Aprendizaje maquina para determinar crédito otorgados a personas físicas y su rentabilidad

Lic. Aranza Alejandra Esteban Avalos

20 de noviembre de 2023

1. Abstract

La capacidad de predecir con precisión los montos de préstamos representa una ventaja tanto para instituciones financieras como para clientes. utilizando herramientas como aprendizaje máquina, se desarrollan análisis con grandes conjuntos de datos para revelar patrones y tendencias en el comportamiento financiero. Este estudio se sumerge aplicando técnicas avanzadas de aprendizaje máquina, como Árboles de Decisión y Regresión Lineal, para predecir los montos de los préstamos en base a ciertas características del conjunto de datos.

Nuestro enfoque se centra en obtener mejor visibilidad a los préstamos personales para en futuro identificar créditos empresariales. Se utiliza un conjunto de datos que contiene información desde los ingresos anuales de los solicitantes hasta sus historiales de crédito. Mediante la implementación de de dichos modelos, buscamos mejorar la eficiencia en la toma de decisiones de préstamos.

A través de este estudio, se busca dar mayor visibilidad al análisis de datos, cuando se combina con herramientas de aprendizaje maquina, puede convertirse en un punto crítico para la toma decisiones más informadas y estratégicas en el ámbito financiero.

2. Introducción

El análisis se centra en los créditos otorgados a personas físicas, con un enfoque específico en la predicción del monto de préstamos y adicional a esto identificar patrones o las diferentes relaciones entre las variables. Se cuenta con una data detallada que incluye variables como el monto del préstamo, la tasa de interés, el ingreso anual del acreditado, el historial de pagos, entre otras, se buscan los patrones y factores clave que influyen en la decisión de otorgar un préstamo y adicional su potencial rentabilidad.

El objetivo es primero, identificar las características de los préstamos y de los acreditados que están más fuertemente asociadas con la aprobación de montos de préstamo adicionales; y en un segundo plano, evaluar la rentabilidad de estos préstamos adicionales a través de un análisis que integra diversas variables financieras y personales. Se emplearán técnicas avanzadas de análisis de datos y modelado predicativo, incluyendo regresión y clasificación, para generar una visión detallada que pueda ser útil en orientar las estrategias de préstamos

3. Descripción de los datos

El conjunto de datos utilizado en este estudio representa un compendio detallado de préstamos personales, abarcando múltiples aspectos financieros y personales de los solicitantes. Cada entrada en este conjunto de datos incluye variables como el monto del préstamo, la tasa de interés, el plazo del préstamo, los ingresos anuales del solicitante, el historial de crédito, entre otros. Estas variables ofrecen una visión integral de las circunstancias financieras y el comportamiento crediticio de los individuos, lo que las hace idóneas para el análisis mediante aprendizaje máquina.

- ***loan_amnt***: Monto del préstamo solicitado. Rango típico de $1,000 a $35,000.
- ***term***: Duración del préstamo en meses. Típicamente 36 o 60 meses.
- ***int_rate***: Tasa de interés del préstamo. Varía, con un rango común de 5.42 % a 24.11 %.
- ***installment***: Cuotas mensuales del préstamo. Rango de $21.74 a $1,288.10.
- ***grade***: Calificación crediticia asignada al préstamo. Generalmente entre 0 y 2.
- ***sub_grade***: Subcategoría de la calificación crediticia. Varía, con un rango de 0 a 34 diferentes comenzando desde la A hasta G con 5 subcategorias cada una
- ***emp_length***: Antigüedad laboral del solicitante en años. Rango de 0 a 10 años.
- ***annual_inc***: Ingresos anuales del solicitante. Varía ampliamente, con un rango de $6,000 a $1,782,000.
- ***purpose***: Propósito del préstamo. Variedad de propósitos representados.
- ***addr_state***: Estado de residencia del solicitante. Variedad de estados representados.
- ***earliest_cr_line***: Fecha del primer crédito reportado. Fechas variadas, algunas negativas.
- ***open_acc***: Número de cuentas de crédito abiertas. Rango de 2 a 34 cuentas.
- ***total_pymnt_inv***: Pagos totales recibidos por el inversor. Rango de $96.96 a $56,475.05.
- ***total_rec_int***: Intereses totales recibidos. Rango de $11.64 a $22,930.84.
- ***last_pymnt_d***: Fecha del último pago realizado. Fechas variadas.
- ***last_pymnt_amnt***: Monto del último pago realizado. Rango de $0.01 a $36,115.20.
- ***last_credit_pull_d***: Fecha de la última consulta de crédito.
- ***default_ind***: Indicador de si el préstamo ha incumplido. Variedad de valores.
- ***sub_grade_encoded***: Codificación numérica de la subcalificación. Rango de 0 a 34.
- ***credit_history***: Duración del historial de crédito. Variedad de valores.

