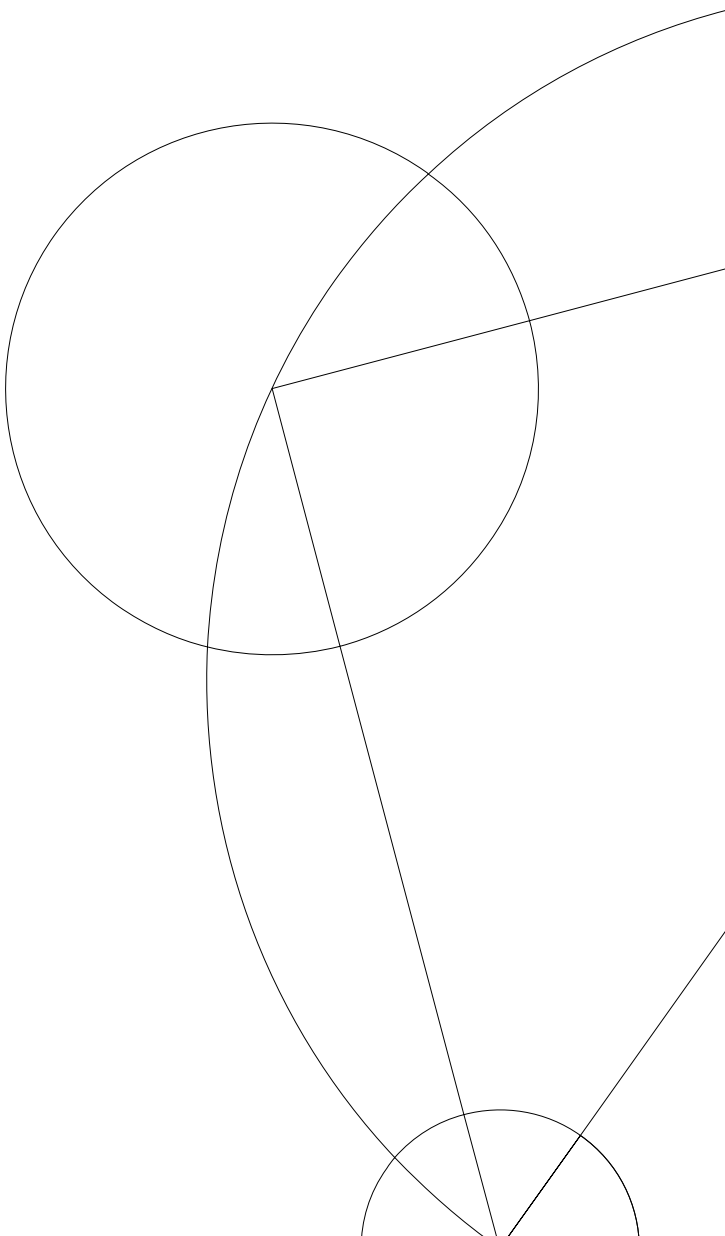
Estudio de datos de campaña de marketing para depósitos a plazo fijo

Alfredo Bistrain, Guillermo Aguilar, Oswaldo Bueno

Introducción a la Ciencia de Datos

Tarea 2



1 de octubre de 2025

Introducción

El conjunto de datos corresponde a campañas de *marketing directo* de una institución bancaria portuguesa. Dichas campañas se llevaron a cabo entre mayo de 2008 y noviembre de 2010, principalmente mediante llamadas telefónicas, con el objetivo de promover un *depósito a plazo fijo*. El conjunto de datos fue elaborado por Paulo Cortez (Universidad de Minho), Sérgio Moro y Paulo Rita (ISCTE-IUL) en 2014, y ha sido descrito en [4].

El conjunto de datos analizado en este trabajo está basado en el conjunto de datos UCI "Bank Marketing" (son casi idénticos a los usados por Moro et al. en [4], pero no se incluyen todas las variables debido a asuntos de privacidad). Los datos considerados se enriquecen con la adición de cinco nuevas características/atributos sociales y económicos (indicadores a nivel nacional de un país con una población de aproximadamente 10 millones), publicados por el Banco de Portugal y disponibles públicamente.

La variable respuesta (y) es binaria y toma los valores *yes* o *no*, indicando si el cliente suscribió o no un depósito a plazo. Esto convierte el problema en uno de clasificación binaria, en el cual se busca explorar el perfil de clientes con mayor probabilidad de realizar un depósito a plazo fijo.

