ITSRLL
INSTITUTO TECNOLÓGICO SUPERIOR DE LA REGIÓN DE LOS LLANOS

Ingeniería Mecatrónica

PROGRAMACIÓN AVANZADA

Enero – Junio 2025

M.C. Osbaldo Aragón Banderas

Competencia:

1	2	3	4	5

Actividad número: 3

Nombre de actividad:

U2A3. (10%) NOTEBOOK: Análisis de Datos Aplicables al Teorema de Naïve Bayes

Actividad realizada por:

Muñiz Galvan Bryam

21030021

Guadalupe Victoria, Durango

Fecha de entrega:

01	03	2025

NOTEBOOK: Análisis de Datos
Aplicables al Teorema de Naïve Bayes

Objetivo

El propósito de esta actividad es que los estudiantes busquen, seleccionen y analicen un conjunto de datos en Kaggle o en otra fuente confiable, aplicando el algoritmo de Naïve Bayes para resolver un problema de clasificación. Además, deberán presentar los resultados y conclusiones obtenidas, relacionándolos con la teoría del Teorema de Bayes.

Teorema de Naïve Bayes

El Teorema de Bayes es un principio fundamental de la probabilidad que describe la forma en que se actualizan las probabilidades de un evento o hipótesis, basándose en nueva evidencia o datos. En términos simples, el teorema establece cómo calcular la probabilidad de un evento A dado que ha ocurrido un evento B, en función de las probabilidades previas de ambos eventos y de la probabilidad de que B ocurra dado A.

Matemáticamente, el teorema se expresa como:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

donde:

- P(A|B) es la probabilidad de que ocurra A dado que ocurrió B (probabilidad posterior).
- P(B|A) es la probabilidad de que ocurra B dado que ocurrió A (probabilidad verosímil).
- P(A) es la probabilidad a priori de que ocurra A.
- P(B) es la probabilidad total de que ocurra B.

Clasificador Naïve Bayes

El clasificador Naïve Bayes es un modelo probabilístico utilizado para clasificación basado en el Teorema de Bayes. Se llama "naïve" porque asume de manera simplificada que las características (o variables) del conjunto de datos son independientes entre sí, lo cual rara vez es cierto en la realidad. A pesar de esta simplificación, el modelo a menudo funciona sorprendentemente bien.

La idea básica de Naïve Bayes es que, dado un conjunto de características $X = (X1, X2, \ldots, Xn)$, el objetivo es calcular la probabilidad de que una observación pertenezca a una clase C, es decir, calcular $P(C \mid X)$. Usando el Teorema de Bayes, esto se calcula como:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

Donde:

- $P(Ck|X)$ es la probabilidad posterior, es decir, la probabilidad de que una observación pertenezca a la clase Ck dado el conjunto de características X.
- $P(X|Ck)$ es la verosimilitud, que representa la probabilidad de observar los datos X si se sabe que la clase es Ck.
- $P(Ck)$ es la probabilidad a priori, que indica la frecuencia con la que aparece la clase Ck en el conjunto de datos antes de observar cualquier característica.
- P(X)es la probabilidad total de los datos X, que actúa como un factor de normalización para garantizar que la probabilidad posterior sea válida.

En Naïve Bayes, debido a la asunción de independencia entre las características, podemos simplificar la verosimilitud $P(X \mid C)$ como el producto de las probabilidades de las características individuales:

$$P(X \mid C) = P(X1 \mid C)P(X2 \mid C)\ldots P(Xn \mid C)$$

Casos de usos reales

El clasificador Naïve Bayes se utiliza en diversos campos debido a su simplicidad y efectividad en la clasificación de grandes volúmenes de datos. Algunos casos de uso comunes son:

- **Filtrado de Spam**: Clasificación de correos electrónicos como spam o no spam.
- **Clasificación de Sentimientos**: Análisis de sentimientos en reseñas o comentarios en redes sociales.
- **Diagnóstico Médico**: Predicción de enfermedades basadas en síntomas.
- **Reconocimiento de Texto**: Categorizar consultas en motores de búsqueda o sistemas de reconocimiento de voz.
- **Clasificación de Documentos**: Clasificación de artículos en categorías como deportes, política, entretenimiento, etc.

