Universidad del Valle
Escuela de Estadística
Programa Académico de Pregrado en Estadística
Técnicas de Minería de Datos y Aprendizaje Automático
Cali, julio de 2023

Laboratorio 3

Aprendizaje Supervisado - Clasificación

Yeimy Tatiana Marín[a], Miguel Enriquez[b]

1. Lectura de Datos, transformación y ajuste de estructura

Al hacer una inspección preliminar a la hoja de datos **Data Clientes Cooperativa** para verificar que no existan datos faltantes (NA), se identifica que el registro 2 presenta un NA en la variable GENERO, por tanto, para este estudio se omite el registro. Además, se modifica la estructura de las variables cualitativas **GENERO, ESTADO_CIVIL, MODALIDAD_PAGO, HIPOTECA, RIESGO** como factores.

2. Análisis exploratorio de los datos

2.1. Estadísticas descriptivas

De la siguiente tabla de resumén descriptivo obtenido para cada variable, se observa que en las cuantitativas el valor mínimo es menor al primer cuartil, el valor máximo es mayor al tercer cuartil y la media es levemente mayor que la mediana, infiriendo de lo anterior, posible presencia de datos atípicos (valores extremos) y una distribución en las observaciones sesgada a la derecha. De las variables cualitativas se observa una participación casi equitativa del género femenino (f) y masculino (m), pero poca participación de clientes con mayor riesgo de incurrir en impago.

EDAD	INGRESOS	GENERO	ESTADO_CIVIL	NUM_HIJOS
Min:18	**Min:**15005	**f:**2077	**divsepwid:**873	**Min:**0
1st Qu:23	**1st Qu:**20497	**m:**2039	**married:**2088	**1st Qu:**1
Mediana:31	**Mediana:**23487		**single:**1155	**Mediana:**1
Media:31.82	**Media:**25572			**Media:**1.453
3rd Qu:41	**3rd Qu:**27565			**3rd Qu:**2
Max:50	**Max:**59944			**Max:**4
NUM_TARJETAS	**MODALIDAD_PAGO**	**HIPOTECA**	**PRESTAMOS**	**RIESGO**
Min:0	**monthly:**2025	**n:**917	**Min:**0	**F:**3312
1st Qu:1	**weekly:**2091	**y:**3199	**1st Qu:**1	**V:**804
Mediana:2			**Mediana:**1	
Media:2.43			**Media:**1.376	
3rd Qu:4			**3rd Qu:**2	
Max:6			**Max:**3	

Tabla 1: Resumén de las variables

[a]Código: 1524344. E-mail: yeimy.marin@correounivalle.edu.co
[b]Código: 2023796. E-mail: miguel.enriquez@correounivalle.edu.co

De los gráficos descriptivos para variables cuantitativas, categóricas y de dispersión bivariado, se observa que entre los 25 a los 45 años se encuentran la mayoría de los clientes tanto los de cumplimiento (F) que en su mayoría están casados, tienen hipoteca y registran su pago los fines de semana, como los clientes que incumplen con el pago. También, se puede observar que existen datos atípicos de los clientes cumplidos en la variable INGRESO (Ver figuras 1,2 y 3).

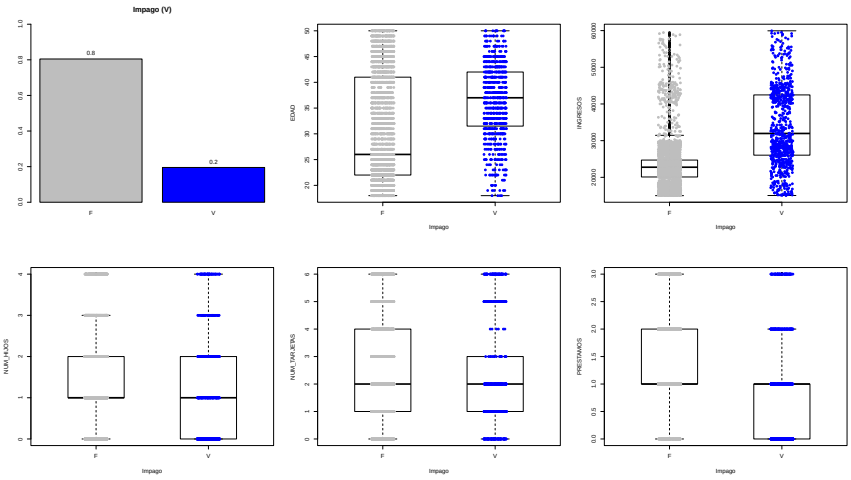

FIGURA 1: Boxplot. visualización gráfica de todas las variables Cuantitativas vs Impago.