En este trabajo se expone un análisis del conjunto de datos en el cual se detallan las variables, su significado, y las distintas problemáticas que los datos presentan.

Tras una limpieza y codificación adecuada de los datos, se implementan modelos de clasificación como *Naive Bayes*, *LDA*, *QDA*, *proyección de Fisher*, *k-NN* y *regresión logística*. Se presentan los resultados de la modelación y se discuten conclusiones al final del trabajo.

Análisis exploratorio de los datos

A continuación se enlistan las variables predictoras del conjunto de datos, así como una breve descripción de ellas y los valores que toman.

Numéricas:

- age (Edad)
- balance (Balance promedio anual, en euros)
- day (Último día de contacto en el mes)
- duration (Duración del último contacto, en segundos)
- campaign (Número de contactos realizados durante la campaña para este cliente, incluye el último contacto)
- pdays (Número de días que pasaron desde el último contacto de una campaña previa)[-1 significa que el cliente no fue previamente contactado]
- previous (Número de contactos realizados antes de la campaña para este cliente)
- emp.var.rate (Tasa de variación del empleo, indicador trimestral)
- cons.price.idx (Índice de precios al consumidor, indicador mensual)
- cons.conf.idx (Índice de confianza del consumidor, indicador mensual)
- euribor3m (Tasa de interés interbancaria europea a 3 meses, indicador diario)
- nr.emplyed (Número de personas empleadas, indicador trimestral)

Binarias:

- default (¿Incumplió con el pago de un crédito o préstamo?) ["no", "yes", "unknown"]
- housing (¿Tiene un préstamo hipotecario?) ["no", "yes", "unknown"]
- loan (¿Tiene un préstamo personal?) ["no", "yes", "unknown"]

- y (¿Realizó un depósito a plazo fijo?) [“no”, “yes”]

Categóricas:

- job (Trabajo actual)[“admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”,“student”, “blue-collar”,“self-employed”,“retired”,“technician”,“services”]
- marital (Estado civil) [“married”,“divorced”,“single”]
- education (Nivel de educación) [“basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”,“professional.course”, “university.degree”, “unknown”]
- contact (Medio de comunicación) [“telephone”, “cellular”]
- month (Último mes de contacto) [“jan”, “feb”, ..., “nov”, “dec”]
- poutcome (Resultado de la última campaña de marketing) [“nonexistent”, “failure”, “success”]

Antes de trabajar con el conjunto de datos se retiró la variable “duration”, pues la descripción de esta menciona que influye mucho en el resultado final (“no” o “sí”) debido a que una duración igual a 0 implica que el cliente no contestó y por lo tanto no realizó un depósito a plazo fijo, mientras que una duración alta de la llamada indica que el cliente estuvo interesado en lo que el banco ofrecía y es altamente probable que realice el depósito. Sin embargo, no podemos conocer la duración de la llamada sino hasta que esta finaliza, por lo tanto, para una modelación realista hemos de no considerar dicha variable.

Inicialmente se buscaron datos faltantes y se encontró que no hay datos de tipo NaN, sin embargo, se observó que hay datos registrados como “unknown” los cuales corresponden a datos desconocidos. Se decidió tratar estos datos como datos faltantes y fueron convertidos a NaN para así poder analizarlos de mejor manera. El porcentaje de datos faltantes encontrados fue el siguiente.

Variable	job	marital	education	default	housing	loan
Porcentaje	0.8 %	0.194 %	4.20 %	20.87 %	2.40 %	2.40 %

Cuadro 1: Porcentaje de datos faltantes en el conjunto de datos

En general el porcentaje de datos faltantes es pequeño (a excepción de la columna default), de manera que se pueden utilizar técnicas de imputación sin gran preocupación.

La columna default contiene un gran porcentaje de datos faltantes, y más que eso, es una variable binaria para la cual se encontró que tiene 32,591 registros y solo 3 de ellos fueron “sí”. Además, cuenta con 8597 datos faltantes, de los cuales probablemente muchos son “sí”, pues la variable “default” registra si el individuo tiene un crédito que ha caído en incumplimiento de pago; dicho esto, es razonable que muchas personas no hayan querido decir que incumplieron con el pago de un crédito o préstamo.