Cuadro 1: Estadísticas descriptivas del conjunto de datos

Columna	**Media**	**Desv. Estándar**	**Mínimo**	**25 %**	**50 %**	**75 %**	**Máximo**
loan_amnt	12,914.28	8,497.64	1,000.00	6,000.00	11,400.00	17,600.00	35,000.00
term	44.06	11.33	36.00	36.00	36.00	60.00	60.00
int_rate	12.45	4.23	5.42	8.90	12.42	15.27	24.11
installment	365.79	227.09	21.74	195.27	323.48	483.38	1,288.10
grade	1.41	0.70	0.00	1.00	2.00	2.00	2.00
sub_grade	9.57	7.00	0.00	4.00	8.00	14.00	34.00
emp_length	5.40	3.54	0.00	2.00	5.00	10.00	10.00
annual_inc	71,080.48	48,185.73	6,000.00	43,200.00	60,000.00	85,000.00	1,782,000.00
purpose	8.81	5.25	0.00	2.00	13.00	13.00	13.00
addr_state	14.63	14.31	3.00	4.00	4.00	27.00	45.00
earliest_cr_line	9,922.50	2,428.80	-2,555.00	8,770.00	10,238.00	11,688.00	13,889.00
open_acc	9.51	4.25	2.00	6.00	9.00	12.00	34.00
total_pymnt_inv	13,831.30	10,102.76	96.96	6,133.00	11,360.34	18,616.80	56,475.05
total_rec_int	2,843.84	3,249.52	11.64	745.01	1,607.72	3,672.00	22,930.84
last_pymnt_d	15,942.66	412.34	14,983.00	15,709.00	16,076.00	16,082.00	16,801.00
last_pymnt_amnt	3,380.40	5,379.24	0.01	281.46	665.57	4,382.62	36,115.20
last_credit_pull_d	16,331.69	472.13	14,981.00	16,076.00	16,441.00	16,801.00	16,801.00
default_ind	6,353.53	2,480.38	1,338.00	4,657.00	5,934.00	7,639.00	18,750.00
sub_grade_encoded	9.57	7.00	0.00	4.00	8.00	14.00	34.00
credit_history	6,353.53	2,480.38	1,338.00	4,657.00	5,934.00	7,639.00	18,750.00

4. Metodología

4.1. Limpieza

La limpieza de datos es un paso crucial en cualquier análisis estadístico y de aprendizaje máquina. En nuestro estudio, este proceso implicó varias fases:

Se identificaron y manejaron los valores faltantes en el conjunto de datos para asegurar la integridad del análisis.

1. Selección de columnas: Selección de las columnas relevantes que se conideraron con mayor relevancia para incluir en el análisis.

2. Validación de Montos de Préstamo: Verificación que el monto del préstamo coincidiera con el monto financiado y se omitio *funded_amnt*.

3. Identificación de Variables con NaN: Variables que contenían valores NaN (valores faltantes) en el conjunto de datos, debido a que estos valores pueden ser problemáticos para análisis posteriores.

4. Procesamiento de la Columna *term*: Eliminación de las palabras y caracteres no numéricos en la columna *term*, para obtener valores numéricos para el plazo del préstamo.

5. Asignación de Valores Numéricos a Variables Categóricas: Asignación de valores numéricos a variables categóricas como *sub_grade*, *grade*, *addr_state*, *purpose*, *application_type*, y *home_ownership*.

6. Se trabajo en hacer normalización para homogeneizar las escalas de las diferentes variables y se realizaron transformaciones necesarias para mejorar la calidad y la relevancia del análisis.

Finalmente se revisaron y corrigieron posibles errores e inconsistencias, como valores atípicos o incoherentes.