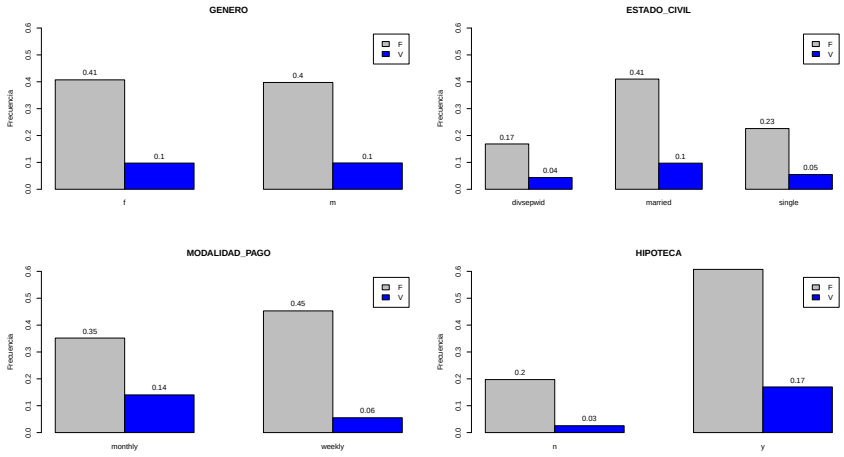

FIGURA 2: Barplot. visualización gráfica de todas las variables Cualitativas vs Impago.

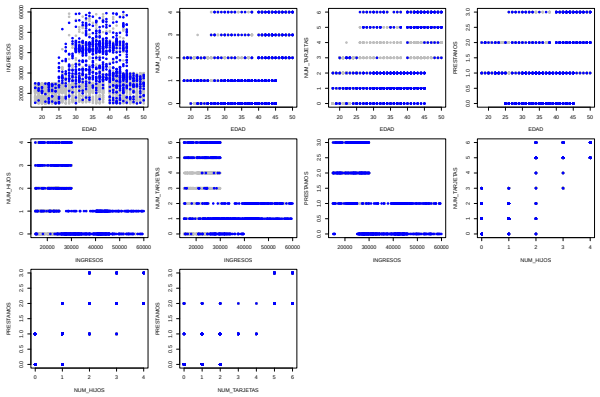

FIGURA 3: Visualización bivariada. gráficos de dispersión X's vs Impago

2.2. Visualización Multivariada - Análisis de Componentes Principales

En las figuras 4 y 5 de representación simultánea, se analiza que la mayoría de los clientes de la Cooperativa son cumplidos con los pagos, aunque no tienen ingresos tan altos como los clientes de impago. También se observa que, entre los clientes cumplidos (F) hay una división interna, entre un grupo pequeño con varios hijos, altos préstamos y números de tarjeta, y el grupo con mayor representación de clientes (F) que es opuesto a este grupo pequeño. Por último, se corrobora que los clientes que incurren en impago no tienen altos préstamos, ni número de tarjetas con la cooperativa.