Debido a la gran desproporción entre “no” y “sí” de la columna “default”, cualquier método de imputación que se lleve acabo creará un gran sesgo, pues es razonable pensar que muchos de los datos faltantes son respuestas afirmativas, pero esto no se puede asegurar.

En la primera parte del análisis se hará una imputación aleatoria de acuerdo a la distribución observada, pero más adelante se considerará el conjunto de datos sin la variable “default” (por razones mencionadas anteriormente) y se comparará el desempeño de los clasificadores en ambos casos.

En este conjunto de datos se tienen distintas variables categóricas no ordinales, así como otras que sí son ordinales (“education”) y algunas del tipo binario. Para las variables categóricas no ordinales se utiliza una codificación one-hot, mientras que la variable “education” se codifica de manera numérica y creciente de acuerdo a los años y grados de estudio. Finalmente, para las variables binarias se codificaron los “no” como 0 y los “sí” como 1. Dichas codificaciones se llevaron acabo para tener solo variables numéricas y así poder hacer uso de los modelos de clasificación sin problema alguno.

Un aspecto importante a resaltar es el hecho de que la variable de respuesta es binaria y tiene una desproporción significativa entre los “no” (0) y los “sí” (1). Esto puede observarse en el siguiente gráfico.

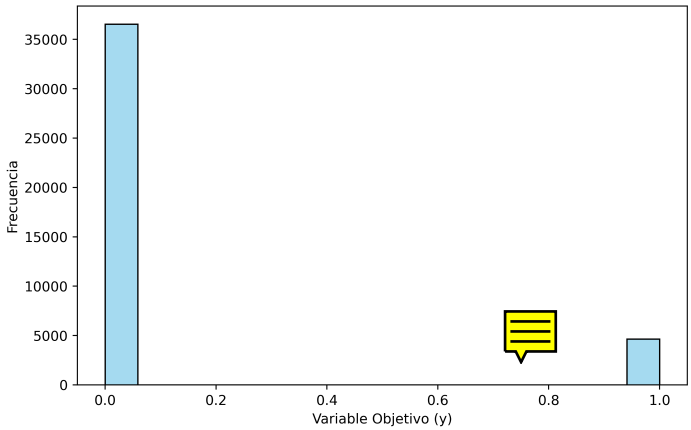


Figura 1: Histograma de la variable de respuesta

Dicha desproporción hará que los clasificadores tengan problemas clasificando correctamente a los individuos que sí deciden realizar un depósito a plazo fijo en el banco.

A continuación se presenta el mapa de correlación de las variables en consideración. A pesar de observar gran correlación (ya sea negativa o positiva) entre algunas de estas variables se ha decidido no eliminar ninguna pues el entrenamiento de los modelos funciona rápidamente aún con este número de variables. Eliminar variables en este análisis debería realizarse con precaución debido al gran número de estas.