4.2. Métodos y metricas de medición

- *F_regression*: Se utiliza para evaluar la importancia estadística de las características en un modelo de regresión, identificando las más relevantes.
- *VarianceThreshold*: Sirve para eliminar características con baja varianza en un conjunto de datos, lo que puede ayudar a simplificar modelos al eliminar variables con poca variabilidad.
- *Random Forest*: Algoritmo de aprendizaje automático que se utiliza para tareas de clasificación y regresión al combinar múltiples árboles de decisión, lo que mejora la precisión y reduce el sobreajuste.
- *Linear Regression*: Modelo de aprendizaje automático que busca establecer una relación lineal entre variables de entrada y una variable de salida, utilizándose principalmente para predicciones numéricas.
- *PCA Linear Regression*: Combina regresión lineal con Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de las características y mejorar la eficiencia del modelo de regresión, útil cuando se tienen muchas variables predictoras. (Figura 2)
- *K-Means*: Un algoritmo de agrupamiento no supervisado que divide un conjunto de datos en grupos *clusteres* basados en similitudes entre los puntos de datos, asignando cada punto al *cluster* más cercano al centroide. (Figura 3 y 4)
- *GridSearchCV* (Hiperparámetros): Técnica de búsqueda exhaustiva que se utiliza para encontrar los mejores hiperparámetros para un modelo de aprendizaje automático al probar diferentes combinaciones en un rango predefinido para optimizar su rendimiento.

Las características muestran variaciones en su importancia dependiendo del método de análisis utilizado. Esto subraya la importancia de utilizar múltiples enfoques para la selección de características y la necesidad de interpretar los resultados en el contexto de cada modelo y su aplicación específica.

Random Forest parece ser el método más robusto y revelador, ya que captura tanto las relaciones lineales como las no lineales y la importancia de las características más sutiles que pueden ser ignoradas por los otros métodos. Adjunto resultados (Cuadro 2, Figura1)

4.3. Pruebas y Selección de Modelos

Se realizo pruebas de hipótesis utilizando técnicas como RidgeCV, VarianceThreshold y fregression para determinar la relevancia de las diferentes variables. Estos métodos ayudaron a identificar las variables más significativas y a refinar nuestro conjunto de datos.

Uso de K-Means para Análisis Exploratorio Como parte de nuestro análisis exploratorio, se aplico el algoritmo K-Means para identificar agrupaciones naturales dentro del conjunto de datos. Este enfoque nos permitió descubrir patrones subyacentes y relaciones entre las variables, informando nuestra selección y aplicación de modelos predictivos.

Posición	Columna	Score	Varianza	Coeficiente	Importancia
3	*term*	3132.836389	1.284657e+02	2225.749823	0.058494
1	*total_pymnt_inv*	37938.163310	1.020657e+08	22.649347	0.027245
9	*loanVSfunded*	631.248277	1.461918e+06	1254.934888	0.020823
8	*int_rate*	1201.441424	1.792676e+01	559.849918	0.004624
7	*sub_grade*	1283.991015	4.900767e+01	-838.278096	0.001386
6	*sub_grade_encoded*	1283.991015	4.900767e+01	-838.278096	0.001299
2	*total_rec_int*	11573.916538	1.055939e+07	998.415225	0.000430
4	*last_pymnt_amnt*	2600.952934	2.893619e+07	195.461074	0.000152
5	*annual_inc*	1655.671414	2.321865e+09	-8.755183	0.000144
13	*open_acc*	262.602180	1.809211e+01	-28.702740	0.000081
10	*default_ind*	542.357197	6.152284e+06	105.483286	0.000071
12	*earliest_cr_line*	538.572498	5.899048e+06	190.835534	0.000070
11	*credit_history*	542.357197	6.152284e+06	105.483286	0.000067
14	*emp_length*	225.453721	1.251704e+01	13.802665	0.000048

Cuadro 2: Resumen de resultados

Installment identificado como un predictor importante, indicando que la estructura de los pagos del préstamo es muy importante para la predicción. *Term* y *loanVSfunded* también destacados como relevantes. *Annual inc* mostró una relación no lineal o compleja, posiblemente debido a la influencia de la regularización.

Con el análisis se puede interpretar que ciertas características financieras son más influyentes en la rentabilidad de los préstamos, con una atención particular en la estructura de pagos y términos del préstamo.

4.4. Modelado y Evaluación

Seleccionamos Árboles de Decisión y Regresión Lineal como nuestros modelos principales:

Árbol de Decisión

Un árbol de decisión se representa como una serie de divisiones basadas en las características. Cada nodo del árbol representa una característica, cada rama representa una decisión y cada hoja representa un resultado. La fórmula general de un árbol de decisión no es tan directa pero la idea básica se puede expresar como una serie de decisiones condicionales

Si $(x_i \leq umbral)$ entonces ir a la rama izquierda, sino ir a la rama derecha

Como resultado obtuve que se capturó relaciones complejas, aunque con un riesgo de sobreajuste.