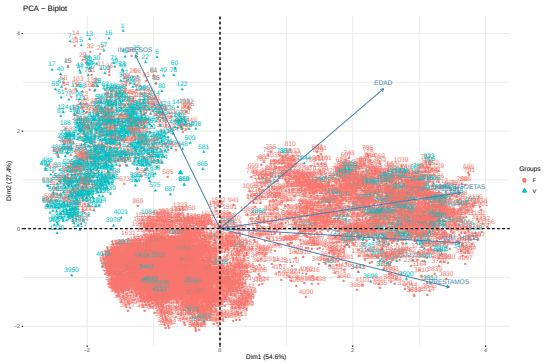

FIGURA 4: Representación simultánea en el plano 1-2

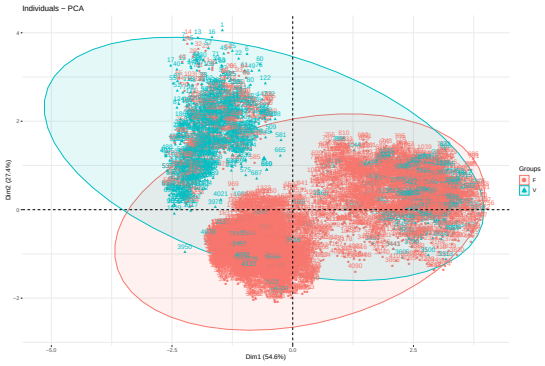

FIGURA 5: Agrupación de clientes por Impago (V) o Cumplimiento (F)

3. Entrenamiento y comparación de modelos de clasificación.

Con el fin de emplear los modelos de clasificación, se realiza la selección del porcentaje para dividir los datos en dos conjuntos más pequeños: 80 % datos de entrenamiento (trainig) y 20 % datos de prueba (test). El subconjunto de datos de entrenamiento será utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se empleará para comprobar el comportamiento del modelo estimado.

3.1 Modelo Estadístico. Regresión Logística (RL).

Se comienza con la selección de las variables que explican la variable predictora **RIESGO (V: Impago)**, por tanto, se inicia con el ajuste de un modelo saturado, es decir, modelo con todas las variables, en el cuál se observa que la variable GENERO no es significativa al modelo. Utilizando entonces, para una

correcta selección de variables en el modelo el método de eliminación hacia atrás (Backward), coincidiendo en eliminar la variable GENERO para un modelo reducido. Luego, con la ayuda de la Curva ROC se selecciona el punto de corte o hiperparamétro para este caso, siendo decisivo como regla de clasificación de los datos como (V: Impago, F: Cumplimiento) en el modelo reducido.

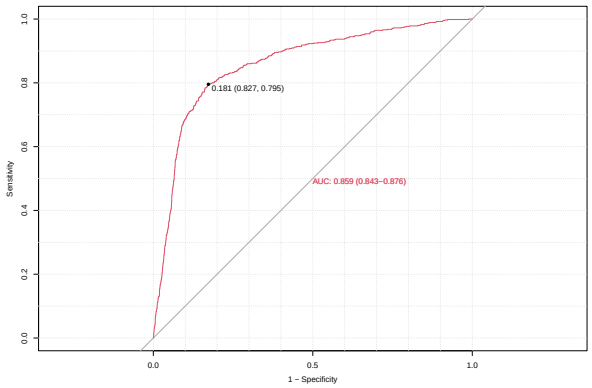

Figura 6: Curva ROC. pc=0.181

3.2 Modelo de aprendizaje automático. Maquina de soporte vectorial (SVM).

Es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión. En este caso lo utilizaremos para clasificar las personas que tengan impago o cumplimiento.

Primero planteamos el modelo espicificando que es de tipo çalsificaciónz que su kernel es "lienal". Una vez obtenido el modelo pasamos a tunear sus hiperparametros para evitar un sobre ajuste o un infra ajuste. Una vez hecho lo anterior obtenemos el modelo ajustado, el cual es el mejor posible para este caso.

3.3 Modelo de ensamble. Random Forest (RF).

Un modelo Random Forest está formado por un conjunto (ensemble) de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping). Esto implica que cada árbol se entrena con unos datos ligeramente distintos.

Para realizar el anterior modelo, comenzamos especificando la cantidad de arboles que queremos que contenga nuestro modelo, que para efectos de este caso es 500. Una vez obtenido los resultados de ese primer modelo, lo tuneamos por medio de su número de variables que se seleccionarán en cada partición de cada árbol del bosque de forma aleatoria por validación cruzada.

Una vez hechos los pasos anteriores hemos encontrado nuestro mejor modelo posible.

