TWBANK

TAIWAN-BANK

Nonpayment of credit cards

Autores:
Erick Aguilar
Andrea Quero

CODERHOUSE

Contenido

1. Contexto y Objetivos

Resumen

Se dispone de un archivo que contiene información sobre factores demográficos, datos crediticios, historial de pago, estados de cuenta y si han caído o no en impago (default) clientes con tarjeta de crédito en Taiwan de Abril 2005 a Septiembre del 2005. Con este archivo se desea encontrar patrones en las personas / estados de cuenta que caen en impago

Objetivo

Se desea encontrar el perfil de los usuarios de tarjetas de crédito que tienen más probabilidades de caer en impago (default). Así mismo, se busca encontrar un patrón o tendencia que sugiera que un usuario está próximo a caer en impago.

Contexto comercial

La banca extiende tarjetas de crédito a un gran número de personas. Al cobrarles intereses y comisiones anuales generan sus ganancias. Sin embargo, si extendieran esas tarjetas a una gran cantidad de personas que no pudieran pagar las deudas adquiridas, entonces el banco podría enfrentar serios problemas de liquidez y gastos inesperados para tratar de recuperar su dinero.

2. Dataset Empleado

- Información demográfica y de estado de cuenta de clientes de TWBank.
- 30000 registros.
- 30 variables:
  - 3 categóricas
  - 27 numéricas,
    - 5 numéricas calculadas.

Columna	Tipo	Datos No-Nulos	Datos Nulos	Datos únicos	Ejemplo
ID	int64	30000	0	30000	1
LIMIT_BAL	float64	30000	0	81	20000
SEX	int64	30000	0	2	2
EDUCATION	int64	30000	0	4	2
MARRIAGE	int64	30000	0	3	1
AGE	int64	30000	0	56	24
PAY_0	int64	30000	0	9	2
PAY_2	int64	30000	0	9	2
PAY_3	int64	30000	0	9	-1
PAY_4	int64	30000	0	9	-1
PAY_5	int64	30000	0	8	-1
PAY_6	int64	30000	0	8	-1
BILL_AMT1	float64	30000	0	22722	3913
BILL_AMT2	float64	30000	0	22345	3102
BILL_AMT3	float64	30000	0	22025	689
BILL_AMT4	float64	30000	0	21547	0
BILL_AMT5	float64	30000	0	21009	0
BILL_AMT6	float64	30000	0	20603	0
PAY_AMT1	float64	30000	0	7942	0
PAY_AMT2	float64	30000	0	7898	689
PAY_AMT3	float64	30000	0	7517	0
PAY_AMT4	float64	30000	0	6936	0
PAY_AMT5	float64	30000	0	6896	0
PAY_AMT6	float64	30000	0	6938	0
default.payment.next.month	int64	30000	0	1	1
PercOfAprDebt	float64	30000	0	24873	0
PercOfMayDebt	float64	30000	0	25186	0
PercOfJunDebt	float64	30000	0	25474	0.03445
PercOfJulDebt	float64	30000	0	25729	0.1551
PercOfAugDebt	float64	30000	0	26225	0.1612

3. EDA: Exploratory Data Analysis

Hipótesis:

