Proyecto 1: Clasificación supervisada (caso práctico)
Introducción a la Ciencia de Datos

Integrantes:	Avendaño Caballero, Joksan; Rodríguez Villagrán, Juan Pablo
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas
Profesor:	Dr. Marco Antonio Aquino López

Resumen

En este trabajo se presentan, a través de un ejemplo aplicado, algunos de los clasificadores supervisados más comunes: Naive Bayes, discriminantes lineal y cuadrático, k vecinos más cercanos y regresión logística. Se muestra que, en términos generales, tienen desempeños similares. Asimismo, se aborda el problema de desbalanceo de clases, ilustrando su impacto y explorando estrategias de tratamiento mediante submuestreo y regresión logística balanceada.
Los datos utilizados provienen de campañas de telemarketing de una institución financiera, y las decisiones de limpieza se justifican en función del mecanismo mediante el cual fueron recabados.

1. Introducción

En los avances recientes de la estadística matemática, dos problemas han orientado de manera significativa la investigación y el desarrollo de algoritmos, como señala Breiman [Bre+84]: el problema de la *regresión* y el de la *clasificación*.

El problema de clasificación especialmente es relevante porque permite proponer modelos estadísticos con niveles bien definidos, lo cual facilita la toma de decisiones. Una analogía ilustrativa es el diagnóstico médico: a partir de ciertos criterios, se determina si un paciente presenta o no una enfermedad. De manera similar, en tareas cotidianas como decidir la madurez de una fruta, se consideran (consciente o inconscientemente) distintos indicadores antes de tomar una decisión.

En términos estadísticos, la clasificación consiste en asignar a un objeto, a partir de sus características observadas, la clase a la que probablemente pertenece dentro de un conjunto predefinido de categorías.

Los datos de esta naturaleza suelen presentar complicaciones intrínsecas, como el *desbalanceo de clases*, que ocurre cuando algunas categorías son mucho más frecuentes que otras. En el ámbito médico, por ejemplo, es habitual que los registros de pacientes sanos superen a los de pacientes enfermos.

Para abordar el problema de clasificación en el contexto *supervisado* —es decir, cuando los datos incluyen la clase a la que pertenece cada observación— se han desarrollado diversas herramientas estadísticas. Entre ellas destacan el *clasificador de Bayes ingenuo (Naive Bayes)*, los discriminantes lineal y cuadrático, el *clasificador de los k vecinos más cercanos* y la *regresión logística*. Una exposición amplia de estos métodos puede encontrarse en [HTF09]

Gracias a las herramientas de cómputo actuales, la implementación y ejecución de estos clasificadores se ha simplificado y es razonable aplicarlos a grandes volúmenes de datos. El presente trabajo tiene como propósito ilustrar la utilidad de estos modelos en el análisis de la base de datos de marketing bancario (*Bank Marketing*), en el contexto de las campañas de telemarketing de una institución financiera.

Los principales objetivos de este trabajo son:

- Identificar y describir los problemas de limpieza de datos que suelen surgir en contextos de clasificación, así como proponer estrategias para su tratamiento.
- Implementar, presentar y comparar distintos clasificadores supervisados, evaluando su desempeño sobre la base de datos de marketing bancarios.

2. Descripción y exploración inicial de los datos

La base de datos [MRC14] surge como parte de un proyecto conjunto del Instituto Universitário de Lisboa y la Universidad de Minho, este primer producto se puede revisar en [MCR19]. El objetivo principal es utilizar técnicas de minería de datos de datos para evaluar qué tan efectivas fueron las campañas de telemarketing de un banco portugués, en el periodo de mayo del 2008 hasta junio del 2013. La metodología de recabación de datos fue la siguiente: un agente humano llamó a una lista de clientes para vender el servicio, o, si el cliente fue quien llamó, se ofrece el servicio. Así, teniendo una serie de características de cada cliente, el resultado final es binario: contacto exitoso o fallido.

Al examinar la base de datos, llama la atención que hay dos juegos de datos distintos: `bank` y `bank-additional`. Cada uno cuenta con una versión corta con el 10 % del total de datos y otra `full`. Para este trabajo se utiliza `bank-additional`, que es el conjunto de datos descrito en [MCR19]. La diferencia principal entre `bank` y `bank-additional` es que el primero incluye la variable de *balance* (información sensible), mientras que en `bank-additional` fue sustituida por indicadores socioeconómicos para preservar la privacidad.

