Proyecto 2: Clasificación Supervisada
Introducción a la Ciencia de Datos

Integrantes:	Sánchez, Hazel; Hernández, Debany ; Canché, Elías
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas (CIMAT)
Profesor:	Dr. Marco Antonio Aquino López

1. Introducción

En el marketing digital y la gestión de relaciones con el cliente, la capacidad de predecir comportamientos del consumidor es una herramienta importante para las empresas. A diferencia del marketing convencional, el marketing digital permite una interacción directa y personalizada con los consumidores facilitando la promoción de productos y servicios, además de la recolección de datos sobre el comportamiento y las preferencias del público objetivo. En un entorno comercial cada vez más competido, esta capacidad de capturar y analizar información se traduce en ventajas de mercado.

La Ciencia de Datos, con técnicas de clasificación supervisada, proporciona herramientas para identificar patrones ocultos, predecir comportamientos futuros y optimizar recursos, también es posible obtener modelos predictivos que identifican clientes con mayor probabilidad de responder positivamente a una campaña, lo cual podría maximizar el retorno de inversión y fortalecer las estrategias de fidelización del cliente.

En este trabajo aplicamos diversos métodos de clasificación supervisada sobre la base de datos "Bank Marketing", la cual documenta campañas de marketing directo realizadas por un banco portugués entre 2008 y 2013, cuyo objetivo fundamental fue identificar clientes propensos a suscribir depósitos a plazo. Además de la implementación técnica, proponemos examinamos cómo la naturaleza de los datos influye en el desempeño de cada clasificador.

2. Exploración inicial de los datos

2.1. *Bank Marketing Dataset.*

El **Bank Marketing Dataset** (`bank-full.csv`) fue desarrollado y publicado por el Instituto de Sistemas e Informática (INESC) de la Universidad de Lisboa, Portugal, como parte de un estudio longitudinal sobre efectividad de campañas de marketing en el sector bancario. Los datos fueron recolectados entre mayo de 2008 y noviembre de 2013 a través de campañas de marketing telefónico realizadas por una institución bancaria portuguesa y contiene información sobre clientes y el resultado de campañas de marketing telefónicas.

La estructura general de la base de datos consta de 45,211 observaciones, 16 variables predictoras y 1 variable de respuesta que indica si el cliente aceptó (`yes`) o no (`no`) contratar el depósito a plazo. En la tabla [1] se describen a detalle las variables, mientras que en las tablas [2] y [3] se muestra un resumen sobre la información de la base de datos respecto a cada variable.

El análisis exploratorio reveló un problema de desbalance en la variable de respuesta, teniendo 88.30 % de "no" contra el 11.7 % de "yes"; dado que es un balance moderado, no se emplearán técnicas para balancear pero en la evaluación de la clasificación se utilizarán métricas como 'precision', 'recall', 'f1-score', 'roc_auc' y no 'accuracy'.

No se encontraron registros repetidos y respecto a los **valores faltantes**, no se detectaron celdas con `NaN`. Sin embargo, algunas variables categóricas presentan la categoría `unknown`, la cual representa información faltante, los detalles de "unknown" por categoría se puede ver en la tabla [4]. Dado que el porcentaje de 'unknown' en la variable job es bajo y no hay un patrón claro que justifique la falta de información entonces pueden ser considerados MCAR. Por su parte, en la variable contact (medio de contacto) no puede faltar aleatoriamente, ya que si un cliente es contactado entonces se debe conocer el medio, por lo tanto esta variable tiene datos MNAR. De la misma manera en education, son MNAR pues algunas veces las personas prefieren no decir su

Cuadro 1: Clasificación de Variables Predictoras del Bank Marketing Dataset

Variable	**Tipo**	**Escala**	**Descripción y Características**
age	Numérica	Continua	Edad del cliente (17-98 años).
job	Categórica	Nominal	Ocupación 12 categorías.
marital	Categórica	Nominal	Estado civil: 'divorced', 'married', 'single', 'unknown'.
education	Categórica	Ordinal	Nivel educativo: 'unknown', 'secondary', 'primary', 'tertiary'.
default	Categórica	Ordinal	Crédito en mora: 'no', 'yes', 'unknown'.
balance	Numérica	Discreta	Saldo promedio anual
housing	Categórica	Ordinal	Préstamo hipotecario: 'no', 'yes', 'unknown'.
loan	Categórica	Ordinal	Préstamo personal: 'no', 'yes', 'unknown'.
contact	Categórica	Nominal	Tipo de contacto: 'cellular', 'telephone'.
day	Numérica	Discreta	Día del mes último contacto.
month	Categórica	Ordinal	Mes último contacto: 'jan' a 'dec'.
duration	Numérica	Continua	Duración de contacto (segundos).
campaign	Numérica	Discreta	Número de contactos en campaña actual.
pdays	Numérica	Discreta	Días desde último contacto.
previous	Numérica	Discreta	Número de contactos anteriores.
poutcome	Categórica	Nominal	Resultado campaña anterior: 'failure', 'nonexistent', 'success'.