Regresión Lineal

La regresión es un enfoque estadístico para modelar la relación entre una variable dependiente en este caso, la rentabilidad del préstamo y una o más variables independientes

Proporcionó un enfoque más simplificado y generalizable. Ambos modelos fueron evaluados usando métricas como MAE, RMSE y R^2 para comparar su eficacia y precisión.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Se comparó la Regresión Lineal con Árboles de Decisión donde los arboles de decisión muestran evidencia de sobreajuste, con un alto rendimiento en el conjunto de entrenamiento y una disminución en los datos de prueba. El sobreajuste es una preocupación para los modelos de árbol. En

cambio las regresión lineal mostró un comportamiento más equilibrado y consistente, con una buena generalización entre los conjuntos de entrenamiento y prueba. (Cuadro 3)

Cuadro 3: Métricas para Modelos de árboles y regresión

Métricas	**Árboles**	**Regresión**
MSE	428.822365	877.750903
RMSE	1658.263721	1672.686004
R2	0.962748	0.962098

5. Exploracion de resultados

El Árbol de Decisión y la Regresión Lineal, se revelaron modelos distintos. La predicción de los montos de préstamos. Cada uno nos contó una historia diferente sobre los patrones y relaciones en nuestros datos.

El Árbol de Decisión, con un MAE y RMSE de cero y un R^2 de 1.0, inicialmente pareció el modelo perfecto, describiendo los datos con una precisión impecable. Sin embargo, este nivel de perfección sugiere un posible sobre ajuste.

Por otro lado, la Regresión Lineal muestra una visión más equilibrada. Con un MAE de 848.55, un RMSE de 1571.36 y un R^2 de 0.9656. Aunque no cubrió cada detalle con la misma precisión del Árbol de Decisión.

El modelo de Árbol de Decisión mostró un rendimiento perfecto en el conjunto de entrenamiento, lo cual puede indicar un sobreajuste, donde el modelo se adapta demasiado bien a los datos de entrenamiento, pero podría no generalizar bien a datos nuevos. Por otro lado, el modelo de Regresión Lineal, aunque menos preciso, mostro resultados más realistas para una situación práctica, con un R^2 muy alto, lo que indica que el modelo puede explicar una gran proporción de la variabilidad en el monto del préstamo.

Esta exploración de los préstamos personales a través del aprendizaje automático abre muchas nuevas ideas y posibilidades. Aprender sobre la importancia de equilibrar la precisión y la generalización, la teoría y la práctica. En última instancia, aunque las herramientas y técnicas pueden ser poderosas, su verdadero valor reside en cómo las aplicamos y en las historias que elegimos contar con ellas.

6. Conclusiones

Este estudio ilustra cómo diferentes modelos de aprendizaje máquina pueden ser utilizados para predecir montos de préstamos. Mientras que el Árbol de Decisión muestra una precisión teóricamente perfecta en el conjunto de entrenamiento, es probable que sufra de sobreajuste. En cambio, la Regresión Lineal, aunque no tan precisa, proporciona un balance más realista entre precisión y capacidad de generalización.

[6] [7] [5] [3] [1] [2] [4]

Referencias

[1] Alberto Benavides. *Aprendizaje No Supervisado*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/4_no_supervisado.ipynb.

[2] Alberto Benavides. *Aprendizaje Supervisado*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/5_supervisado.ipynb.

[3] Alberto Benavides. *Métricas de Desempeño - Aprendizaje Automático*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/7_metricas_desempeno.ipynb.

[4] Alberto Benavides. *Selección de Características*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/3_sel_carac.ipynb.

[5] scikit-learn. *Clustering*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/modules/clustering.html.

[6] scikit-learn. *Comparing Successive Halving and Random Search*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_successive_halving_heatmap.html.

[7] scikit-learn. *Supervised Learning*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/supervised_learning.html.