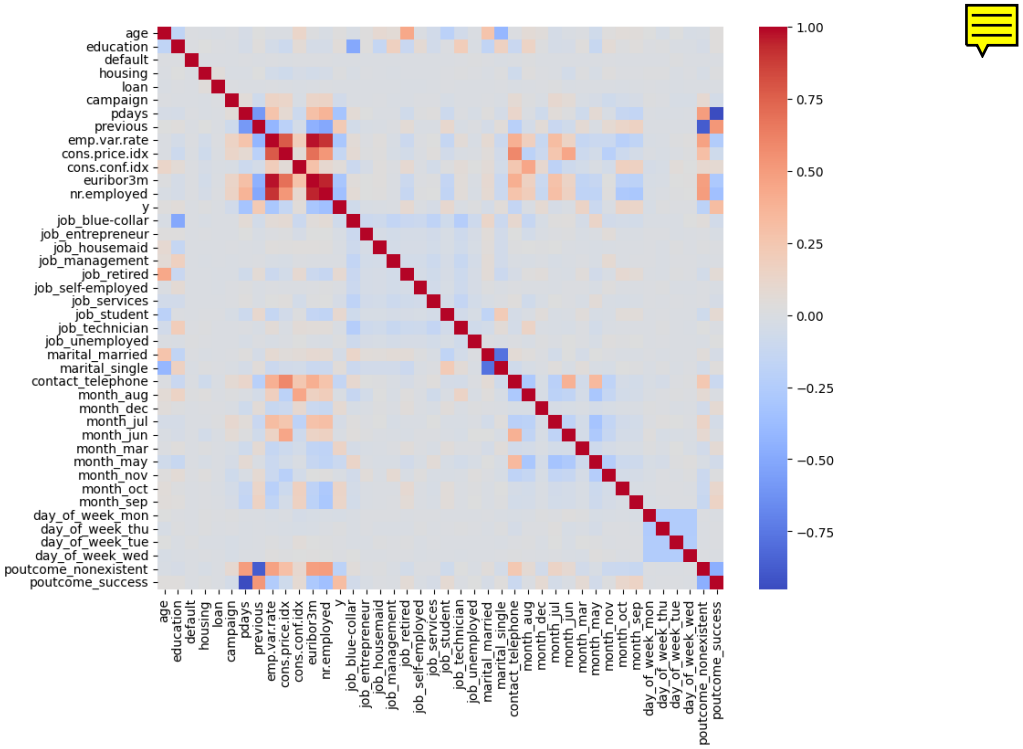


Figura 2: Matriz de correlación

Modelación y resultados

Para la modelación se separaron los datos (de manera aleatoria, pero conservando la proporción de la variable de respuesta) en un conjunto que será el de entrenamiento y otro que será el de prueba. Los tamaños de dichos conjuntos de datos serán el 70 % y el 30 % del tamaño del conjunto completo, respectivamente.

Una vez que se cuenta con el conjunto de entrenamiento, se está en condiciones de entrenar modelos de clasificación pues los datos han sido codificados de manera adecuada y se ha lidiado con los datos faltantes.

Los modelos entrenados son Naive Bayes, LDA, QDA, k-NN (donde se utilizó $k = 5$), proyección de Fisher y regresión logística (ver [1]). Para dichos modelos de clasificación se evaluaron las métricas de *accuracy, precision, recall, F1*

Modelo	**Accuracy**	**Precision**	**Recall**	**F1**
Naive Bayes	0.871	0.427	0.439	0.433
LDA	0.893	0.538	0.369	0.438
QDA	0.879	0.465	0.466	0.465
k-NN (k=5)	0.893	0.545	0.287	0.376
Fisher	0.893	0.538	0.369	0.438
Regresión logística	0.901	0.690	0.218	0.331

Cuadro 2: Comparación de modelos

La gráfica de separación de clases a través de la proyección de Fisher es la siguiente, en la cual podemos observar como estos siguen bastante mezclados.

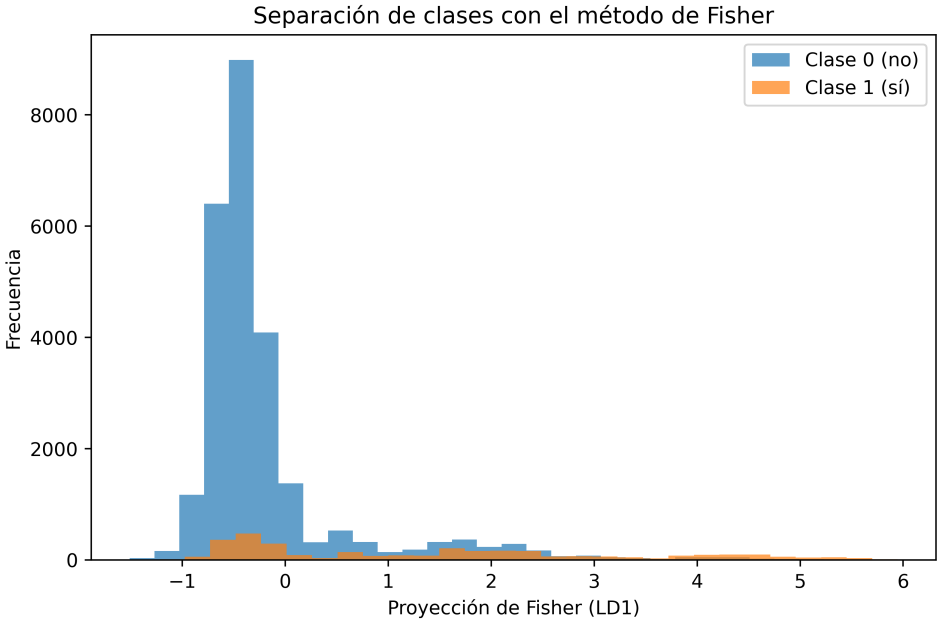


Figura 3: Separación de clases con proyección de Fisher

Asimismo, se realizó una validación cruzada k-fold con $k = 5$. En este proceso se analizaron las mismas métricas que antes. En la siguiente tabla se muestran los resultados.