Variables Demográficas

I. El género afecta la probabilidad de caer más en impago

II. Las personas con un menor nivel de estudios tienden a caer más en impagos

III. Las personas casadas tienden a caer más en impago

IV. Las personas jóvenes tienden a caer más en impago

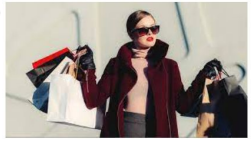
Variables históricas y financieras

V. Las personas cuyo saldo está más próximo a su límite de crédito tienden a caer más en impago

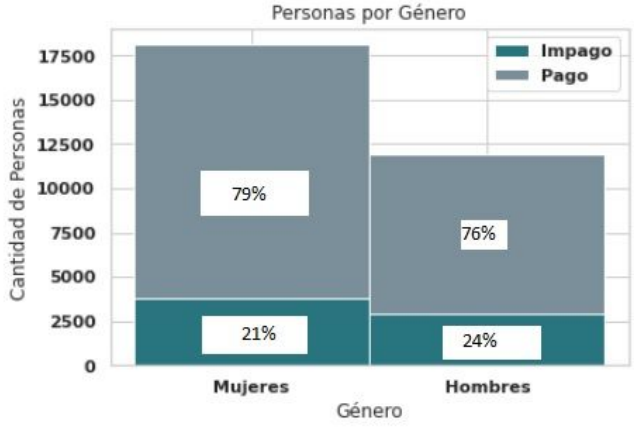
VI. Las personas que se han retrasado más en sus pagos tienden a caer más en impago

3. Análisis exploratorio de datos

I. ¿El género afecta la probabilidad de caer más en impago?

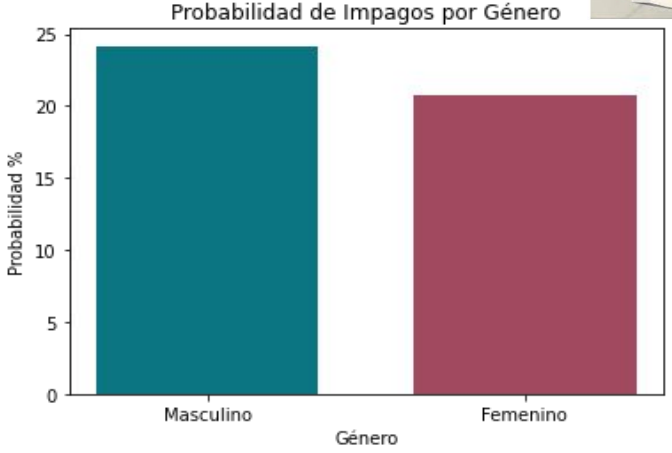


De las personas que cayeron en impago, 3,763 fueron del sexo Femenino, mientras que sólo 2,873 fueron del sexo Masculino. Entonces, ¿podemos concluir que las mujeres caen más en impago? Noooo!



Necesitamos conocer la cantidad total de hombres y de mujeres que contiene el data set.

- 18,112 registros de mujeres
- 11,888 registros de hombres

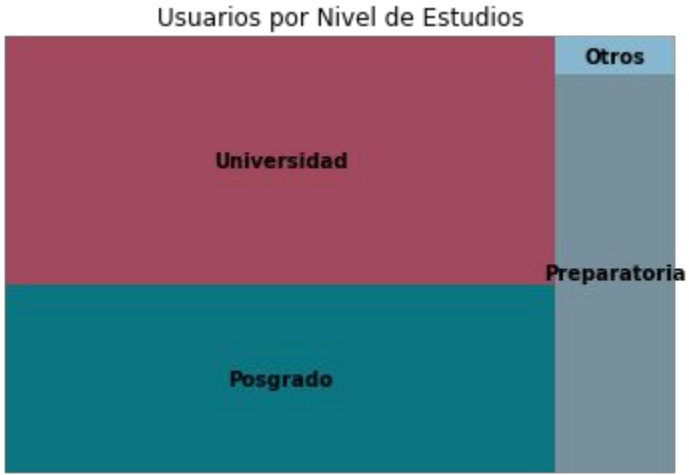


La probabilidad de que un hombre caiga en impago es del 24.16% mientras que la probabilidad de que una mujer lo haga es del 20.77%..