Cuadro 2: Estadísticas Descriptivas de Variables Numéricas

Variable	**Media**	**Desv. Est.**	**Mínimo**	**Máximo**
age	40.94	10.62	17	98
balance	1362.27	3044.77	-8019	102127
duration	258.28	259.28	0	4918
campaign	2.76	3.10	1	63
pdays	962.48	186.91	0	999
previous	0.17	0.49	0	7

Cuadro 3: Resumen Estadístico de Variables Categóricas

Variable	**Categoría**	**Conteo**	**Porcentaje**	**Variable**	**Categoría**	**Conteo**	**Porcentaje**
job	blue-collar	9732	21.53	marital	married	27214	60.19
job	management	9458	20.92	marital	single	12790	28.29
job	technician	7597	16.80	marital	divorced	5207	11.52
job	admin.	5171	11.44	education	secondary	23202	51.32
job	services	4154	9.19	education	tertiary	13301	29.42
job	retired	2264	5.01	education	primary	6851	15.15
job	self-employed	1579	3.49	education	unknown	1857	4.11
job	entrepreneur	1487	3.29	housing	yes	25130	55.58
job	unemployed	1303	2.88	housing	no	20081	44.42
job	housemaid	1240	2.74	month	may	13766	30.45
job	student	938	2.07	month	jul	6895	15.25
job	unknown	288	0.64	month	aug	6247	13.82
loan	yes	7244	16.02	month	jun	5341	11.81
loan	no	37967	83.98	month	nov	3970	8.78
contact	cellular	29285	64.77	month	apr	2932	6.49
contact	unknown	13020	28.80	month	feb	2649	5.86
contact	telephone	2906	6.43	month	jan	1403	3.10
poutcome	unknown	36959	81.75	month	oct	738	1.63
poutcome	failure	4901	10.84	month	sep	579	1.28
poutcome	other	1840	4.07	month	mar	477	1.06
poutcome	success	1511	3.34	month	dec	214	0.47
default	no	44396	98.20				
default	yes	815	1.80				

grado académico por no sentirse cómodos. Por último, poutcome tiene datos faltantes MNAR pues al considerar a las campañas, independientes entre sí, entonces los clientes pueden ser la primera vez que participan, por lo que, en este sentido, pueden incluso considerarse 'nonexistent'.

Dada la discusión anterior, se toma la decisión de imputar los datos mediante la moda, para las variables job, education; mientras que para contact y poutcome se consideran categorías, pues pueden haber situaciones que no se describan con las categorías que ya existen y esto puede ser informativo para los modelos de clasificación.

Cuadro 4: **Unknown** por variable

Variable	Conteo	Porcentaje	Variable	Conteo	Porcentaje
job	288	0.637013	education	1857	4.107407
contact	13020	28.798301	poutcome	36959	81.747805

3. Preprocesamiento General

El preprocesamiento tiene como finalidad transformar la base de datos **Bank Marketing** (`bank-full.csv` en una matriz de diseño adecuado para el modelado supervisado. En particular, buscamos

i) Representar de forma numérica las variables categóricas preservando su información,

ii) Homogenizar la escala de las variables numéricas para algoritmos sensibles a la magnitud,

iii) Separar correctamente predictores X y respuesta y evitando fugas de información.

Como paso inicial, se normalizó el formato de las variables categóricas (*trimming* de espacios y conversión a minúsculas) para evitar duplicidades por diferencias de uso de mayúsculas o espacios, mientras que la variable de respuesta `y` se codificó como binaria (0 = `no`, 1 = `yes`). También se revisaron las correlaciones de "y" con las variables numéricas, las cuales están en la tabla [5].

Cuadro 5: Correlaciones con "y"

Variable	Correlación	Variable	Correlación	Variable	Correlación
duration	0.394521	campaign	0.073172	day	0.028348
pdays	0.103621	balance	0.052838 age	0.025155	
previous	0.093236				

Por descripción de la variable 'duration', esta se obtiene después del contacto con el cliente, por lo tanto no tiene una verdadera capacidad predictiva y además, como se puede observar en la tabla [5], 'duration' tiene alta correlación con 'y' por lo que puede viciar el análisis, por lo tanto se toma la decisión que eliminarla.

Como último paso del procesamiento se adecuaron los datos para cada método:

- **Naive Bayes:** En este caso las numéricas no necesitaban un reescalamiento. Las categóricas nominales se codificaron utilizando One-Hot, mientras que para las categóricas ordinales se utilizó OrdinalEncoder.
- **LDA:** En este caso las numéricas sí necesitan reescalamiento, el cual fue StandardScaler. Las categóricas nominales se codificaron utilizando One-Hot, mientras que para las categóricas ordinales se utilizó OrdinalEncoder.
- **QDA**: El QDA no necesita reescalamiento en las numéricas por lo tanto la configuración se mantuvo igual que en Naive Bayes.
- **k-NN**: Este método necesita que todas las variables estén en la misma escala, por lo que las numéricas se reescalaron nuevamente.