El estudio original recabó alrededor de 150 variables, pero como se describe en [MCR19], tras un proceso de selección de variables se decidió utilizar únicamente dos juegos de características: un juego de 22 variables sobre los agentes de telefonía, las cuales se describen en el artículo, y un juego de 21 variables sobre los clientes. Para este trabajo, se utilizan las variables de los clientes, mostrados en la tabla 1, conformado de 41188 registros, para determinar qué características influyen en que un cliente contrate o no el servicio del banco.

Atributo	**Descripción**	**Tipo**	**Posibles valores**
age	Edad del cliente	Numérica	Valores reales (años)
job	Tipo de empleo	Categórica	admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown
marital	Estado civil	Categórica	divorced, married, single, unknown
education	Nivel educativo	Categórica	basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown
default	¿Tiene créditos en incumplimiento?	Categórica	no, yes, unknown
housing	¿Tiene crédito hipotecario?	Categórica	no, yes, unknown
loan	¿Tiene préstamo personal?	Categórica	no, yes, unknown
contact	Medio de contacto	Categórica	cellular, telephone
month	Mes del último contacto	Categórica	jan, feb, mar, ..., nov, dec
day_of_week	Día de la semana del último contacto	Categórica	mon, tue, wed, thu, fri
duration	Duración del último contacto (segundos)	Numérica	Valores enteros ≥ 0 (nota: solo válida ex post, no debe usarse en modelos predictivos realistas)
campaign	Número de contactos en esta campaña (incluye el último)	Numérica	Valores enteros ≥ 1
pdays	Días desde el último contacto en campaña previa	Numérica	Enteros; 999 indica que no fue contactado antes
previous	Número de contactos antes de esta campaña	Numérica	Valores enteros ≥ 0
poutcome	Resultado de la campaña previa	Categórica	failure, nonexistent, success
emp.var.rate	Tasa de variación del empleo (trimestral)	Numérica	Valores reales
cons.price.idx	Índice de precios al consumidor (mensual)	Numérica	Valores reales
cons.conf.idx	Índice de confianza del consumidor (mensual)	Numérica	Valores reales
euribor3m	Tasa Euribor a 3 meses (diaria)	Numérica	Valores reales
nr.employed	Número de empleados (trimestral)	Numérica	Valores reales
y	¿El cliente contrató un depósito a plazo?	Binaria	yes, no

Cuadro 1: Detalle de variables del conjunto de datos Bank Marketing

3. Detección de problemas en los datos

Antes de proceder al modelado, es fundamental identificar problemas potenciales en la base de datos. En particular, se detectaron dos situaciones principales: el desbalanceo de clases y la presencia de datos faltantes. En el resumen de las variables categóricas presentado en la tabla 2 se pueden ver ambas situaciones.

Atributo	**Conteo por categoría**
job	admin.: 10422 blue-collar: 9254 technician: 6743 services: 3969 management: 2924 retired: 1720 entrepreneur: 1456 self-employed: 1421 housemaid: 1060 unemployed: 1014 student: 875 unknown: 330
marital	married: 24928 single: 11568 divorced: 4612 unknown: 80
education	university.degree: 12168 high.school: 9515 basic.9y: 6045 professional.course: 5243 basic.4y: 4176 basic.6y: 2292 unknown: 1731 illiterate: 18
default	no: 32588 unknown: 8597 yes: 3
housing	yes: 21576 no: 18622 unknown: 990
loan	no: 33950 yes: 6248 unknown: 990
contact	cellular: 26144 telephone: 15044
month	may: 13769 jul: 7174 aug: 6178 jun: 5318 nov: 4101 apr: 2632 oct: 718 sep: 570 mar: 546 dec: 182
day_of_week	thu: 8623 mon: 8514 wed: 8134 tue: 8090 fri: 7827
poutcome	nonexistent: 35563 failure: 4252 success: 1373
y	no: 36548 yes: 4640