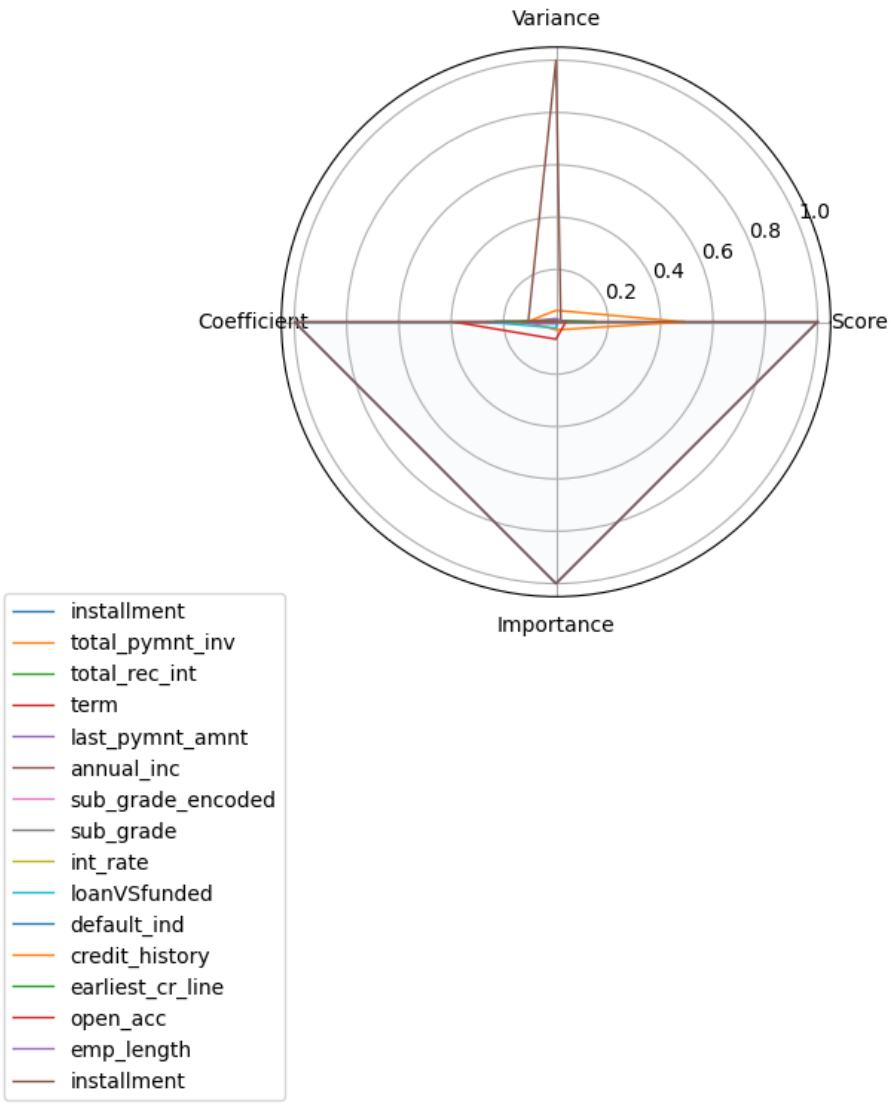

Figura 1: resumen métodos

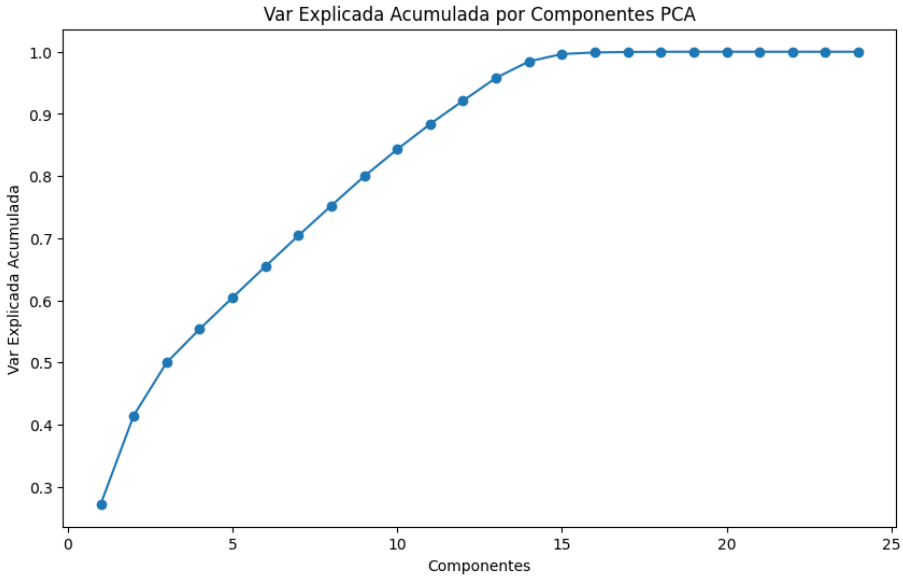

Figura 2: PCA

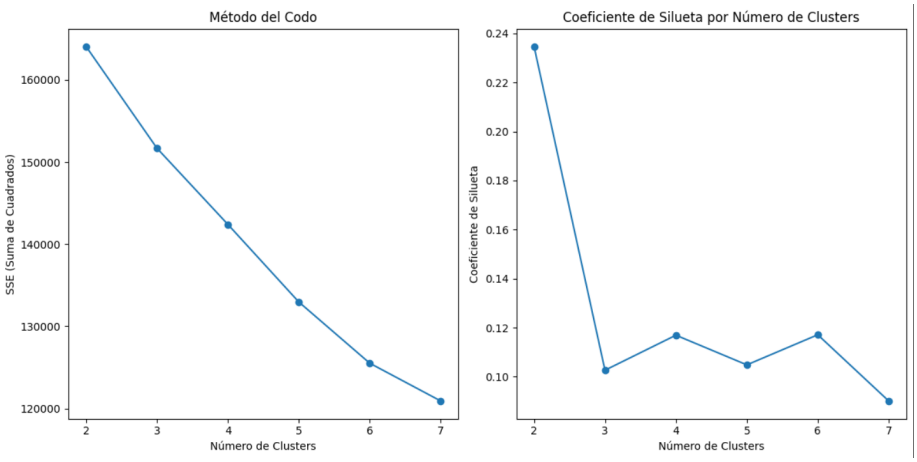