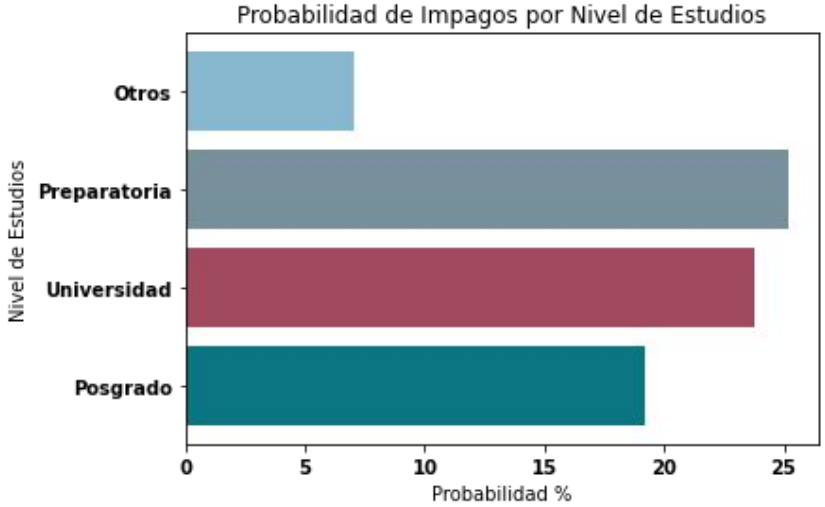
Con esto, es posible confirmar la hipótesis alterna: **Los hombres tienden a caer más en impago que las mujeres.**

3. Análisis exploratorio de datos

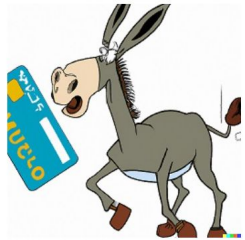
II. ¿Las personas con un menor nivel de estudios tienden a caer más en impagos?



La mayoría de los usuarios tienen estudios universitarios, seguidos por los que tienen estudios de posgrado, los que tienen preparatoria y otros.



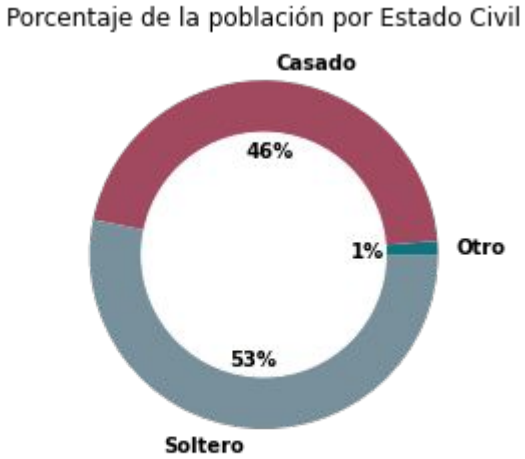
Al descartar la categoría Otros, con el gráfico anterior se puede concluir que **a menor nivel de estudios, mayor la probabilidad de caer en impago.**



3. Análisis exploratorio de datos

III. ¿Las personas casadas tienden a caer más en impago?

Es factible pensar que el estado civil de los clientes tiene impacto en su probabilidad de caer en impago. ¿Corresponde esta hipótesis con lo mostrado por el Dataset?



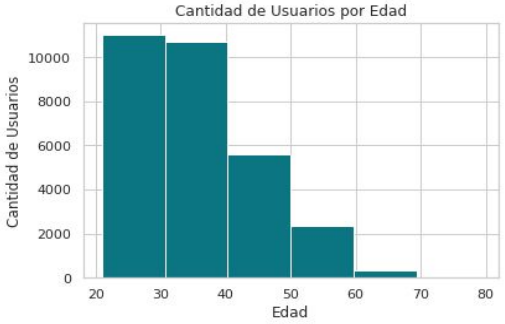
Total de la población: 30,000 personas

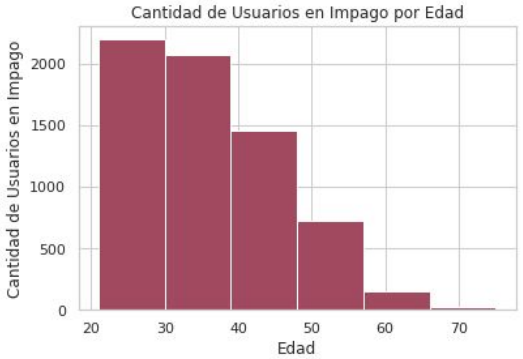
Se observa prácticamente el mismo nivel de probabilidad de caer en impago para las personas casadas o con un estado civil en la categoría *Otro*. Por lo tanto, se rechaza la Hipótesis en cuestión, pero se concluye que **las personas solteras tienen un poco menos probabilidades de caer en impago**

