciones con la campaña de marketing y que adquirieron el depósito a plazo. De esta manera, en el futuro pueden enfocarse a los clientes con mayor potencial. Los modelos utilizados son k-NN, LDA, QDA, Naive Bayes y el Criterio de Fisher.

Puesto que la mayoría de dichos métodos dependen de una noción de "distancia", es conveniente escalar las variables numéricas, las cuales representan mediciones de distintos fenómenos; y de esta manera podemos evitar que la presencia de valores muy grandes en una de ellas opaque la influencia que puedan tener el resto. Además, es conveniente hacer una codificación de las variables categóricas, siendo la más adecuada en la mayoría de las ocasiones la "one-hot", puesto que muchas de las categorías no tienen un orden intrínseco, a excepción de la educación, que puede codificarse con "label-encoding".

Con las consideraciones anteriores, al realizar el entrenamiento y evaluación de las clasificaciones mencionadas con un conjunto de prueba de tamaño igual al $30\,\%$ de la información total, obtuvimos las matrices de confusión de la Figura 10. En una primera instancia entrenamos el k-NN con $k = 5$, pero más adelante veremos una manera de obtener el óptimo.

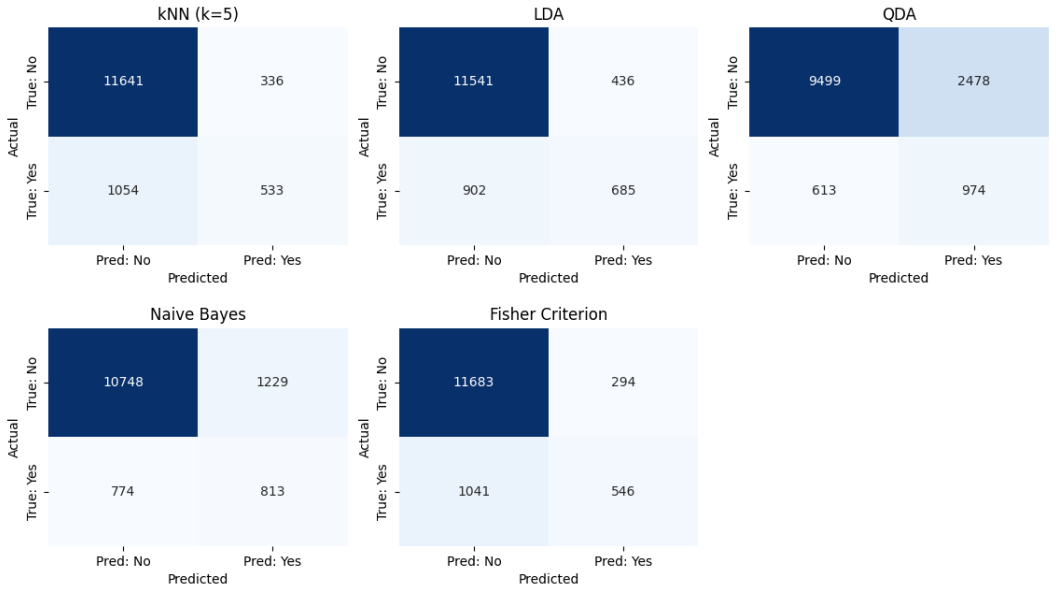

Figura 10: Matrices de confusión para distintos modelos de clasificación.

Para evaluar el desempeño de dichos clasificadores, podemos calcular las métricas como exactitud, sensibilidad, especificidad, F1 y AUC. Los resultados obtenidos son los que se muestran en el Cuadro 2.

Clasificador	Exactitud	Sensibilidad	Especificidad	F1	AUC
kNN (k=5)	0.898	0.336	0.972	0.43	0.826
LDA	0.901	0.432	0.964	0.51	0.904
QDA	0.772	0.614	0.793	0.39	0.764
Naive Bayes	0.852	0.512	0.897	0.45	0.807
Criterio de Fisher	0.902	0.344	0.975	0.45	0.904

Cuadro 2: Comparación de desempeño de los clasificadores con distintas métricas.

Recordemos que la exactitud mide la proporción de clientes que se clasificaron de manera correcta. En este caso el Criterio de Fisher y el LDA son los que tienen los valores mayores, pero no sería adecuado quedarse simplemente con esta métrica, puesto que hay una mayor cantidad de clientes que rechazaron el producto a la cantidad que lo adquirieron. Así que si el objetivo es tratar de predecir si un cliente aceptará o no, hay que considerar las otras métricas. La sensibilidad corresponde a la proporción de positivos que fue correctamente identificada. En este sentido los clasificadores no tuvieron un muy buen rendimiento, lo cual puede explicarse porque la proporción de positivos es muy chica con respecto a los negativos (solo 5289 clientes aceptaron el producto, lo que corresponde a un $11.698\,\%$). Considerando lo anterior, los resultados del QDA son bastante buenos bajo esta métrica.