4. Recomendación del modelo que mejor clasifique al cliente con grandes probabilidades de impago.

En el punto 3 se ha tomado la decisión metodológica de escoger tres modelos de distinta clase cada uno para hallar el modelo que más se ajuste al problema. Los modelos de clasificación escogidos corresponden a los siguientes:

1. Estadístico (Regresión logística).
2. Algoritmo de aprendizaje (Máquina de soporte vectorial).

3. Aprendizaje automático (Máquina de soporte vectorial).

En esta ocasión el caso de estudio tiene como objetivo crear un modelo que clasifique a priori a los clientes que quieran realizar un crédito y tengan una mayor probabilidad de caer en impagos. Para hallar dicho modelo nos enfocaremos en la sensibilidad como medida principal, ya que está, la cual nos informa la capacidad del modelo de clasificar a los verdaderos positivos, que en este caso serían las personas que verdaderamente tienen una gran posibilidad de caer en impagos.

Ahora bien, más allá de la medida de sensibilidad también se debe de tener en cuenta que cada modelo tiene sus propias características, y esto genera un trade-off entre lo que queramos encontrar y lo que nos pueda brindar cada modelo.

Modelo	Enfoque	Fundamento teórico	Ventajas
Estadistico	Los modelos estadísticos se basan en teoría estadística y métodos inferenciales para explicar las relaciones entre las variables y hacer inferencias sobre los parámetros del modelo.	Los modelos estadísticos se basan en supuestos y principios estadísticos bien establecidos	Interpretación: Los modelos estadísticos a menudo permiten una interpretación más fácil de los coeficientes o parámetros del modelo, lo que facilita la comprensión de las relaciones entre las variables.
Aprendizaje automatizado	Los modelos de Aprendizaje automatizadose centran en desarrollar algoritmos y técnicas que permitan a las máquinas aprender patrones y tomar decisiones o hacer predicciones automáticamente, sin una programación explícita.	Los modelos de machine learning se basan en métodos de aprendizaje automático y procesamiento de datos para encontrar patrones ocultos en los datos y construir modelos predictivos o descriptivos.	Escalabilidad: Los modelos de machine learning son especialmente adecuados para grandes conjuntos de datos y problemas complejos, ya que pueden manejar relaciones no lineales y adaptarse a datos de alta dimensionalidad.
Ensamble	Los modelos de ensamble combinan múltiples modelos individuales más simples para mejorar la precisión y el rendimiento general del modelo.	Los modelos de ensamble pueden combinar modelos estadísticos o de machine learning mediante técnicas como el promedio de predicciones, la votación o la combinación ponderada de modelos.	Reducción del sesgo y la varianza: Los modelos de ensamble pueden reducir el sesgo y la varianza inherentes a los modelos individuales, lo que puede mejorar la generalización y la precisión del modelo final.

Tabla 2: Diferencia de los modelos

Ahora que se tiene claro la diferencias entre modelos, pasaremos a analizar la sensibilidad de los modelos arrojada con los datos de entrenamiento y de testeo.

Modelo	Sensibilidad
Regresión Logistica	0.7953846
SVM	0.6723077
Random Forest	0.6938462

Tabla 3: Diferencia de sensibilidad, datos de entramiento

Modelo	Sensibilidad
Regresión Logistica	0.7272727
SVM	0.5974026
Random Forest	0.6038961

Tabla 4: Diferencia de sensibilidad, datos de testeo

Como podemos ver tanto en la TABLA 3 como en la TABLA 4, el modelo que mejor clasifica a los clientes con mayor probabilidad de caer en impagos es el modelo estadístico de regresión logística. Por lo cual, se le aconsejaría al banco acogerse a este modelo ya que también tiene una excelente capacidad explicativa, siendo esto una ventaja porque puede facilitar el entendimiento en el proceso de un cliente que caiga en impago de forma mucho más práctica e intuitiva. Esto le sirve a la cooperativa para entender mucho mejor las cualidades de su negocio y proponer soluciones efectivas para evitar ese tipo de clientes y maximizar sus ganancias con clientes con un buen historial de pago.