Cuadro 2: Distribución de variables categóricas del conjunto Bank Marketing

3.1. Desbalanceo de clases

De los conteos en la variable de respuesta y, es inmediato notar que la cantidad de personas que no toman el servicio es nueve veces mayor que la cantidad de las que sí, es decir, están a un ratio de 1:9. Esto indica que hay mucha más información de las personas que no toman el servicio que de las que sí lo toman, lo que hará que sea muy fácil clasificar a quienes dirán que no. Algunas propuestas para reducir este sesgo, las cuales se pueden revisar con más detalle en [Cha+02] o en [HG09], son las siguientes

- *Submuestreo de la clase predominante.* La idea tras esta técnica es a partir de la muestra mayoritaria obtener una submuestra de un orden comparable al de la minoritaria. La principal ventaja es que reduce el sesgo hacia la clase mayoritaria y entrena más rápido. Un riesgo que se corre es la pérdida de información si la clase mayoritaria es muy heterogénea. Una manera de disminuir este riesgo es no hacer las poblaciones del mismo orden pero sí hacer que el desbalanceo sea menor, usualmente a ratios de 2:1 o 3:1. Una regla empírica para el submuestreo es hacerlo cuando la clase mayoritaria es muy grande, del orden de 10^3, y homogénea.

- *Sobremuestreo de la clase que escasea.* Contraria a la técnica anterior, en ésta se incrementa artificialmente el tamaño de la muestra minoritaria. La manera más básica de hacer esto es replicando observaciones, pero en [Cha+02] se presenta SMOTE, una propuesta que genera puntos sintéticos a partir de interpolaciones aleatorias. La principal ventaja de este método es que no se pierde información, aunque esto puede generar ruido y sobreajuste si se sobreduplica demasiado un mismo patrón. No es recomendable exceder el tamaño de la clase mayoritaria, pues esto introduce sobreajuste.

- *Una combinación de los dos métodos anteriores.* La complicación que se puede tener en esta técnica es decidir qué tanto aplicar cada método. La ventaja es que permite manejar datos heterogéneos sin perder demasiada información.

- *Clasificación pesada.* Si el método de clasificación que se utiliza surge de una función de pérdida, ésta se puede modificar agregando pesos que penalicen más los errores en la clase minoritaria. Las principales ventajas de este método son que los datos no se modifican y es un método fácil de implementar. La parte en la que se debe tener cuidado es en la elección de pesos. Una elección común es asignar pesos proporcionales al inverso de la frecuencia de cada clase.

Ultimadamente, en este trabajo se opta por hacer submuestreo de la clase predominante utilizando un ratio de 2:1. De la tabla 2 se puede notar que, salvo por la columna de *loan*, los conteos están repartidos de manera aproximadamente uniforme, por lo que se puede tomar la hipótesis de población aproximadamente homogénea en la clase. Adicionalmente, para el clasificador de regresión logística se utiliza la clasificación pesada con el tamaño de cada clase.

3.2. Datos faltantes

En la tabla 2 se aprecia que hay seis variables que presentan datos *unknown*. Porcentualmente, los datos con *unknown* se distribuyen de la siguiente manera, en la tabla 3. Notemos que situaciones como la de la variable *default* plantean un reto adicional.

Columna	**No. de faltantes**	**Porcentaje**
job	330	0.80 %
marital	80	0.19 %
education	1731	4.20 %
default	8597	20.87 %
housing	990	2.40 %
loan	990	2.40 %
contact	0	0.00 %
month	0	0.00 %
day_of_week	0	0.00 %
poutcome	0	0.00 %

Cuadro 3: Valores faltantes en variables categóricas del conjunto Bank Marketing

El tratamiento que se da a cada uno de ellos es dependiente del caso, a continuación se dan las consideraciones hechas sobre estos datos faltantes.