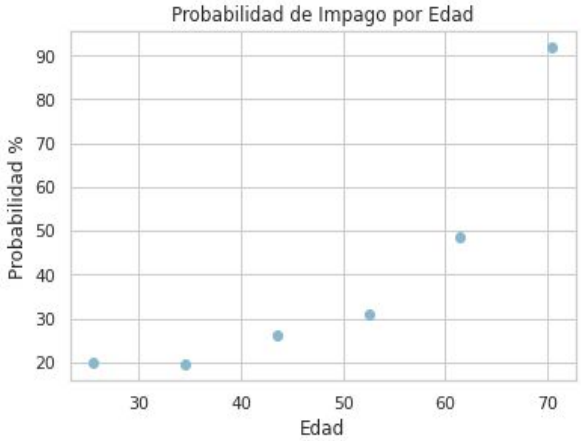
3. Análisis exploratorio de datos

IV. ¿Las personas jóvenes tienden a caer más en impago?

La edad suele considerarse un factor decisivo en la capacidad de mantener solvencia en créditos bancarios. ¿Realmente la edad es un factor a considerar en las probabilidades de caer en impago?





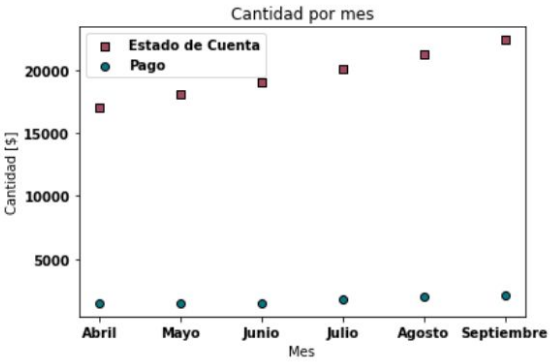


La probabilidad de caer en impago tiende a aumentar de manera exponencial conforme el usuario tiene una mayor edad. Por lo tanto se rechaza la Hipótesis

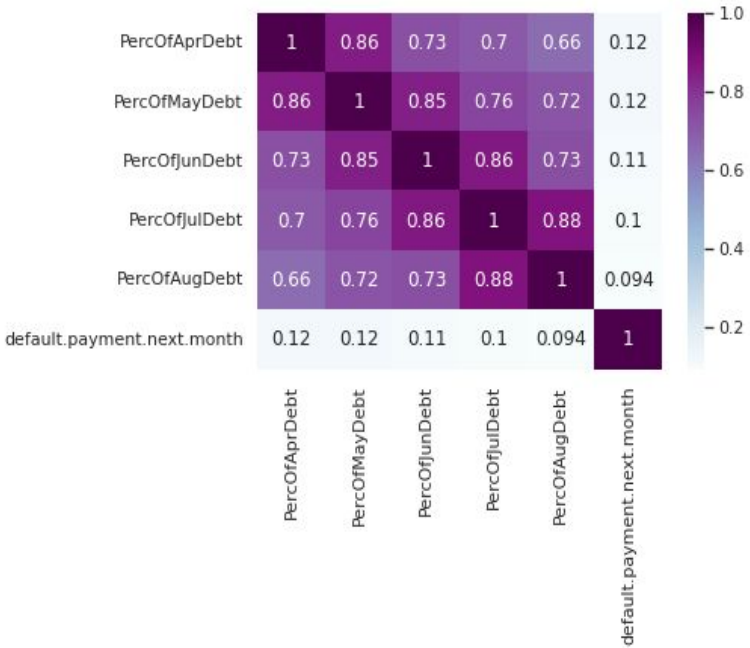
3. Análisis exploratorio de datos

V. ¿Las personas cuya deuda está más próxima a su límite de crédito tienden a caer más en impago?

Se presume que cuando una persona tiene la deuda más próxima a su límite de crédito es porque no ha podido saldar el total de la misma y ésta ha ido y seguirá creciendo hasta que finalmente la persona caiga en impago.



