15 de julio del 2025

Proyecto

Matemáticas: fundamento de Machine Learning

Introducción

Este proyecto busca mostrar cómo las matemáticas son la base de los modelos de Machine Learning, específicamente de los árboles de decisión. Más allá de usar librerías de programación que resuelven todo de forma automática, el propósito es entender paso a paso qué ocurre “tras bambalinas”. De esta manera, se ofrece una visión clara y accesible de cómo los datos se transforman en reglas lógicas de clasificación, combinando fundamentos matemáticos con aplicaciones tecnológicas.

Objetivos

- Mostrar cómo las matemáticas son la base sobre la cual se construyen los algoritmos de Machine Learning, más allá de las librerías y plataformas tecnológicas.
- Visualizar cómo se generan las decisiones y los resultados, sin necesidad de depender únicamente de la programación
- Ir más allá de la aplicación práctica de modelos y reflexionar sobre lo que realmente ocurre detrás de la “caja negra” de los algoritmos.
- Ofrecer una visión accesible que sirva como punto de partida para estudiantes, curiosos y profesionales que deseen reforzar o redescubrir las bases matemáticas y estadísticas que sostienen el campo de la Ciencia de Datos.

Contenido

1. Planteamiento del problema

En la gestión de clientes bancarios, uno de los retos es anticipar si un cliente aceptará o no un préstamo personal, ya que no basta con observar sus ingresos o edad de forma aislada. Aunque los modelos de Machine Learning permiten clasificar este comportamiento, con frecuencia se priorizan los resultados y se deja de lado la comprensión de los procedimientos matemáticos y estadísticos que los sostienen. Este proyecto, utilizando un árbol de decisión, busca mostrar de manera clara y accesible cómo se construye esa clasificación.

2. Dataset

Este conjunto de datos contiene información de 5,000 clientes de un banco, con variables demográficas y financieras como la edad, años de experiencia, ingresos anuales, tamaño de la familia, nivel educativo, gasto promedio con tarjeta de crédito, valor de hipoteca y si posee distintos productos financieros (cuenta de valores, certificado de depósito, etc.). La variable objetivo (Personal Loan) indica si el cliente aceptó o no un préstamo personal ofrecido en la última campaña del banco.

Variable (Inglés)	Descripción (Español)
ID	ID del cliente
Age	Edad del cliente en años cumplidos
Experience	Años de experiencia profesional
Income	Ingreso anual del cliente (en miles de dólares)
ZIPCode	Código postal del domicilio del cliente
Family	Tamaño de la familia del cliente
CCAvg	Gasto promedio mensual en tarjetas de crédito (en miles de dólares)
Education	Nivel educativo. 1: Pregrado; 2: Graduado; 3: Avanzado/Profesional
Mortgage	Valor de la hipoteca de la vivienda, si la tiene (en miles de dólares)
Personal Loan	¿El cliente aceptó el préstamo personal ofrecido en la última campaña? (No = 0, Sí = 1)
Securities Account	¿El cliente tiene una cuenta de valores con el banco? (No = 0, Sí = 1)
CD Account	¿El cliente tiene una cuenta de certificado de depósito (CD) con el banco? (No = 0, Sí = 1)
Online	¿El cliente utiliza servicios de banca en línea? (No = 0, Sí = 1)
CreditCard	¿El cliente usa una tarjeta de crédito emitida por UniversalBank? (No = 0, Sí = 1)

Enlace: https://www.kaggle.com/code/pritech/bank-personal-loan-modelling

3. Modelo de Clasificación

Un modelo de clasificación es una fórmula estadística que separa los datos en grupos o categorías. Aprende de ejemplos previos y, cuando recibe un nuevo caso, dice a qué grupo pertenece. Por ejemplo: con datos de clientes, puede predecir si aceptará o no aceptará un préstamo.

Para este proyecto considero un árbol de decisión porque, a diferencia de otros modelos de clasificación, permite observar de forma clara las reglas que llevan a clasificar a un cliente, haciendo más transparente e interpretativo el proceso mostrando cómo a partir de variables como ingresos, edad o nivel educativo se construyen pasos lógicos que terminan en un resultado (“acepta” o “no acepta” el préstamo).

3.1 Conceptos básicos de un árbol de clasificación

Algunos conceptos fundamentales para entender el trabajo con un árbol de decisión para clasificación son:

Nodo: Es un punto de decisión dentro del árbol. En él se hace una pregunta sobre una variable (por ejemplo, “¿Ingreso > 50,000?”).

Raíz: Es el nodo inicial del árbol, donde comienza la primera decisión.

Hojas: Son los nodos finales, donde ya no se dividen más los datos. Cada hoja representa un resultado o clasificación.

Impureza: Mide qué tan “mezclados” están los datos en un nodo. Si hay de todo (clientes que aceptan y que no aceptan el préstamo), el nodo es impuro. Si todos son iguales (todos aceptan o todos rechazan), el nodo es puro.

Entropía: Es una forma de medir la impureza. A mayor entropía, más desorden hay en el nodo (mezcla de clases). A menor entropía, más orden (una sola clase domina).

Índice de Gini: Otra manera de medir la impureza. Dice la probabilidad de clasificar mal a un dato si se asignara aleatoriamente según la proporción de clases en el nodo. Cuanto más cercano a 0, más puro.

Ganancia: Es cuánto mejora la pureza al dividir un nodo. Si una división reduce bastante la mezcla de clases, tiene alta ganancia.

Profundidad: Qué tan “extenso” llega el árbol, es decir, la cantidad de decisiones desde la raíz hasta la hoja.

Regla de decisión: Es el camino de condiciones que sigue un dato desde la raíz hasta una hoja. Ejemplo: *Ingreso > 50,000 → Edad > 40 → Acepta el préstamo*.

3.2 Variables para el modelo

Para un modelo de clasificación es importante conocer las siguientes variables:

- Variable respuesta (Y):
  También llamada **variable dependiente** u **objetivo**. Es la que se quiere predecir. En este caso, corresponde a **Personal Loan**, que toma valor **0** si el cliente no aceptó el préstamo y **1** si sí lo aceptó.

- Variables predictoras (X):
  También llamadas **variables independientes** o **explicativas**. Son aquellas que se usan para intentar predecir la variable respuesta.
  Para este dataset serían:
  - Age
  - Income
  - Family
  - CCAvg
  - CreditCard

3.3 Balance en la variable respuesta

En este paso se observa qué proporción hay de cada categoría de la variable objetivo (por ejemplo, cuántos clientes aceptaron el préstamo y cuántos no). Este análisis inicial es sencillo pero clave, porque nos dice si los grupos están balanceados o si hay uno mucho más grande que otro, lo cual influye en cómo se construirá y se interpretará el modelo de clasificación.

En el dataset se puede observar un desbalance a favor de los clientes que no aceptaron el préstamo en la última campaña.

Personal Loan	Conteo	%
0	4520	90.4%
1	480	9.6%
Total general	**5000**	**100%**

Por lo general se aplican técnicas para balancear las clases, pero para este proyecto se procede a usar el conjunto de datos como se presenta.

3.4 Evaluar la impureza del primer nodo (nodo raíz)

el paso que sigue es **medir la pureza del nodo inicial**. Esto se hace aplicando un criterio como el **índice Gini**, que nos indican qué tan mezclados están los grupos en la variable objetivo. La idea es simple: si las clases están muy mezcladas, el valor de impureza será alto, y si están dominadas por un solo grupo, el valor será bajo.

Gini = $1 - \sum p_i^2$

Donde:

p_i^2 es la proporción de cada clase en el conjunto de datos

Entonces:

Gini = $1 - (0.904^2 + 0.096^2) = 0.174$

Interpretación: el 0.174 refleja una mezcla baja de clases, esto significa que el nodo raíz tiene una pureza alta, ya que la gran mayoría de los clientes pertenecen a la clase 0 (no aceptaron el préstamo).

3.5 Cortes en una variable predictora

El siguiente paso consiste en probar todas las variables predictoras, una por una, en las que se les calcula puntos de corte en sus datos.

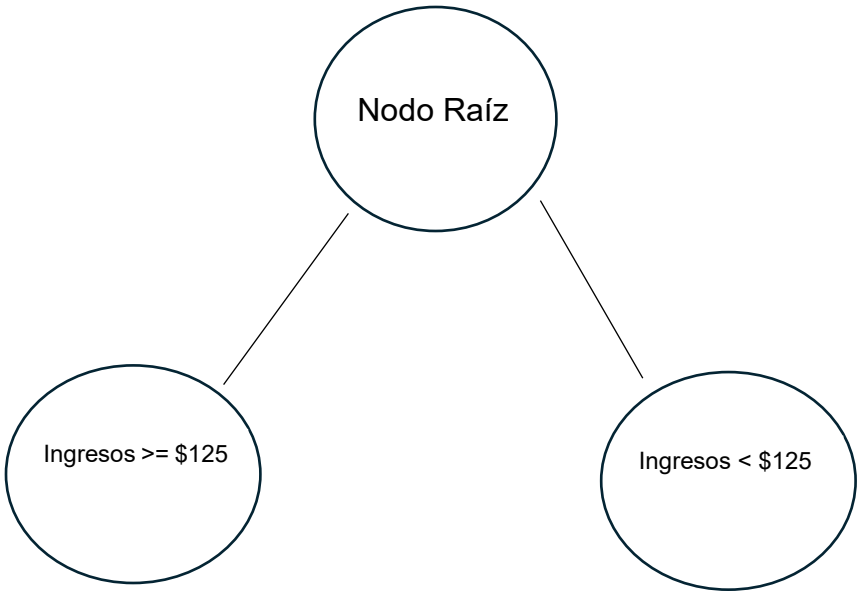
Primero se ordenan de menor a mayor y se calcula el punto medio entre cada valor:

$$t_j = \frac{x_j + x_{j+1}}{2}$$

Donde cada t_j es un posible corte

3.6 Separación de los datos y creación de los nodos hijos

Para este ejemplo se usará la variable Ingresos (Income) en el corte 125, esto significa que se dividirá el conjunto de datos en dos partes (nodos hijos), el nodo de la izquierda estarán los clientes que tengan ingresos iguales o superiores a $125 en el nodo de la izquierda los clientes que tengan menos de $125



3.7 Gini Ponderado y Ganancia

Gini Ponderado: mide cuán mezclados quedaron los dos hijos, teniendo en cuenta el tamaño de cada grupo.

Ganancia: es la mejora en pureza que se obtiene después de hacer un corte en el árbol. Se calcula comparando la pureza del nodo padre con la pureza promedio de los hijos. Si la ganancia es alta, significa que el corte logró que los nodos hijos quedasen mucho más puros que el padre; si es baja, el corte no ayudó a separar mejor las clases.

- Cálculo del Gini para cada nodo, recordando que:
  Gini = $1 - (p_1^2 + p_2^2)$

  Nodo Izquierdo:

Income >= 125	conteo	%
0	101	48.8%
1	106	51.2%
Grand Total	**207**	**100%**

$$Gini_{izq} = 1 - (0.488^2 + 0.512^2) = 0.4997$$

  Nodo Derecho:

Row Labels	conteo	%
0	4419	92.20%
1	374	7.80%
Grand Total	**4793**	**100%**

$$Gini_{der} = 1 - (0.922^2 + 0.078^2) = 0.1438$$

- Ambos Gini son necesarios para calcular el Gini Ponderado

  $$Gini_{ponderado} = \frac{n_{izq}}{n_{total}} * Gini_{izq} + \frac{n_{der}}{n_{total}} * Gini_{der}$$

  Donde n son el total de clientes o filas

  Entonces:

  $$Gini_{ponderado} = \frac{207}{5000} * 0.4997 + \frac{4793}{5000} * 0.1438 = 0.1585$$

  $$Gini_{ponderado} = 0.1585$$

- Para finalmente calcular la ganancia

  $$Ganancia = Gini_{padre} - Gini_{ponderado}$$

  Ganancia = 0.174 – 0.1585 = 0.0155

  Interpretación: La ganancia de 1.55% significa que, al dividir los datos, se logró mejorar un poco la claridad: la mayoría de los registros quedaron bien agrupados porque el grupo grande es muy puro, pero el grupo pequeño sigue mezclado y por eso la mejora total no es tan alta.

En estos cálculos se realizan con todos los cortes posibles de cada una de las variables para evaluar cuáles proporcionan la mejor opción de clasificación o separación de clases por nodo. Hacer estos cálculos es manualmente es prácticamente imposible, en cambio, con lenguajes de programación, se hace de forma automática: el algoritmo prueba todas las combinaciones de variables y cortes, calcula cuál da la mayor ganancia en pureza y elige el mejor.

Conclusión

El árbol de decisión permite observar cómo las matemáticas están en el corazón de la Ciencia de Datos. Conceptos como impureza, índice Gini, entropía y ganancia muestran que detrás de cada separación de datos hay un razonamiento estadístico que busca mejorar la capacidad de predicción del modelo.

Desde esta perspectiva, el modelado no es una “caja negra”, sino la unión de métodos estadísticos clásicos con la potencia de la programación. Así, se logra transformar datos en conocimiento y ofrecer modelos capaces de tomar decisiones útiles en la práctica.