- **job** y **marital**. Porcentualmente sus valores faltantes son muy pocos. Se decide que faltan debido a un mecanismo MAR, ya que las personas podrían no responder debido a otras variables, como lo es la edad o alguna situación particular. Para no descartar renglones del conjunto de datos, se decide *imputar* datos para eliminar los unknowns. El mecanismo de imputación elegido es imputar la moda.
- **education**. Probablemente por algún estigma social, las personas con menor nivel educativo prefieran no hablarlo. Se decide que estos datos faltan debido a un mecanismo MNAR. Se decide por tratar los faltantes como una propia clase de *unknown*.
- **default**. Esta variable surge de la pregunta de incumplimientos de crédito. Las personas con mal historial crediticio tienen más probabilidad de no reportar este dato. Se decide que estos datos faltan debido a un mecanismo MNAR. Se decide por tratar los faltantes como una propia clase de *unknown*, además de que, al ser un 20.87 % de los datos se perdería una cantidad importante de información.
- **housing**. Esta variable surge de la pregunta sobre préstamos hipotecarios. Por la misma razón de la variable anterior, se decide que faltan debido a un mecanismo MNAR y se trata *unknown* como su propia clase.
- **loan**. Esta variable surge de la pregunta sobre préstamos personales, así que se decide exactamente lo mismo que en **housing**.

3.3. Otras consideraciones

Para este trabajo se consideró, como se explicó en la sección 2, que las variables de la base de datos *ya pasó por un proceso de selección de variables*. Por ello, se decidió no hacer otro proceso de selección de variables.

Sobre la codificación de los datos, por un lado, la manera de registrar los datos faltantes computacionalmente es difícil de manejar ya que típicamente el software estadístico espera que éstos estén registrados como `NaN` o parecidos. El manejo de esta situación ya se explicó previamente. Por otro lado, la variable `pdays` registra un valor de 999 cuando no se había llamado antes, es decir, un *código de censura*. Este tipo de codificación puede distorsionar distribuciones numéricas. Una alternativa, que no se explora en este trabajo, para atender este problema es definir una variable dicotómica que responda a la pregunta de si se llamó antes o no.

No se hizo proceso alguno de detección de *outliers*. Las variables numéricas y estadísticas numéricas sugieren comportamientos aproximadamente regulares en éstas. En lo que concierne a variables categóricas, no existe en sí un concepto de *outlier*, pero sí el de categorías "raras" o "atípicas". Un estudio más cuidadoso de este conjunto de datos atendería especialmente variables con potenciales outliers contextuales como duration o categorías atípicas como los que respondieron *sí* en *default*.

4. Procesamiento de datos

Los métodos de clasificación utilizados son: Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-Nearest Neighbors y regresión logística. Una exposición detallada de estos métodos se puede revisar en [HTF09] o en la parte 2 de este estudio en la que se someten los clasificadores a datos simulados con estructuras diversas para probarlos. Para elegir el valor de k, se utiliza un procedimiento de validación cruzada 5-fold con valores de 1 hasta mín$\{100, \lfloor\sqrt{n}\rfloor\}$, donde k se elige a partir del valor que maximice el parámetro del F-score sobre el conjunto de entrenamiento. El rango de iteración es con el fin de suavizar el ruido pero sin diluir la influencia de la minoría al ser un desbalance considerable; mientras que el criterio del F-score se utiliza ya que contiene información de la precisión y sensibilidad de predicción.

Para estudiar el desempeño de los clasificadores, se separa el conjunto de datos en un conjunto de entrenamiento (70 % de los datos) y uno de prueba (30 % de los datos), así las observaciones se distribuyen como: 28831 para entrenamiento y 12357 para prueba. Al evaluar el modelo entrenado en el conjunto de prueba se obtienen las métricas de desempeño del modelo.

4.1. Caso desbalanceado

Aquí la única limpieza hecha es la imputación de la moda en los faltantes de las columnas **job** y **marital**. Para la elección del valor de k usa la validación cruzada con valores del 1 hasta 100, con el criterio de maximizar el F-score, véase figura 1. Algunas consideraciones que se tuvieron son

- Quadratic Discriminant Analysis. Al ejecutar el algoritmo, se tiene problemas de colinealidad, así que se agrega un factor de regularización con un peso de 0.1 y con ello mitigar el problema.
- k-Nearest Neighbors. El valor óptimo de k bajo el criterio de validación cruzada maximizando el indicador $F\text{-}score$ es $k = 41$.