Figura 3: Método de codo

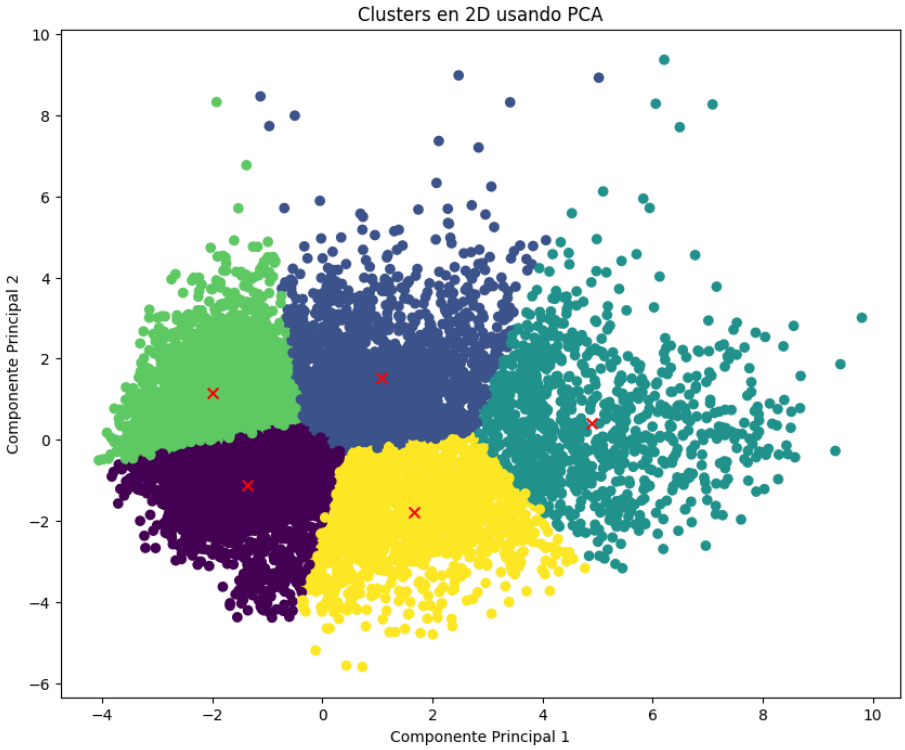

Figura 4: Clustering

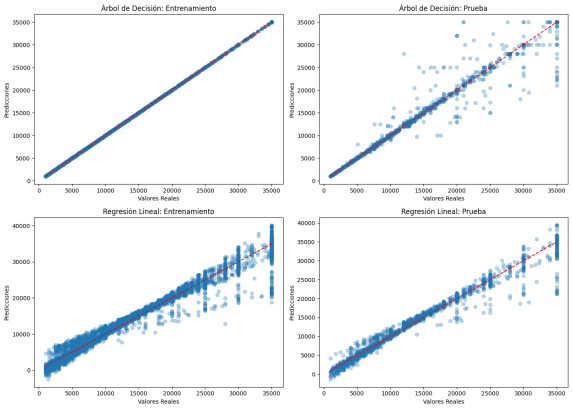

Figura 5: Arboles de decisión vs Regresion