La mediana del saldo de los Estados de Cuenta aumentan en mayor proporción de lo que aumentan los pagos a través del tiempo

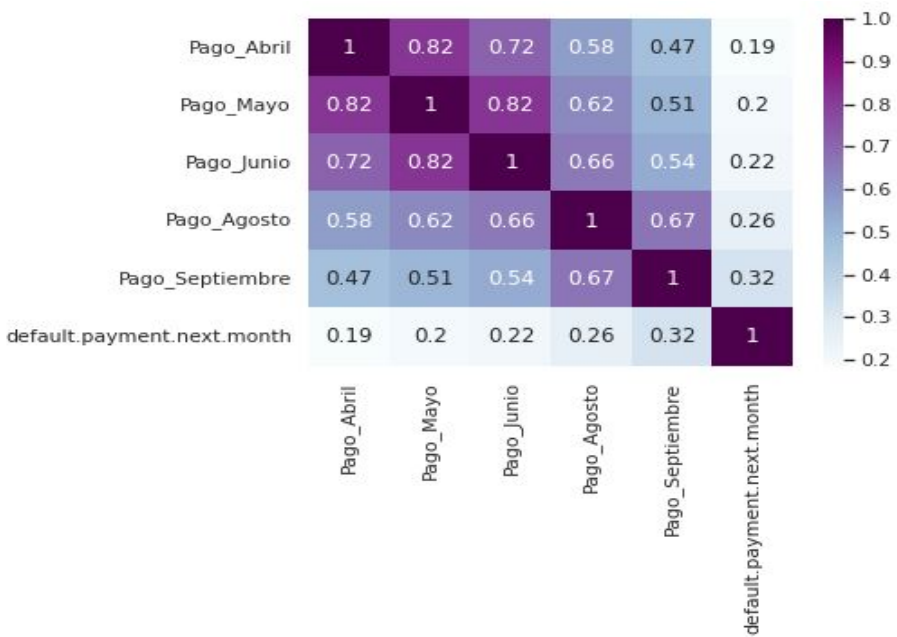


La correlación es muy baja. **No hay evidencia estadística fuerte que indique que las personas cuya deuda está más próxima a su límite de crédito tienden a caer más en impago.**

3. Análisis exploratorio de datos

VI. ¿Las personas que se han retrasado más en sus pagos tienden a caer más en impago?

Al retrasarse en el pago, la deuda sigue aumentando conforme pasa el tiempo.
¿A las personas les será cada vez más difícil saldar la deuda y llegará el momento de que caigan en impago?



Dado que se obtiene un coeficiente de correlación entre a) el estado de pago en Septiembre y b) la condición de impago de 0.39, esto indica que hay una correlación mediana entre ambas variables. Por tal motivo, no se rechaza la hipótesis en estudio. Es decir, s**e encuentra cierta relación entre los meses de retraso en Septiembre y el estado de impago**

3. Conclusiones del Análisis exploratorio de datos

I

En el proceso de aprobación de tarjetas, se recomienda al banco poner atención especial en las solicitudes de personas del sexo masculino, con un nivel de estudio bajo, no solteras y de edad avanzada, ya que este grupo tiene más probabilidades de caer en impago.