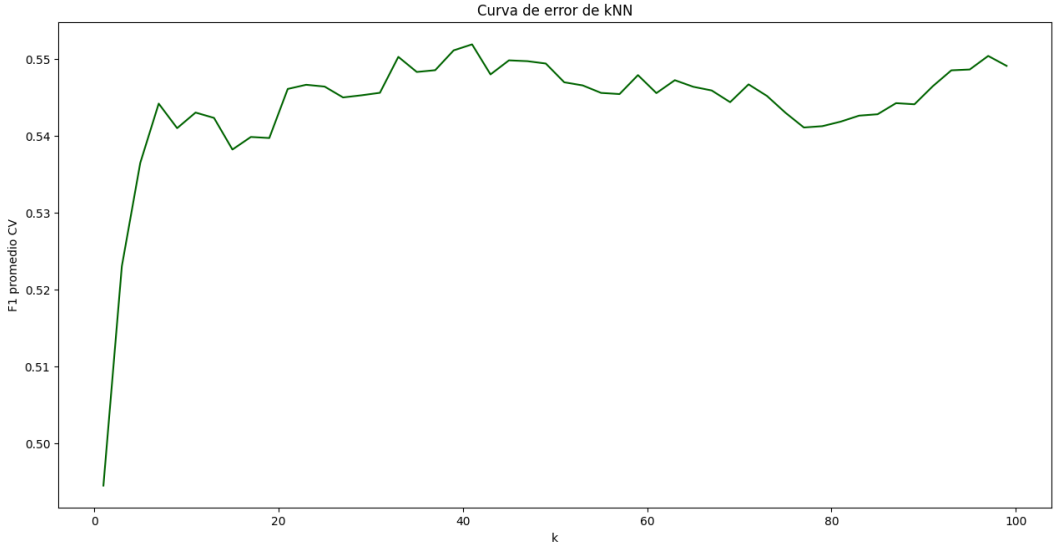

Figura 1: Curva de error del método k-NN tomando como score a F_1. En el rango de valores considerado, el máximo se alcanza en $k = 41$.

En la tabla 4 se presentan simultáneamente las métricas de desempeño de los clasificadores. De aquí se puede notar que todos tienen desempeños muy parecidos, siendo algunos mejores a los demás respecto a determinados criterios.

Modelo	**Acc**	**Precisión**	**Recall**	**Especificidad**	**F1**	**AUC**
Naive Bayes	0.875	0.880	0.875	0.924	0.877	0.847
LDA	0.911	0.904	0.911	0.963	0.906	0.936
QDA	0.885	0.896	0.885	0.921	0.889	0.911
k-NN ($k = 41$)	0.911	0.903	0.911	0.967	0.905	0.927
Reg. Logística (pesos=None)	0.909	0.897	0.909	0.976	0.898	0.931

Cuadro 4: Comparación de métricas de desempeño de los clasificadores utilizados.

A modo de complemento, en la tabla 5 se presentan las matrices de confusión. Éstas sugieren que el conjunto de prueba también está compuesto por clases desbalanceadas, lo que favorece que los clasificadores clasifiquen bien a la clase predominante, justificando intuitivamente los valores de rendimiento presentados en la tabla 4.

Naive Bayes

10127	838
711	681

LDA

10558	407
693	699

QDA

10096	869
558	834

k-NN ($k = 41$)

10601	364
730	662

Reg. Logística

10703	262
857	535

Cuadro 5: Matrices de confusión de los clasificadores aplicados.

Como indicador robusto de la precisión, se presenta en la tabla 6 la *accuracy* dada con el método de validación cruzada. Esta tabla refuerza la idea de que los clasificadores tienen desempeños comparables.

Modelo	**Accuracy medio**	**Desv. Est.**
Naive Bayes	0.874	0.002
LDA	0.910	0.001
QDA	0.882	0.004
k-NN ($k = 41$)	0.912	0.003
Regresión Logística (pesos=None)	0.910	0.002

Cuadro 6: Accuracy promedio de cada clasificador en el caso desbalanceado utilizando validación cruzada 5-fold.

4.2. Caso balanceado

Antes de proceder al submuestreo, se revisa la regresión logística pesada tomando los pesos *inversos al tamaño de cada clase*. En las tablas 7, 8 y 9 se presentan indicadores del desempeño de este clasificador.

Modelo	**Acc**	**Precisión**	**Recall**	**Especificidad**	**F1**	**AUC**
Reg. Logística (pesos="balanced")	0.862	0.924	0.862	0.857	0.881	0.939

Cuadro 7: Métricas de desempeño del clasificador de regresión logística con pesos balanceados.

Reg. Logística	
9402	1563
142	1250

Cuadro 8: Matriz de confusión de la regresión logística balanceada.