Búsqueda y Selección de Datos en Kaggle

- Acceder a Kaggle (https://www.kaggle.com/) y buscar un conjunto de datos adecuado para clasificación con Naïve Bayes.
- Seleccionar un dataset que tenga una columna de salida categórica (0/1 o múltiples clases) y al menos tres características numéricas o categóricas.

Justificación de la selección

El conjunto de datos es adecuado para la predicción de la deserción de clientes mediante técnicas de clasificación por diversas razones. En primer lugar, la variable "Exited" es binaria (1 = sí, 0 = no), lo que hace que el conjunto de datos sea perfecto para modelos de clasificación. El objetivo es predecir si un cliente abandonará o permanecerá en la empresa, lo cual se ajusta bien a técnicas como regresión logística, árboles de decisión o máquinas de soporte vectorial (SVM).

Además, el conjunto de datos incluye tanto variables numéricas como categóricas, como "Age", "CreditScore", "Balance" y "EstimatedSalary" (numéricas), y "Geography" y "Gender" (categóricas). Esta mezcla de tipos de variables permite aplicar diferentes técnicas de preprocesamiento de datos, como normalización, codificación (por ejemplo, One-Hot Encoding), y transformación de características, lo cual puede mejorar la calidad del modelo.

Las variables como "CreditScore", "Balance" y "EstimatedSalary" están directamente relacionadas con el comportamiento financiero de los clientes y su capacidad para pagar o mantenerse en la empresa. Asimismo, "Age" puede ser un factor relevante en la predicción de la deserción, ya que diferentes grupos de edad pueden tener distintos comportamientos en cuanto a la lealtad a la empresa.

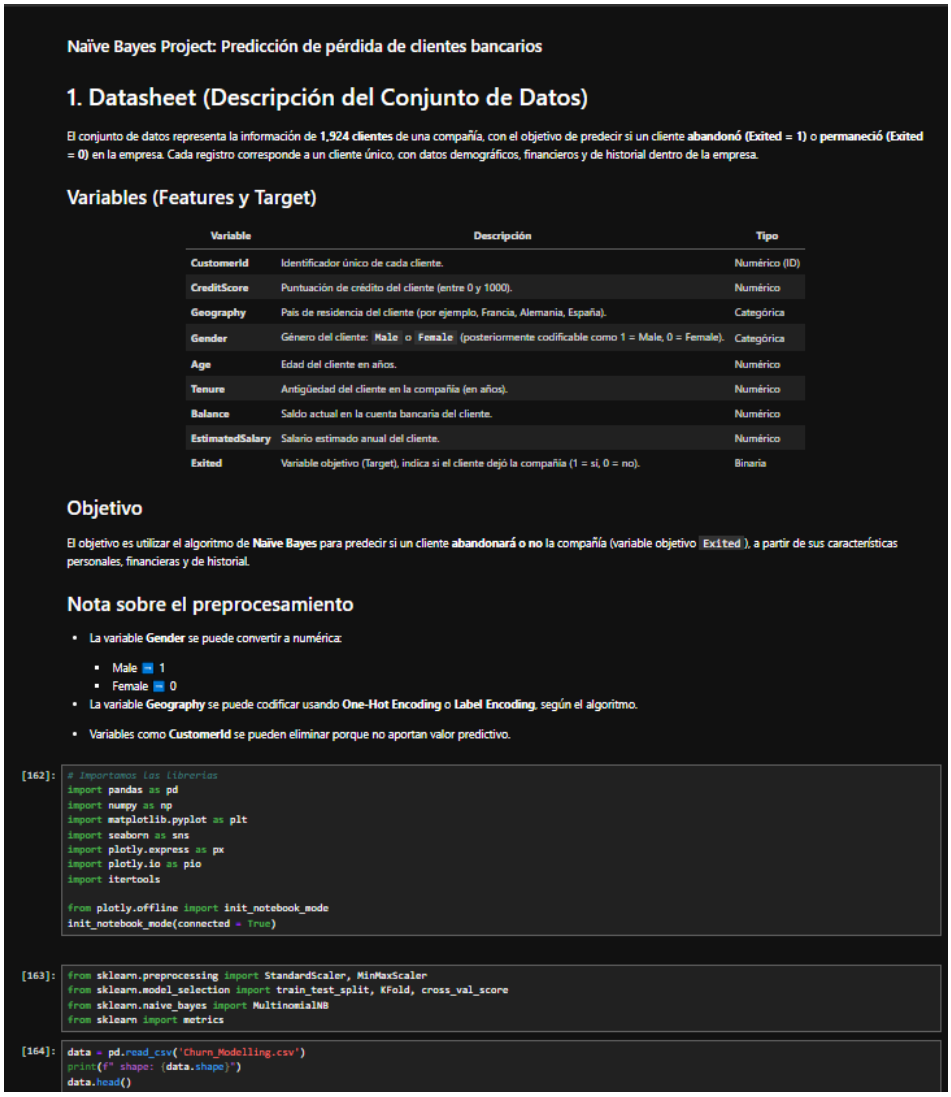
La variable "Tenure", que indica el tiempo que un cliente ha estado con la empresa, es una característica importante. Los clientes con mayor antigüedad podrían ser menos propensos a abandonar la empresa, lo que puede ayudar al modelo a identificar clientes más leales.