Por otra parte, la especificidad corresponde a la proporción de negativos que fue correctamente identificada, y en este sentido esperaríamos que los clasificadores no tuvieran mucho problema, puesto que hay una gran cantidad de datos negativos de los que pudieron aprender. Aquí vemos que el Criterio de Fisher vuelve a resaltar, lo cual tiene sentido, puesto que era el que mayor exactitud había tenido, y una gran parte de los datos son negativos, aunque el k-NN y el LDA también tienen un buen rendimiento. Ahora bien, el F1-Score permite ver cuál de los estimadores tiene un mejor balance entre precisión (no tantos falsos positivos) y sensibilidad (poder identificar suficientes positivos). En este sentido el LDA presenta un mejor balance, lo que puede implicar que los datos se encuentran separados de manera aproximadamente lineal. Por último, valores de AUC cercanos a 0.5 indican que el modelo no es mucho mejor que "adivinar" de manera aleatoria, y cercanos a 1 indican que es un muy buen clasificador, y en esta métrica los de mejores resultados son el LDA y el Criterio de Fisher. De nuevo, esto tiene sentido porque ambos se relacionan con cantidades lineales: un hiperplano de separación, como en el LDA, o una proyección sobre un eje de una única dimensión.

El análisis anterior se hizo para $k = 5$ en el k-NN. Sin embargo, esta elección puede resultar de cierto modo arbitraria, y surge la necesidad de justificar cuál podría ser un buen valor para k. En la Figura 11 se muestra el desempeño del clasificador k-NN para diferentes valores de k, en términos de exactitud y AUC. Como se puede observar, a partir de un valor de $k = 15$ no se presenta una mejora significativa para la exactitud ni el AUC, por lo que podemos considerar este modelo para su comparación con los otros.

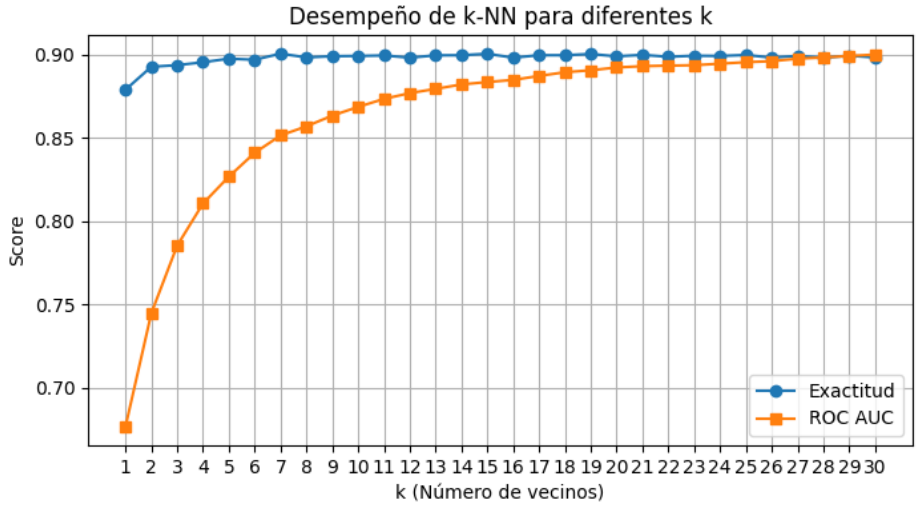

Figura 11: Comparación de rendimiento de k-NN para diferentes valores de k.

Las métricas obtenidas para $k = 15$ en k-NN se presentan en el Cuadro 3. Como vemos, dicho estimador es muy bueno para detectar los casos negativos, pues tiene alta especificidad y una buena exactitud. Sin embargo, tiene muy poca sensibilidad, por lo que no es tan bueno para distinguir los casos positivos, lo cual sugiere que, para la naturaleza de los datos, el método k-NN no parece ser tan adecuado.

Clasificador	Exactitud	Sensibilidad	Especificidad	F1	AUC
kNN (k=15)	0.90	0.30	0.98	0.41	0.882

Cuadro 3: Desempeño de kNN con $k = 15$.