Modelo	**Accuracy**	**Precisión**	**Recall**	**F1**
Naive Bayes	0.869 ± 0.002	0.422 ± 0.009	0.440 ± 0.020	0.431 ± 0.013
LDA	0.890 ± 0.002	0.518 ± 0.012	0.377 ± 0.015	0.436 ± 0.011
QDA	0.877 ± 0.003	0.455 ± 0.014	0.462 ± 0.017	0.458 ± 0.014
k-NN (k=5)	0.890 ± 0.004	0.521 ± 0.033	0.289 ± 0.015	0.371 ± 0.019
Fisher	0.890 ± 0.002	0.518 ± 0.012	0.377 ± 0.015	0.436 ± 0.011
Regresión logística	0.900 ± 0.001	0.654 ± 0.009	0.229 ± 0.019	0.339 ± 0.021

Cuadro 3: Resultados de validación cruzada (5-fold)

A pesar de que la regresión logística está clasificando peor a los datos "sí", nótese que tiene una precisión más alta que los demás modelos. Es decir, aquellos datos que clasífica como "sí", suelen estar más en lo correcto que en el resto de los modelos vistos en clase.

En general, todos los modelos que se consideraron tienen *exactitud (accuracy)* muy similar, pero no basta con ver solo esta métrica, pues sus *sensibilidades (recall)* y *precisiones* cambian mucho entre sí.

Ahora se presentan las matrices de confusión (para el conjunto de prueba) asociadas a los modelos de clasificación. Se obtienen resultados similares a lo esperado inicialmente debido a la desproporción que hay entre los "no" y los "sí" en la variable de respuesta. Obsérvese como el modelo cuadrático y Naive Bayes realizan una mejor clasificación de la clase "sí", mientras que los demás lo hacen peor pero manteniendo una mejor clasificación de la clase "no". Asimismo, obsérvese cómo el modelo de regresión logística clasifica aún mejor a la clase "no", pero clasifica peor a la clase "sí", lo cual es causado por la desproporción en la variable objetivo.