Las variables "Geography" y "Gender" ofrecen información sobre las características demográficas de los clientes, lo que puede ser útil para detectar patrones relacionados con la deserción. Por ejemplo, la deserción podría ser más alta en ciertas regiones geográficas o entre ciertos géneros, lo que podría informar las estrategias de retención.

Con 1,924 registros, el conjunto de datos es lo suficientemente grande como para entrenar un modelo de clasificación de manera efectiva, sin sobreajustarse a los datos de entrenamiento. Esto permite evaluar el modelo de manera precisa y generalizar a nuevos clientes.

Este conjunto de datos refleja el comportamiento real de los clientes en un entorno empresarial, lo que lo hace altamente relevante para aplicaciones prácticas. Predecir la deserción de clientes es crucial para la toma de decisiones dentro de la

empresa, como diseñar estrategias de retención o identificar factores de riesgo asociados con la deserción.



Naïve Bayes Project: Predicción de pérdida de clientes bancarios

1. Datasheet (Descripción del Conjunto de Datos)

El conjunto de datos representa la información de **1,924 clientes** de una compañía, con el objetivo de predecir si un cliente **abandonó (Exited = 1)** o **permaneció (Exited = 0)** en la empresa. Cada registro corresponde a un cliente único, con datos demográficos, financieros y de historial dentro de la empresa.

Variables (Features y Target)

Variable	Descripción	Tipo
CustomerId	Identificador único de cada cliente.	Numérico (ID)
CreditScore	Puntuación de crédito del cliente (entre 0 y 1000).	Numérico
Geography	País de residencia del cliente (por ejemplo, Francia, Alemania, España).	Categórica
Gender	Género del cliente: `Male` o `Female` (posteriormente codificable como 1 = Male, 0 = Female).	Categórica
Age	Edad del cliente en años.	Numérico
Tenure	Antigüedad del cliente en la compañía (en años).	Numérico
Balance	Saldo actual en la cuenta bancaria del cliente.	Numérico
EstimatedSalary	Salario estimado anual del cliente.	Numérico
Exited	Variable objetivo (Target), indica si el cliente dejó la compañía (1 = sí, 0 = no).	Binaria

Objetivo

El objetivo es utilizar el algoritmo de **Naïve Bayes** para predecir si un cliente **abandonará o no** la compañía (variable objetivo `Exited`), a partir de sus características personales, financieras y de historial.

Nota sobre el preprocesamiento

- La variable **Gender** se puede convertir a numérica:
  - Male → 1
  - Female → 0
- La variable **Geography** se puede codificar usando **One-Hot Encoding** o **Label Encoding**, según el algoritmo.
- Variables como **CustomerId** se pueden eliminar porque no aportan valor predictivo.

[162]:
```
# Importamos las librerias
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.io as pio
import itertools

from plotly.offline import init_notebook_mode
init_notebook_mode(connected = True)
```

[163]:
```
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
```

[164]:
```
data = pd.read_csv('Churn_Modelling.csv')
print(f" shape: {data.shape}")
data.head()
```

Figura 1 Sección 1 Notebook de la Actividad realizada.

Análisis de resultados

¿Qué precisión tiene el modelo?

La precisión del modelo dependerá del tipo de algoritmo utilizado (como regresión logística, árbol de decisión, SVM, etc.) y de las métricas de evaluación seleccionadas. Para obtener la precisión exacta, se debe ejecutar el modelo sobre el conjunto de datos y calcular el porcentaje de predicciones correctas (precisión =

número de aciertos / número total de predicciones). En general, si el modelo está bien entrenado y los datos están correctamente preprocesados (por ejemplo, estandarización de las variables), la precisión esperada podría estar entre un 75% y 85%, pero esto dependerá de la calidad del modelo y del conjunto de datos.