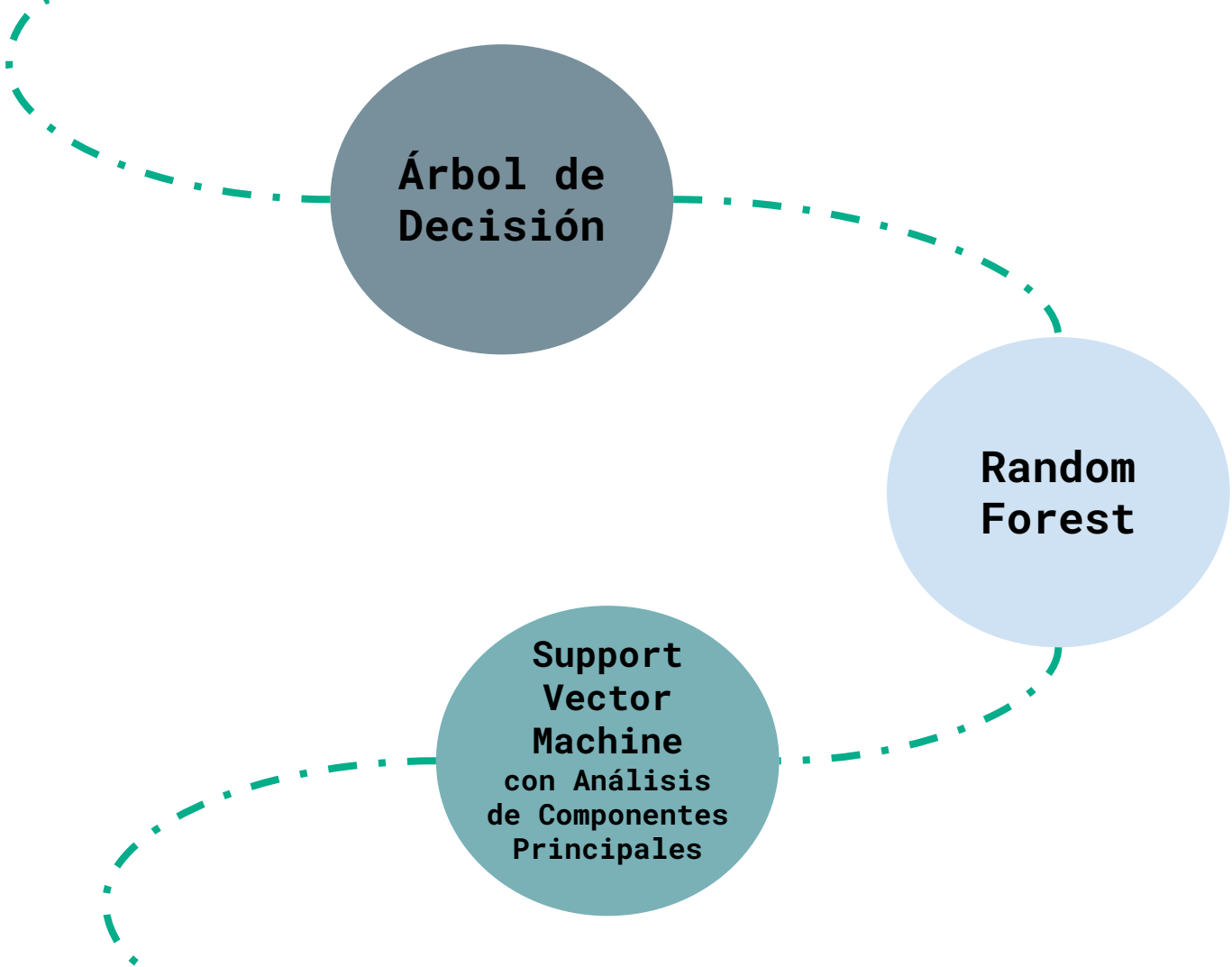
II

En el monitoreo de las cuentas, se recomienda al banco poner atención detallada en aquellos usuarios que empiezan a retrasarse en su pago, ya que es posible que lleguen a caer en impago.