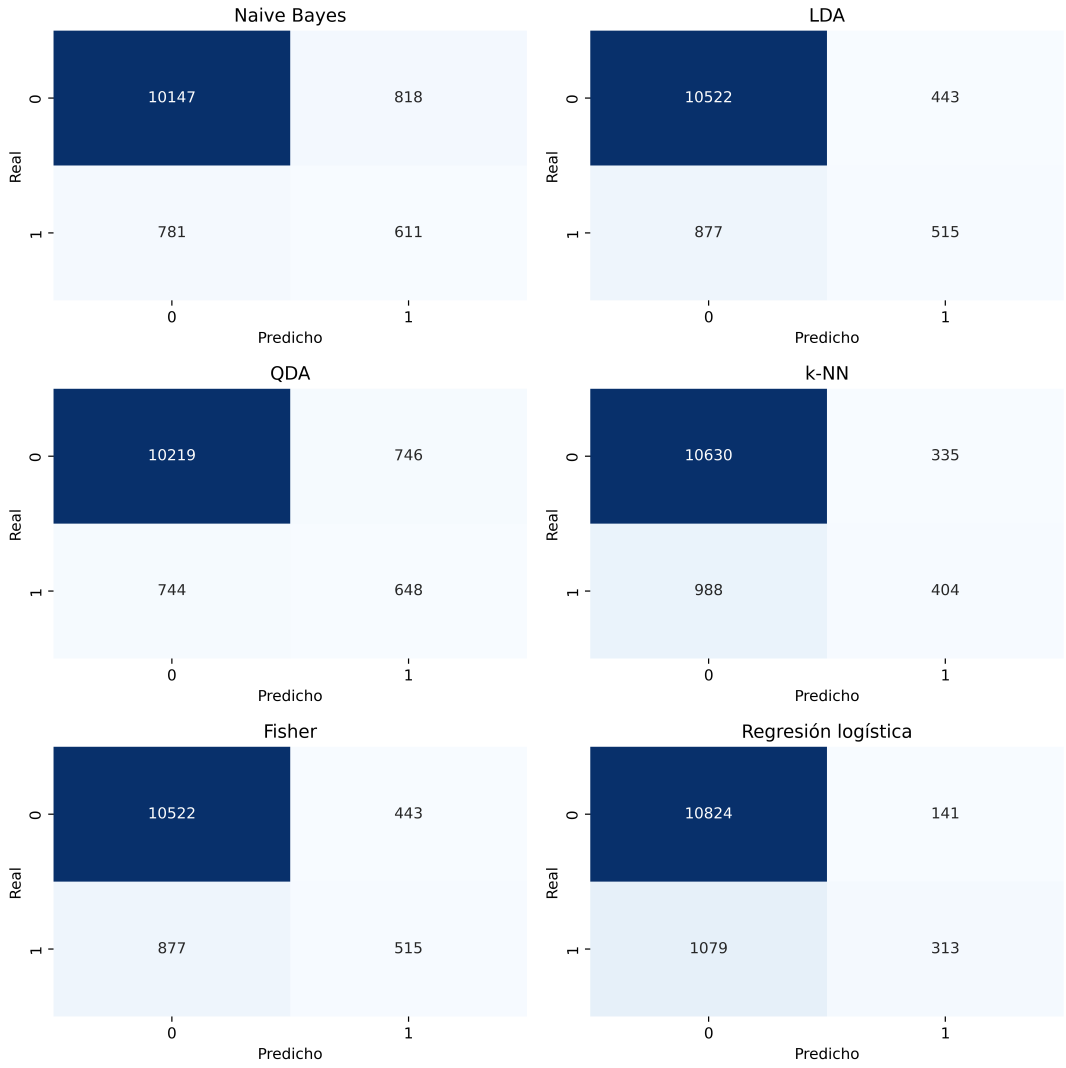


Figura 4: Matrices de confusión de los modelos entrenados

Retomando la discusión inicial acerca de la variable "default", se presentan las matrices de confusión asociadas al conjunto de datos dejando esta columna por fuera. Obsérvese como los resultados de la clasificación son prácticamente iguales a excepción del modelo cuadrático, el cual empeoró su clasificación de la clase "no". Esto muestra como, a pesar de que la variable default pareciera inducir mucho sesgo o ruido en los modelos, el quitarla no tiene un efecto positivo en los resultados.