Sobre la Codificación de variables categóricas

Las variables cualitativas (`job`, `marital`, `education`, `contact`, `month`, `poutcome`, entre otras) se trataron como nominales y se transformaron mediante **One-Hot Encoding** (creación de indicadores binarios por categoría). Esta estrategia nos ayuda a:

- Evitar imponer un orden arbitrario sobre categorías no ordinales.
- Permite a los modelos lineales asignar efectos específicos por categoría.
- Facilita la interpretación al nivel de categoría (coeficientes por indicador).

Para garantizar una matriz de diseño de *rango completo*, se utilizó la opción `drop_first=True`, eliminando una categoría de referencia por variable (conocido como *dummy variable trap*). En términos prácticos, si `marital` tiene categorías {`married`, `single`, `divorced`}, tras la codificación con referencia en `married` se generan dos columnas (`marital_single`, `marital_divorced`); un registro con ambos indicadores en 0 se interpreta como la categoría de referencia (`married`).

Escalado de variables numéricas

Las variables numéricas (`age`, `balance`, `day`, `duration`[1], `campaign`, `pdays`, `previous`, etc.) se estandarizaron con **StandardScaler**, es decir, cada columna x se transformó a:

$$z = \frac{x - \mu}{\sigma},$$

donde μ es la media y σ la desviación estándar de la variable. Esta normalización centra las variables en media 0 y varianza unitaria, lo que:

1. Evita que magnitudes grandes (`balance`) dominen sobre otras (`campaign`).
2. Mejora la estabilidad numérica y la convergencia en modelos como Regresión Logística y SVM.
3. Hace comparables los coeficientes (en modelos lineales) en términos de desviaciones estándar.

La interpretación de un valor estandarizado es directa: un $z = 1{,}2$ en `age` indica que el individuo está 1,2 desviaciones estándar *por encima* de la media de edad; un $z = -0{,}5$ en `balance` indica que está 0,5 desviaciones estándar *por debajo* de su media. X y longitud de y.

4. Modelado

En esta sección se muestran los resultados obtenidos para cada uno de los métodos en la tabla [6]

Modelo	Exactitud	Sensibilidad	Especificidad	Precisión	F1-Score	AUC-ROC
Naive Bayes	0.8368	0.4418	0.8910	0.3573	0.3951	0.7369
LDA	0.8889	0.2860	0.9716	0.5799	0.3831	0.7528
QDA	0.8669	0.3868	0.9327	0.4410	0.4121	0.7488
k-NN (k=50)	0.8871	0.1054	0.9943	0.7188	0.1839	0.7672

Cuadro 6: Resultados

De los resultados podemos observar que k-NN tiene el problemas de clasificación, k-NN está siendo conservador, predice "no" casi siempre,

- **Sensibilidad muy baja (0.1054):** Solo detecta el 10.5 % de los clientes que realmente subscribirán depósitos
- **Especificidad muy alta (0.9943):** Es modelo es excelente identificando quiénes NO subscribirán
- **Precisión alta (0.7188):** Cuando predice "sí", tiene un 72 % de acierto
- **AUC-ROC más alto (0.7672):** Tiene la mejor capacidad discriminativa general

[1]Recordemos que la variable `duration` se registra al final de la llamada y puede inducir fuga de información si se usa para predecir el éxito de la campaña antes o durante la llamada. Por ello, aunque se reporta en la exploración, suele *excluirse* como predictor en modelos destinados a la toma de decisión previa.

El desbalance de clases afecta significativamente. Todos los modelos muestran esta tendencia:

- **Alta especificidad** ($> 89\,\%$): Buenos identificando la clase mayoritaria
- **Baja sensibilidad** ($< 45\,\%$): Malos identificando la clase minoritaria

El clasificador Naive Bayes es el más balanceado,

- Mejor sensibilidad (0.4418)
- F1-Score más alto (0.3951)
- Buen equilibrio entre identificar ambas clases

LDA y QDA muestran patrones interesantes,

- **LDA:** Muy conservador (sensibilidad 0.2860) pero alta precisión
- **QDA:** Mejor balance que LDA pero peor que Naive Bayes

5. Conclusiones

Naive Bayes es el mejor modelo considerando el contexto de marketing bancario, donde identificar clientes potenciales (sensibilidad) es crucial, incluso a costa de algunos falsos positivos. k-NN necesita ajustes, su alta precisión es engañosa porque casi nunca predice la clase positiva.

Para una evaluación honesta del desempeño, el ajuste (*fit*) de transformaciones se realizó exclusivamente sobre el conjunto de entrenamiento y luego aplicarse (*transform*) al de validación/prueba. Esto incluye el cálculo de medias y desviaciones del escalado así como el mapeo de categorías en One-Hot. En la implementación, esto se logró de forma segura mediante *pipelines* que encadenan *encoder* y *scaler* con el estimador final. Asimismo, se recomienda evaluar la exclusión de `duration` para escenarios de predicción en tiempo real y considerar técnicas para manejar el desbalance de clases (métricas como F1/recall, *class weights* o remuestreo).