III

En el monitoreo de las cuentas, se sugiere al banco poner atención especial en los usuarios del sexo masculino, con un nivel de estudio bajo, no solteras y de edad avanzada, que empiezan a atrasarse con su pago, ya que el riesgo de que caigan en impago es muy alto

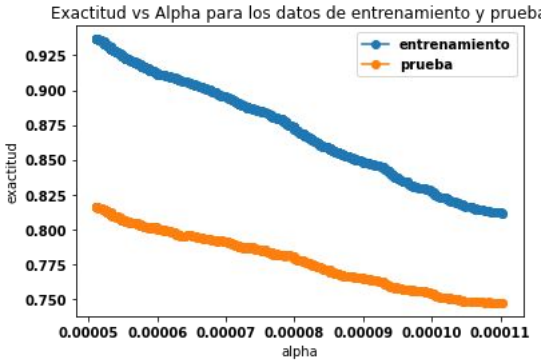
4. Modelos de Machine Learning



4. Árbol de decisión

Inicialmente se realizó el árbol de decisión sin incluir la variable Edad, y posteriormente utilizando esta variable dado que anteriormente se observó que la probabilidad de caer en impago aumenta exponencialmente con la edad.

Búsqueda de un valor óptimo para Alpha y "podar" el árbol.



Predicción de la variable impago sin incluir la variable edad con árbol de decisión.

	Exhaustividad para el caso de impago	Exactitud
test	0.95	0.86
Train	1.00	0.98

⚠️ Posible Sobreajuste

Predicción de la variable impago sin incluir la variable edad con árbol de decisión. Alpha= 0.00008

	Exhaustividad para el caso de impago	Exactitud
test	0.88	0.81
Train	0.95	0.91

✔ Sin Sobreajuste.
✔ Mejor desempeño

4. Árbol de decisión

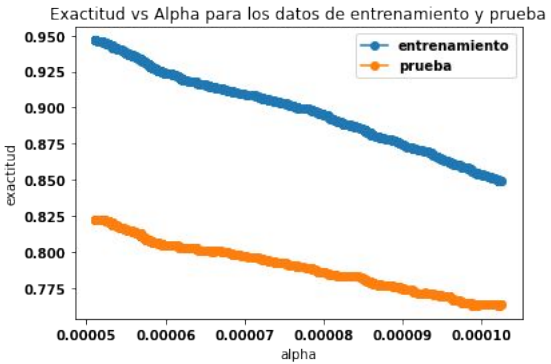
Al incluir la variable edad la probabilidad de caer en impago aumenta exponencialmente con la edad, por lo que se incluye esta variable en el modelo:

Predicción de la variable impago sin incluir la variable edad con árbol de decisión incluyendo Edad.

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.95	0.85
Train	1.00	0.99

⚠️ Posible Sobreajuste

Búsqueda de un valor óptimo para Alpha y "podar" el árbol.



Alpha de 0.000085 otorga una exactitud del 80%

Predicción de la variable impago sin incluir la variable edad con árbol de decisión incluyendo Edad. Alpha= 0.000085

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.84	0.78
Train	0.92	0.89

✔ Sin Sobreajuste.
✔ Mejor desempeño

4. Árbol de decisión- comparación

- La exhaustividad obtenida del árbol podado sin la variable edad (88%) en comparación al incluir la variable edad (84%), muestra que no tiene una influencia fuerte en el resultado (clasificación como pago o impago).

- El árbol de decisión va seleccionando las variables que tienen más influencia para la clasificación de los resultados. En el árbol creado tomando en cuenta la variable edad, se aprecia que el comportamiento de pagos (PAY, PercofDeb) tiene un peso mucho mayor que las variables demográficas (AGE, SEX, etc), ya que PAY y PercofDeb aparecen en los primeros nodos.