Modelo	**Accuracy medio**	**Desv. Est.**
Regresión Logística (pesos="balanced")	0.860	0.003

Cuadro 9: Accuracy promedio de la regresión logística balanceada utilizando validación cruzada 5-fold.

En general, el rendimiento no es tan alto como en el caso anterior. El problema que presenta este método radica de nuevo en el ratio de desbalanceo tan alto que se tiene, así los pesos asignados a la clase mayoritaria son considerablemente menores.

Aquí se hizo la imputación de moda y el submuestreo con ratio objetivo de 2:1. Se utiliza el ratio 2:1 porque, aunque originalmente haya un sesgo debido al ratio 9:1, el desbalanceo es parte de la estructura intrínseca del problema. El tamaño de la clase mayoritaria es ahora de 6496 observaciones. Para la elección del valor de k usa la validación cruzada con valores del 1 hasta 100, con el criterio de maximizar el F-score, lo que gráficamente se puede ver en la figura 2. Se tuvieron las mismas consideraciones que en el caso desbalanceado

- Quadratic Discriminant Analysis. Al ejecutar el algoritmo, se tiene problemas de colinealidad, así que se agrega un factor de regularización con un peso de 0.1 y con ello mitigar el problema.
- k-Nearest Neighbors. El valor óptimo de k bajo el criterio de validación cruzada maximizando el indicador F-*score* es $k = 27$.

En la tabla 10 se presentan simultáneamente las métricas de desempeño de los clasificadores. De aquí se puede notar que todos tienen desempeños muy parecidos, siendo algunos mejores a los demás respecto a determinados criterios.
A modo de complemento, en la tabla 11 se presentan las matrices de confusión. Éstas sugieren que el conjunto de prueba también está compuesto por clases desbalanceadas, lo que favorece que los clasificadores clasifiquen bien a la clase predominante, justificando intuitivamente los valores de rendimiento presentados en la tabla 10.

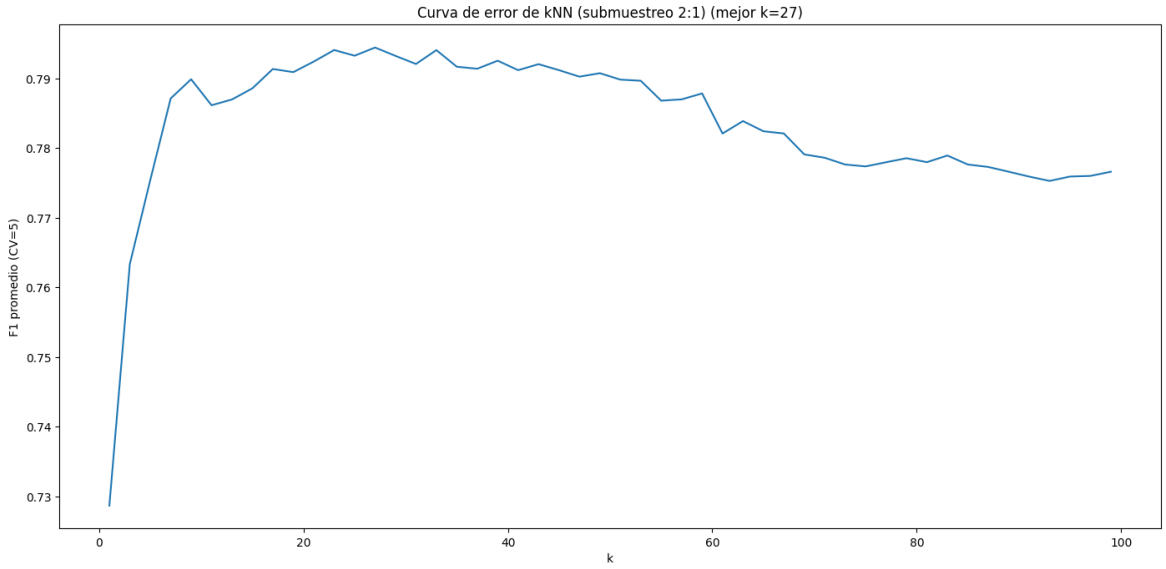

Figura 2: Curva de error del método k-NN tomando como score a F_1. En el rango de valores considerado, el máximo se alcanza en $k = 27$.

