ITSRLL
INSTITUTO TECNOLÓGICO SUPERIOR DE LA REGIÓN DE LOS LLANOS

Ingeniería Mecatrónica

PROGRAMACIÓN AVANZADA

Enero – Junio 2025

M.C. Osbaldo Aragón Banderas

UNIDAD:

1	**2**	3	4	5

Actividad número: 3

Nombre de actividad:

NOTEBOOK: Análisis de Datos Aplicables al Teorema de Naïve Bayes

Actividad realizada por:

Melissa Gómez Rentería.

Guadalupe Victoria, Durango

Fecha de entrega:

01	03	2025

¿QUÉ ES EL TEOREMA DE NAIVE BAYES?

El Teorema de Bayes es una regla matemática que nos permite calcular la probabilidad de que algo ocurra basándonos en información previa. En términos simples, nos ayuda a actualizar nuestras creencias cuando obtenemos nueva información.

Por ejemplo, si se escucha que va a llover mañana. Antes de revisar el clima, surge una idea general de qué tan probable es que llueva basándose en la época del año. Sin embargo, si luego logra apreciarse que el cielo está nublado y el viento está fuerte, esa información nueva puede hacer pensar que la lluvia es aún más probable. Eso es, en esencia, lo que hace el Teorema de Bayes, ajustar la probabilidad de un evento con base en nueva evidencia.

ECUACIÓN GENERAL DE NAÏVE BAYES

La fórmula matemática es la siguiente:

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

- P(A | B): es la probabilidad posterior, es decir, qué tan probable es que ocurra A después de haber observado B.
- P(B | A) es la verosimilitud, es decir, qué tan probable es observar B si A es cierto.
- P(A) es la probabilidad previa, es decir, qué tan probable era A antes de considerar B.
- P(B) es la probabilidad total de que ocurra B, sin importar si A ocurre o no.

Ahora, el clasificador Naïve Bayes se basa en este teorema para hacer predicciones, pero con una suposición clave: asume que todas las características son independientes entre sí. Esto no siempre es cierto en la vida real, pero simplifica mucho los cálculos y hace que el modelo sea rápido y eficiente.

La ecuación general del Naïve Bayes se expresa así:

$$P(C|X_1, X_2, ..., X_n) = P(C) \cdot \prod_{i=1}^{n} P(X_i|C)$$

¿Qué significa cada parte de esta ecuación?

$P(C \mid X_1, X_2, ..., X_n)$ → La probabilidad de que un dato pertenezca a una categoría específica dado un conjunto de características observadas. Por ejemplo, si estamos clasificando correos, esta sería la probabilidad de que un correo sea spam basándonos en palabras clave y otros factores.

$P(C)$ → La probabilidad previa de la categoría. Si estamos clasificando enfermedades, esto sería qué tan común es esa enfermedad en la población en general.

$P(X_i \mid C)$ → La probabilidad de observar cada característica X_i si estamos en la categoría C. Por ejemplo, si una persona tiene fiebre, ¿qué tan probable es que tenga gripe?

Para clasificar un nuevo dato, el modelo calcula esta probabilidad para cada posible categoría y elige la que tenga el mayor valor.

CASOS DE USO REALES

Este modelo es muy eficiente en muchos problemas del mundo real porque es rápido, fácil de implementar y no requiere tantos datos como otros algoritmos más complejos.

1. Filtrado de spam en correos electrónicos

Servicios como Gmail usan Naïve Bayes para clasificar automáticamente los correos como spam o no spam. Analizan palabras clave, la dirección del remitente y otros factores para calcular la probabilidad de que un correo sea spam.

2. Diagnóstico médico

Se usa en hospitales para predecir la probabilidad de que un paciente tenga una enfermedad basándose en síntomas y antecedentes médicos. Por ejemplo, puede ayudar a diagnosticar diabetes analizando factores como nivel de azúcar en sangre, presión arterial y peso.

3. Clasificación de opiniones en redes sociales

Empresas de marketing analizan comentarios y reseñas de productos para detectar si son positivos, negativos o neutrales. Esto se usa en redes sociales como Twitter o en plataformas de reseñas como Amazon.

4. Sistemas de recomendación

Naïve Bayes ayuda a predecir qué productos o películas podrían interesarte basándose en tus preferencias pasadas. Por ejemplo, Netflix puede clasificar géneros de películas que podrían gustarte según las que has visto antes.

5. Detección de fraudes en transacciones bancarias

Los bancos utilizan este modelo para detectar actividades sospechosas en cuentas bancarias. Si una compra parece inusual con respecto al historial del usuario, el banco puede marcar la transacción como potencialmente fraudulenta.

ELECCIÓN DE DATASET (DIAGNÓSTICO DE SOP)

El dataset tiene una variable objetivo bien definida, que es la columna "PCOS (Y/N)":

- 1: significa que la paciente tiene SOP.
- 0: significa que no tiene SOP.

El Naïve Bayes es un modelo de clasificación que funciona muy bien cuando hay que decidir entre dos categorías basándose en varias características. En este caso, nos ayuda a predecir si una paciente tiene SOP o no, basándose en sus datos médicos.

El dataset tiene varias características médicas relevantes, para hacer un buen diagnóstico, se necesita información sobre varios factores de salud. Este dataset incluye características como:

- Edad de la paciente
- Índice de Masa Corporal (IMC)
- Niveles hormonales (testosterona)
- Frecuencia del ciclo menstrual
- Historial de fertilidad

Tener varias características hace que el modelo pueda encontrar patrones y mejorar la precisión del diagnóstico.

El modelo de Naïve Bayes se usa mucho en el campo de la medicina porque es fácil de entrenar y rápido pues no necesita muchos datos ni mucho poder de cómputo, además que maneja bien diferentes tipos de datos y puede trabajar con números (como el IMC) y con valores categóricos (como "Sí/No" en historial menstrual).

A diferencia de modelos más complejos, como redes neuronales, Naïve Bayes permite ver claramente qué factores influyen más en la predicción.

Algunos modelos avanzados, como redes neuronales, requieren mucha potencia de procesamiento y grandes volúmenes de datos para funcionar bien.

En cambio, Naïve Bayes puede entrenarse con menos datos y sigue siendo preciso.

Este dataset no es demasiado grande ni demasiado pequeño, lo que lo hace perfecto para este tipo de modelo.

LINK DE CONJUNTOS DE DATOS

https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset