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	08	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

Naïve Bayes Project: Predicción de pérdida de clientes bancarios

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 9,996 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.

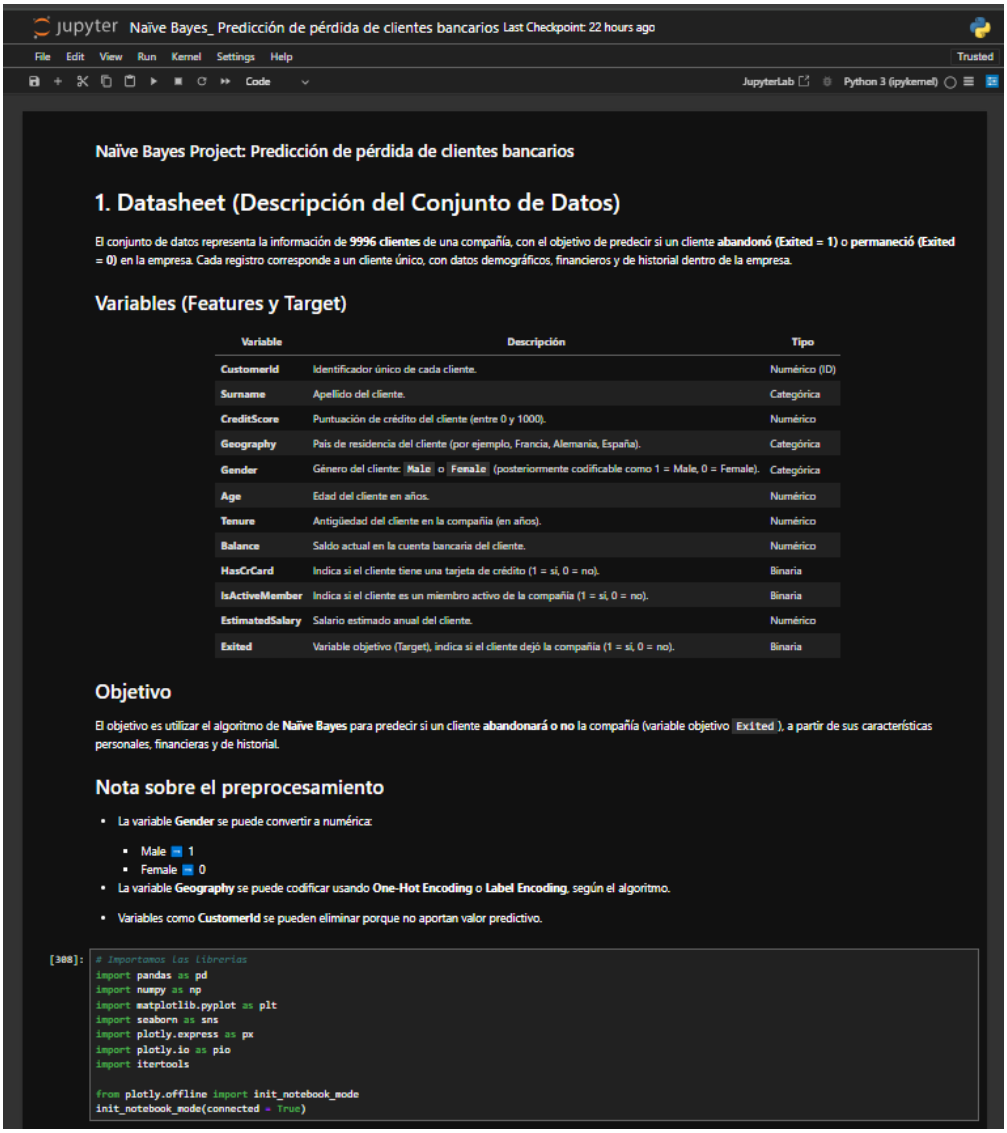

Figura 1 Sección principal Notebook de la Actividad realizada.

Análisis de resultados

¿Qué precisión tiene el modelo?

El modelo final tiene una precisión del 70%, lo que indica que el 70% de las predicciones fueron correctas. Aunque ha mejorado con respecto a versiones anteriores, todavía hay margen de mejora, especialmente en la detección de clientes que abandonan.

¿Cuáles fueron los errores más comunes en la clasificación?

Los principales errores fueron los falsos negativos, es decir, clientes que el modelo predijo que se quedarían, pero en realidad se fueron. Además, la precisión para la clase "Exited (1)" fue solo del 37%, lo que muestra dificultades en la detección de estos clientes. La principal causa es el desbalance de clases, ya que solo el 20% de los clientes han abandonado, haciendo que el modelo favorezca la predicción de la clase mayoritaria ("No Exited").

¿Qué conclusiones se pueden extraer del análisis?

El factor más importante en la predicción de abandono es la edad y la actividad del cliente, mientras que el salario y el género no influyen significativamente. A pesar de un accuracy del 70%, la baja precisión en la clase "Exited (1)" indica que el modelo aún no es totalmente fiable para detectar clientes que se irán.

Comparar los resultados con las expectativas iniciales y discutir posibles mejoras

Inicialmente se esperaba que factores como el salario o la puntuación de crédito fueran determinantes en la salida de clientes, pero el análisis demostró que la edad y la actividad tienen mayor peso. Para mejorar el modelo, se pueden aplicar técnicas de balanceo de datos como SMOTE, probar modelos más avanzados como Random Forest o XGBoost, optimizar hiperparámetros y considerar nuevas variables que ayuden a capturar mejor los patrones de abandono.

Conclusiones

El análisis de los datos ha revelado que la edad y la actividad del cliente son los factores más influyentes en la retención, mientras que variables como el salario y la puntuación de crédito tienen un impacto menor. A pesar de que el modelo final alcanzó un 70% de precisión, aún presenta dificultades para identificar correctamente a los clientes que abandonan, lo que se debe en gran parte al desbalance de clases.

Para mejorar la capacidad predictiva del modelo, es necesario implementar estrategias como balanceo de datos, ajuste de hiperparámetros y el uso de algoritmos más avanzados. Además, sería recomendable explorar nuevas variables que ayuden a identificar con mayor precisión los patrones de abandono, permitiendo a la empresa diseñar estrategias más efectivas para la retención de clientes.

Link Github

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