Modelo	**Acc**	**Precisión**	**Recall**	**Especificidad**	**F1**	**AUC**
Naive Bayes (2:1)	0.871	0.880	0.871	0.918	0.875	0.848
LDA	0.893	0.916	0.893	0.910	0.901	0.937
QDA	0.884	0.904	0.884	0.912	0.892	0.911
k-NN ($k = 27$)	0.880	0.918	0.880	0.911	0.905	0.929
Reg. Logística (pesos=None)	0.897	0.921	0.897	0.911	0.905	0.938
Reg. Logística (pesos="balanced")	0.862	0.924	0.862	0.858	0.880	0.938

Cuadro 10: Comparación de métricas de desempeño de los clasificadores utilizados.

Naive Bayes

10067	898
693	699

LDA

9980	985
343	1049

QDA

9996	969
460	932

k-NN ($k = 27$)

9993	972
305	1087

Reg. Logística

10703	262
857	535

Reg. Logística balanced

9406	1559
146	1246

Cuadro 11: Matrices de confusión de los clasificadores aplicados.

5. Conclusiones y discusión

Primeramente, no hay manera única de tratar o de definir un comportamiento "atípico" cuando se trabajan datos categóricos. En este ejemplo desarrollado, al tratarse de un conjunto de datos obtenidos por una encuesta, se aprecia que es un comportamiento típico que los datos falten de manera *no aleatoria* al dar la opción a la población de *no responder*. El problema del desbalanceo también resulta natural para los datos, además de que el sesgo hacia la clase mayoritaria permite identificar con más facilidad *cuándo no van a tomar el servicio*, aunque no ocurra lo mismo con *cuándo sí*.

En el caso desbalanceado, las métricas de desempeño y las matrices de confusión ilustran que los clasificadores funcionan muy bien para el conjunto de entrenamiento de el conjunto de datos. Así, los clasificadores dan resultados aproximadamente equivalentes con una precisión muy parecida. Esto puede deberse a, entre otras

razones, la homogeneidad de los datos debida al desbalanceo.

En el primer caso balanceado, con un ratio 9:1, la ponderación desplaza la pérdida hacia la clase minoritaria, lo que típicamente reducce *accuracy* en favor de mayor sensibilidad a *yes*. El caso balanceado obtenido por submuestreo muestra un desempeño muy parecido al caso desbalanceado lo cual permite argumentar a favor del uso del caso completamente desbalanceado –que es preferible porque usa toda la información–. Aunque sí se aprecia sensibilidad de los clasificadores ante este preprocesamiento de los datos. Esto no es de sorprender ya que es en sí un cambio estructural en la composición de los datos.

Dentro de esto es importante mencionar las complicaciones de trabajar con datos categóricos. En trabajos previos, al justificar que datos son outliers, se depende de alguna noción de distancia, pero ésta es complicada de definir de una manera contextualmente razonable para datos categóricos.

En atención a comentarios hechos a lo largo del trabajo, podría ser relevante el efecto de categorizar variables con *valores de censura* como es el caso de la variable *pdays*. También podrían tomarse decisiones a partir de un proceso de selección de variables con el fin de medir la influencia de cada variable en la respuesta final. Finalmente, un área de oportunidad que no se mencionó, podría utilizarse la técnica de las curvas ROC como otro mecanismo de evaluación del desempeño de los clasificadores.

Referencias

[Bre+84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen y Charles J. Stone. *Classification and Regression Trees*. Chapman Hall, 1984.

[Cha+02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall y W Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". En: *Journal of Artificial Intelligence Research* 16 (2002), págs. 321-357.

[HG09] Haibo He y Edwardo A Garcia. "Learning from imbalanced data". En: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), págs. 1263-1284.

[HTF09] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[MCR19] Sérgio Moro, Paulo Cortez y Paulo Rita. "A Data-Driven Approach to Predict the Success of Bank Telemarketing". En: (2019). DOI: dx.doi.org/10.1016/j.dss.2014.03.001.

[MRC14] Sérgio Moro, P. Rita y P. Cortez. *Bank Marketing*. 2014. DOI: 10.5880/GFZ.4.3.2023.002. URL: https://doi.org/10.24432/C5K306.