Selección de Variables

Otra cuestión de interés en este trabajo es identificar cuáles son las características con mayor peso para clasificar de manera adecuada los datos, es decir, determinar las variables más importantes. Como se mencionó anteriormente, una manera de evaluar la relevancia de las variables es verificar su multicolinealidad mediante la matriz de correlación, ya que esto permite basar la selección en propiedades estadísticas de las variables, independientemente del modelo de clasificación.
En este caso, como se puede ver en la Figura 1 las únicas variables numéricas que presentan correlación relevante son *previous* y *pdays*. Sin embargo, dado que la correlación no es alta y solo involucra una variable, no resulta conveniente eliminarla.
A partir de lo anterior, y de la revisión en la literatura [2] y [3], encontramos que un método adecuado para la clasificación y jerarquización de variables predictoras es el Random Forest, propuesto por Breiman (2001). Este método consiste en combinar múltiples árboles de decisión con el fin de mejorar la capacidad predictiva y reducir el sobreajuste. Su funcionamiento, a grandes rasgos, es el siguiente:

1. Se generan múltiples subconjuntos de datos de entrenamiento mediante muestreo con reemplazo (bootstrap).

2. En cada subconjunto se ajusta un árbol de decisión.

3. En el caso de clasificación, el resultado final se obtiene mediante voto mayoritario; en regresión, a través del promedio de las predicciones.

En Random Forest no se consideran todas las variables predictoras en cada división, sino un subconjunto aleatorio de ellas. Este mecanismo reduce la correlación entre los árboles y produce un ensamble más estable y preciso. Además de su buen desempeño predictivo, Random Forest permite evaluar la importancia de las variables. El enfoque principal que usa es el siguiente:

Reducción de impureza (Mean Decrease in Impurity, MDI): Cada vez que una variable es utilizada para dividir un nodo, se mide la mejora lograda en el criterio de división usando la medida de Gini [1], la cual nos determina qué tan mezcladas están las clases en el nodo. De esta manera, lo que hace el algoritmo es evaluar en todas las divisiones posibles y selecciona la que minimiza la medida de Gini (buscamos que sea cercana a cero).

Este método se encuentra programado en el objeto RandomForestClassifier de Python. Así que implementándolo para nuestros datos obtenemos los resultado de la Figura 12.

Esto nos permite inferir que una posible selección de variables predictoras corresponde a las primeras 7 con mayor importancia, ya que a partir de ese punto el aporte de las demás variables no muestra variaciones significativas. Las variables seleccionadas son: *“num_duration”, “num_balance”, “num_age”, “num_day”, “num_poutcome_sucess”, “num_campaign” y “num_pdays”*. Una posible interpretación de estas variables seleccionadas es la siguiente: la variable *num_duration* refleja la duración de la llamada de contacto, lo cual tiene sentido, ya que un mayor tiempo puede asociarse con un mayor interés de la persona. El *balance* sugiere que el promedio de saldo disponible podría influir en la decisión, lo cual es razonable al considerar proyectos de inversión como los depósitos a plazo. La variable *age* constituye a un factor demográfico relevante, lo cual nos hace pensar que tal vez en ciertos rangos de edad las personas suelen mostrar una mayor o menor disposición a participar en este tipo de productos. Por su parte, *num_days* podría indicar que en determinado días del mes , como las quincenas, existe mayor o menor interés. Finalmente las variables *num_campaign* y *num_pdays* pueden asociarse con la frecuencia e historial de contactos previos, factores que también aportan información valiosa al modelo.