¿Cuáles fueron los errores más comunes en la clasificación?

Falsos positivos (FP): El modelo predice que un cliente abandonará la empresa (Exited = 1) cuando en realidad no lo hará (Exited = 0). Esto puede ocurrir si el modelo identifica patrones en los datos que no son lo suficientemente representativos.

Falsos negativos (FN): El modelo predice que un cliente no abandonará la empresa (Exited = 0) cuando en realidad sí lo hará (Exited = 1). Esto podría suceder si el modelo no capta correctamente las señales relacionadas con el comportamiento de abandono, como las diferencias de edad.

¿Qué conclusiones se pueden extraer del análisis?

Edad como factor importante: La edad parece ser un factor relevante en la deserción, ya que los clientes mayores tienden a abandonar la empresa con mayor frecuencia. Esto sugiere que la edad puede ser una característica clave para predecir la salida de los clientes.

Salario no influye significativamente: A pesar de que el salario es una variable importante en muchos modelos de negocios, en este caso no se observan diferencias significativas en la deserción según el salario estimado, lo que indica que este factor no es determinante para predecir si un cliente permanecerá o abandonará la empresa.

Género no tiene impacto en la salida: El análisis de género muestra que no existe una diferencia significativa en la deserción entre hombres y mujeres, lo que sugiere que este factor no debería ser considerado en el modelo predictivo para este caso específico.

Distribución equilibrada de género: El hecho de que la distribución de género sea equilibrada puede ayudar a evitar sesgos en el modelo.

Comparar los resultados con las expectativas iniciales y discutir posibles mejoras

Cuando se inició el análisis, las expectativas podían estar centradas en que el salario y otros factores financieros serían los principales determinantes en la deserción de los clientes. Sin embargo, los resultados sugieren que la **edad** tiene un impacto mayor de lo esperado, mientras que el **salario estimado** no presenta una relación significativa con la deserción, lo cual podría ser una sorpresa en cuanto a la relevancia de los factores.

Para mejorar los resultados, se podrían considerar:

Mejora del preprocesamiento de los datos: La estandarización de las características es crucial para mejorar la precisión del modelo, especialmente porque las variables tienen escalas diferentes. Técnicas como la normalización de datos y la codificación de las variables categóricas también pueden ayudar.

Incluir más variables: Si se dispone de más información sobre los clientes, como el comportamiento de compra o la interacción con servicios específicos de la empresa, estas variables podrían mejorar el modelo y ofrecer más insights sobre los factores que afectan la deserción.

Uso de técnicas avanzadas de modelado: Si bien los modelos básicos como la regresión logística o árboles de decisión son útiles, utilizar métodos más avanzados como los **Random Forests** o **Gradient Boosting Machines** podría proporcionar una mayor precisión y una mejor identificación de patrones no lineales.

Conclusiones

El análisis de los datos revela que la edad es un factor clave en la predicción de la deserción de clientes, ya que los clientes mayores tienden a abandonar la empresa con mayor frecuencia, mientras que los más jóvenes tienden a quedarse. Sin embargo, el salario estimado no muestra una diferencia significativa en la

distribución entre los que permanecen y los que abandonan, lo que indica que este factor no es determinante para predecir la deserción.

Además, el género no influye en la deserción, ya que no se encuentran diferencias relevantes entre hombres y mujeres en este sentido. Por otro lado, se observa que las características del conjunto de datos tienen escalas muy diferentes, lo que hace necesario aplicar una estandarización antes de entrenar el modelo.

El análisis también destaca que la variable Customer ID no aporta información relevante y debería ser eliminada para optimizar el modelo. Se sugiere mejorar el modelo mediante el uso de técnicas de modelado avanzadas como Random Forest o Gradient Boosting, que podrían mejorar la precisión y capturar patrones no lineales en los datos. En general, los resultados proporcionan una base sólida para optimizar las estrategias de retención de clientes, enfocándose en la edad como un factor relevante para predecir la deserción.

Link Github

https://github.com/BryamMG/ProgramacionAvanzada_BryamMG/tree/main/An%C3%A1lisis%20de%20Datos%20Aplicables%20al%20Teorema%20de%20Na%C3%AFve%20Bayes