4. Random Forest

Primera Aproximación:

Predicción de la variable impago mediante Random Forest

	Exhaustividad para impago	Exactitud
test	0.95	0.87
Train	1.0	0.98

⚠️ Posible Sobreajuste

Predicción de la variable impago mediante Random Forest optimizado con RandomizedSearchCV

	Exhaustividad para impago	Exactitud
test	0.89	0.86
Train	0.97	0.97

✔ Requerimiento computaciónal menor

Predicción de la variable impago mediante Random Forest optimizado con GridSearchCV

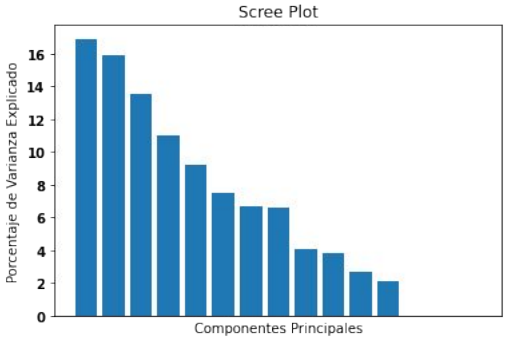
	Exhaustividad para impago	Exactitud
test	0.88	0.85
Train	0.97	0.96

⚠️ Resultado similar al obtenido con RandomizedSearchCV

4. Random Forest - comparación

- La exhaustividad obtenida en la primera aproximación fue la mejor.
- Posterior al ajuste de los hiperparámetros con ayuda de la función RandomizedSearchCV se puede observar que esta exhaustividad prácticamente no cambió.
- Una búsqueda más refinada con ayuda de la función GridSearchCV arrojó una exhaustividad similar a la obtenida con RandomizedSearchCV.
- Una búsqueda aleatoria (RandomizedSearchCV), dependiendo del número de muestras que tome, puede ayudar a encontrar hiperparámetros mejorados con un costo computacional menor comparado con el requerido por GridSearchCV.

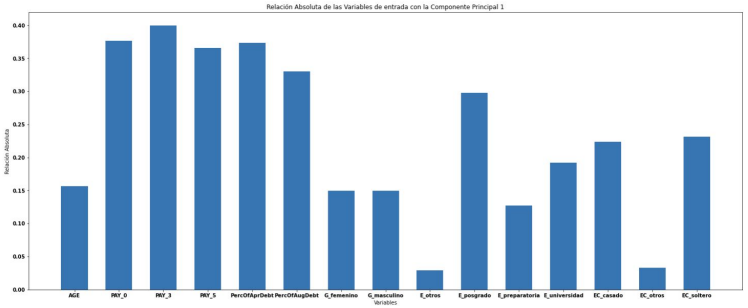
4. Support Vector Machine con Análisis de Componentes Principales



Se tomaron componentes principales que tuvieran un efecto en la variabilidad mayor o igual al 10%. Por lo tanto únicamente se tomaron PC1, PC2, PC3 y PC4.

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.54	0.42
Train	0.62	0.71

⚠️ Posible Sobreajuste



Relación Absoluta de las Variables de Entrada con la Componente Principal 1

Las variables con mayor influencia son las relacionadas al comportamiento de pago (PAY, PercOfDebt) y no las variables demográficas como nivel de estudio, estado civil, edad, etc. consideradas anteriormente.

5. Conclusiones

I. Las variables históricas y financieras tienen una mayor influencia sobre la salida (condición de pago o impago) que las demográficas.

II. Al balancear (oversampling) la variable de salida se observó una mejora tanto en el recall como en la precisión de los casos positivos.

III. Se logró obtener un buen modelo predictivo a través de un árbol de decisión podado, con una exhaustividad (recall) del 84%.
Al usar un random forest en lugar de un sólo árbol,la exhaustividad aumentó del 84% al 88%.

IV. El modelo creado usando Support Vector Machine con Análisis de Componentes Principales no resultó ser muy útil para este problema y fue el modelo que más recursos computacionales requirió, pero que, a la vez, el que demostró el menor desempeño.

V. El **modelo seleccionado** sería **Random Forest:** demostró un mejor desempeño en la exhaustividad para el caso de impago. De acuerdo a los datos obtenidos, este modelo sería capaz de predecir el 88% de los casos que caerán en impago.

TWBANK
TAIWAN-BANK

Gracias

CODERHOUSE