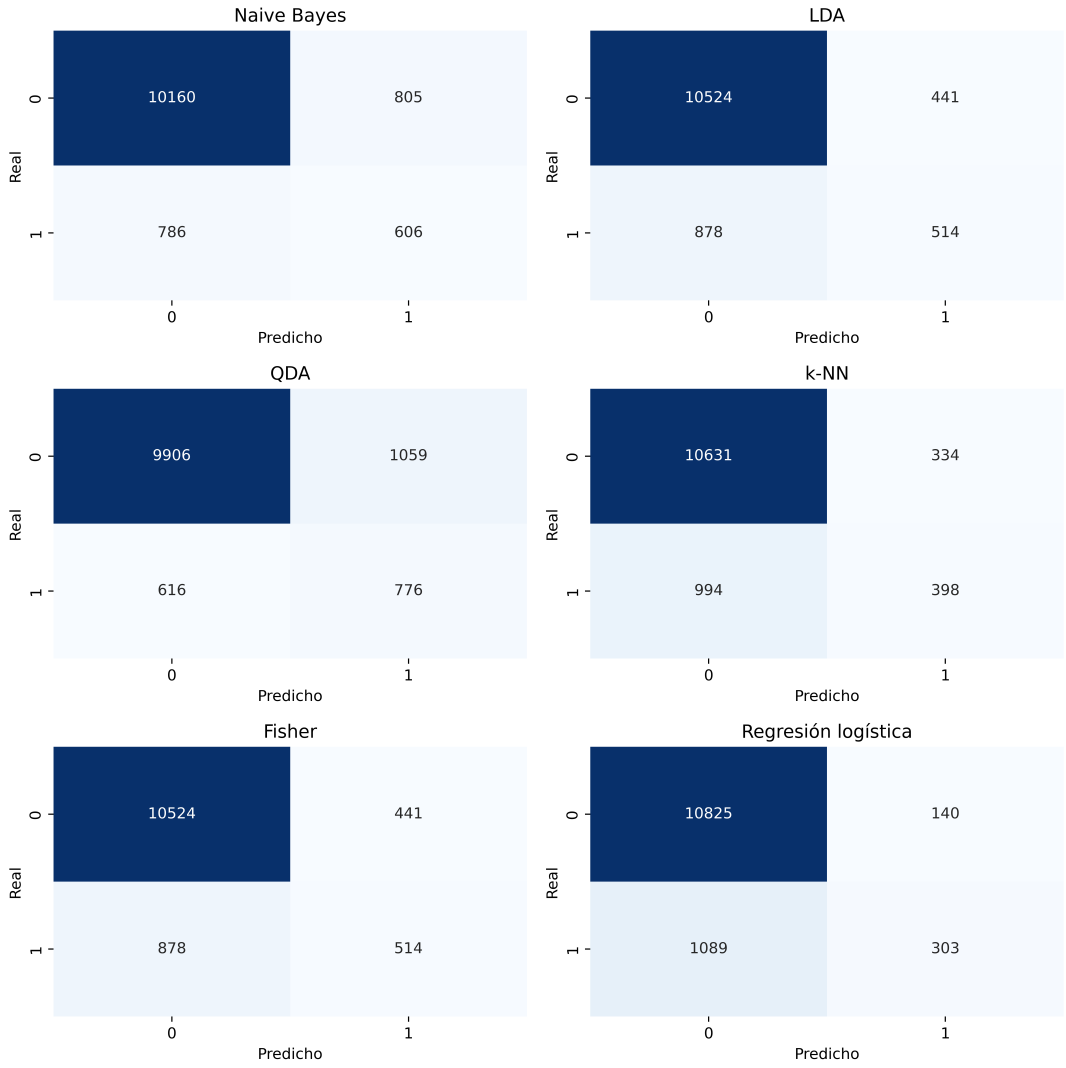


Figura 5: Matrices de confusión de los modelos entrenados sin la variable "default"

Conclusiones

Los modelos de clasificación entrenados mostraron un buen desempeño en general, a excepción de la regresión logística, la cual realizó una mala clasificación de la clase "sí", pero una muy buena de la clase "no". La elección de uno de estos modelos deberá realizarse en base al objetivo específico del banco, es decir, si tienen más interés en identificar aquellos clientes que no harán un depósito o si en identificar a aquellos que sí. Asimismo, ha de considerarse qué tanto están dispuestos a equivocarse con una de las clases anteriores.

Un punto importante a mencionar es que la variable "default" parecía inducir mucho ruido o sesgo en nuestra clasificación, pero al eliminarla no se observaron cambios significativos. Por otro lado, internamente se trato de eliminar otras variables que parecieran no tener mucha importancia para la variable de respuesta, pero al eliminarlas la clasificación empeoraba.

Como trabajo futuro sería interesante eliminar cuidadosamente aquellas variables que estén altamente relacionadas con otras y verificar si los tiempos (que ya son muy bajos) de ejecución mejoran o si hay algún cambio notorio en la clasificación. Para la eliminación de variables (y a su vez mayor entendimiento de ellas) puede hacerse uso también de los *weight of evidence* (ver [3, p. 192-193]), los cuales nos muestran cómo se comporta la variable objetivo a través de las distintas clases de las variables categóricas (y también con variables numéricas, pero estas hay que discretizarlas en grupos).

Una dificultad de este conjunto de datos es la alta dimensionalidad, la cual nos limita la visualización

de cómo algunas variables pueden estar relacionadas con la clasificación, así como el gran número de variables categóricas. En este caso la proyección de Fisher fue útil para visualizar qué tan "separados" se encontraban los datos. En conjuntos de datos de dimensión pequeña con variables de naturaleza numérica es más fácil analizar gráficamente cómo las variables afectan a la variable objetivo.

Trabajar con datos cuya variable objetivo tiene una desproporción significativa afecta el desempeño de los clasificadores. Encontrar técnicas adecuadas para disminuir dicho efecto es importante, por ejemplo, en el caso de la regresión logística pueden asignarse pesos a ambas clases (ver [2]) y así mejorar la clasificación de la menor clase, aunque empeoraría la de la mayor de ellas.

Referencias

[1] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

[2] King, G., & Zeng, L. (2001). *Logistic regression in rare events data*. Political analysis, 9(2), 137-163.

[3] Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford university press.

[4] S. Moro, P. Cortez & P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001