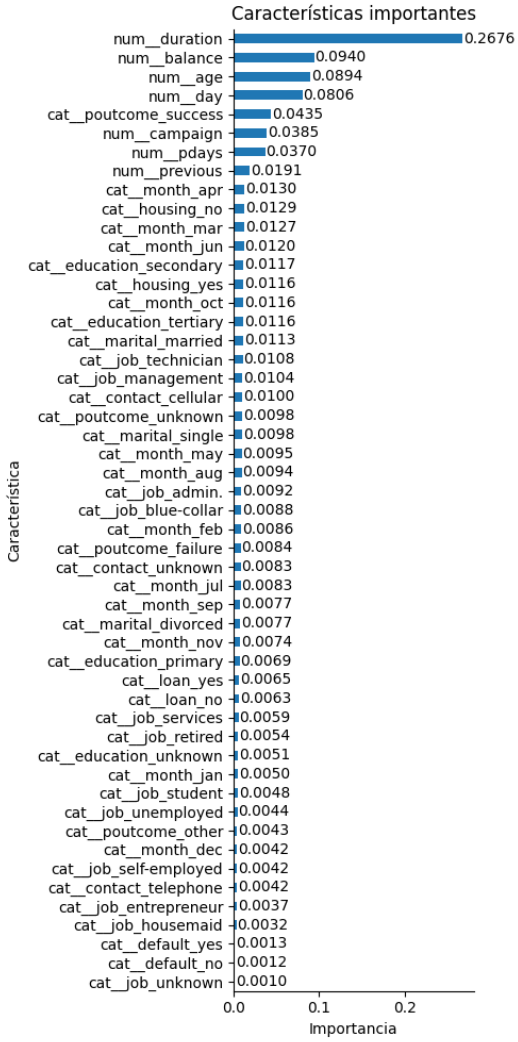

Figura 12: Importancia de las características determinada por Random Forest.

Probando nuevamente los modelos de predicción pero solo con estas variables obtenemos los resultados de ajuste en el Cuadro 4.

Clasificador	Exactitud	Sensibilidad	Especificidad	F1	AUC
kNN (k=15)	0.897	0.32	0.973	0.42	0.843
LDA	0.899	0.37	0.969	0.46	0.855
QDA	0.883	0.31	0.959	0.39	0.692
Naive Bayes	0.863	0.44	0.920	0.43	0.810
Criterio de Fisher	0.897	0.29	0.978	0.39	0.856

Cuadro 4: Comparación de desempeño de los clasificadores con distintas métricas para un subconjunto de variables seleccionadas.

De los resultados obtenidos podemos observar que, en todos los modelos, la especificidad mejoró ligeramente, excepto en el caso de QDA, donde la mejora fue considerable. Esto significa que QDA ahora resulta muy eficaz para identificar los verdaderos negativos. Sin embargo, recordemos que nuestro principal interés está en la sensibilidad.

En este aspecto, la mayoría de los modelos se mantuvieron estables, salvo nuevamente QDA, cuya sensibilidad disminuyó de manera significativa. Esto evidencia que, con las variables utilizadas, QDA pierde las ventajas que previamente mostraba frente a los demás clasificadores, pero hemos ganado una reducción de variables significativa que puede disminuir el tiempo de entrenamiento de dichos modelos.

Análisis Crítico

El análisis comparativo previo muestra que los métodos lineales, como el LDA y el Criterio de Fisher, ofrecen el mejor balance entre precisión y sensibilidad. El QDA, en cambio, aunque no alcanza los niveles de exactitud, presenta un desempeño más alto en sensibilidad, el cual resulta ser el más adecuado si el objetivo es identificar con mayor certeza a los clientes que aceptarán el producto. Métodos como k-NN se ven limitados por el desbalance de clases, lo que dificulta la detección de casos positivos aún cuando se ajusta el parámetro k. En este contexto, la exactitud por sí sola no resulta suficiente como criterio de evaluación, siendo más apropiado atender a métricas como el F1 y AUC, que permiten valorar mejor el comportamiento del clasificador frente a datos desbalanceados.

Además, la selección de variables no mostró un impacto significativo, ya que el desempeño de los modelos resultó muy similar. Esta técnica podría ser útil en escenarios donde se deba trabajar con grandes volúmenes de datos y las 16 características disponibles resulten ser muy pesadas para el análisis. Sin embargo, en este caso no parece ser estrictamente necesaria. Cabe destacar que, de acuerdo con la literatura [4], estos datos provienen originalmente de una base de datos con mayor número de variables predictoras, lo que sugiere que las variables con las que contamos ya poseen una relevancia considerable para los métodos de clasificación empleados.

Una manera de extender el análisis es considerar modelos más flexibles como en Random Forest o incluso clasificación no supervisada con redes neuronales, los cuales podrían ofrecer mejoras significativas para datos desbalanceados como en este caso.

Referencias

[1] Departamento de Matemática Aplicada, ETSI Navales, UPM. (2021). *02.4 Métodos de Clasificación – Árboles de Decisión*. Recuperado de `https://dcain.etsin.upm.es/~carlos/bookAA/02.4_MetodosdeClasificacion-ArbolesdeDecision.html`

[2] Amat Rodrigo, J. (2020). *Random Forest con Python*. Cienciadedatos.net. Recuperado de `https://cienciadedatos.net/documentos/py08_random_forest_python.html`

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[4] Moro, S., Cortez, P., & Rita, P. (2014). *A data-driven approach to predict the success of bank telemarketing*. Decision Support Systems, 62, 22–31. `https://doi.org/10.1016/j.dss.2014.03.001`