Laboratorio 1: La maldición de la dimensionalidad

Prof. Rosa Yuliana Gabriela Paccotacya Yanque rpaccotacya@unsa.edu.pe

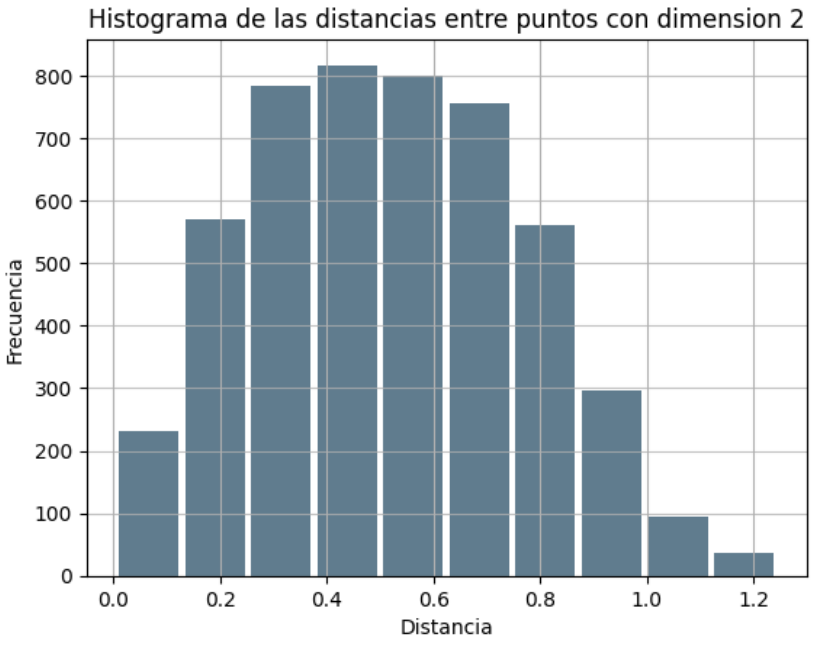
Objetivo

El objetivo de este laboratorio es analizar cómo el espacio cambia a medida que la dimensionalidad (cantidad de atributos o features) de los datos aumenta y cuán desafiante esto puede ser.

Descripción del laboratorio

Realizaremos diversos experimentos en C++ donde trabajaremos con conjuntos de datos de diferentes dimensiones (10, 50, 100, 500, 1000, 2000, 5000). Para cada conjunto de datos de determinada dimensión d se debe:

- Generar **100 puntos aleatorios** entre 0 y 1 de dimensión d (Hint:https://en.cppreference.com/w/cpp/numeric/random/uniform_real_distribution)
- Calcular la distancia entre todos los pares de puntos (Distancia Euclidiana) (Hint 4950 distancias)
- Generar un histograma (pueden usar Python) de las distancias obtenidas para cada dimensión como el de la figura mostrada a continuación:



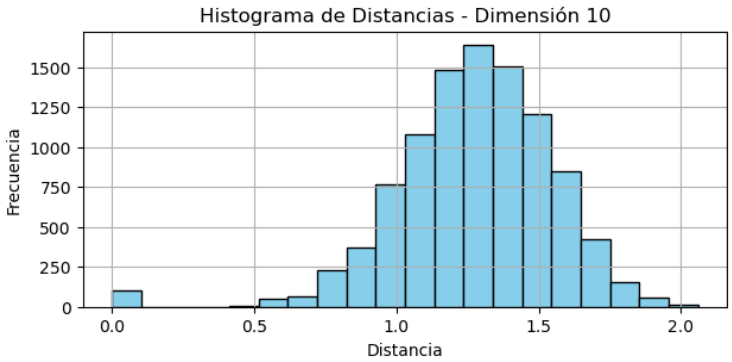
Entregable:

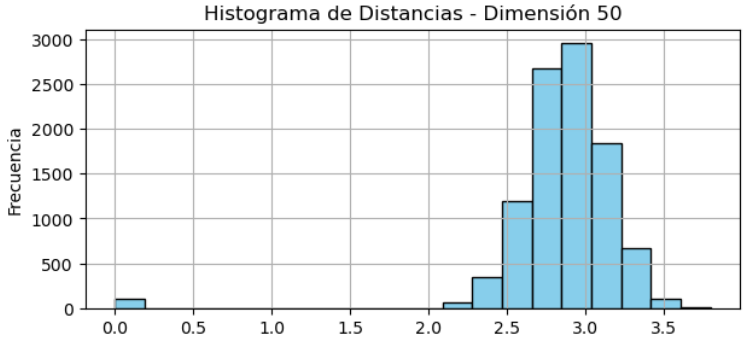
- Informe conteniendo las gráficas obtenidas para cada dimensión: 10, 50, 100, 500, 1000, 2000, 5000 y el análisis correspondiente.
    - Al final del informe incluir un link a github donde se encuentre el código e informe.
    - El informe debe ser detallado explicando que está ocurriendo (analizar los cambios en la distribución de las distancias, los valores en los ejes x y y, etc.) y deben incluir citas bibliográficas en caso las necesiten.
- **Fecha límite de entrega: 17/09/2023 10:00pm**

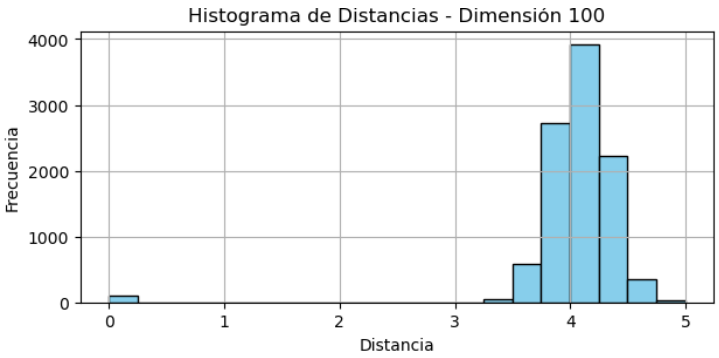
Evaluación:

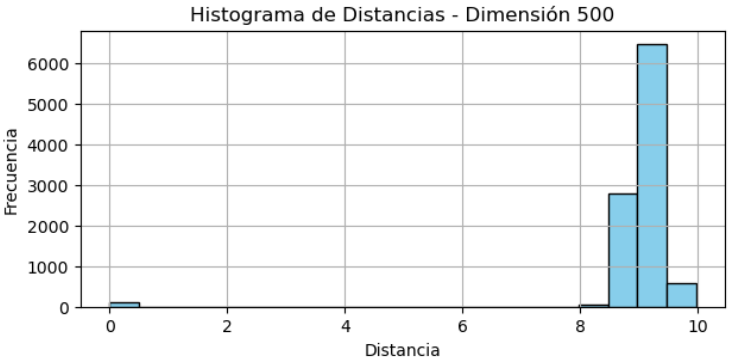
- Informe
- Código legible
- No exista plagio o sus variantes
    - En caso de copia o plagio o similares, todos los alumnos implicados tendrán nota 0 en este laboratorio y sanción en toda la evaluación de los laboratorios.

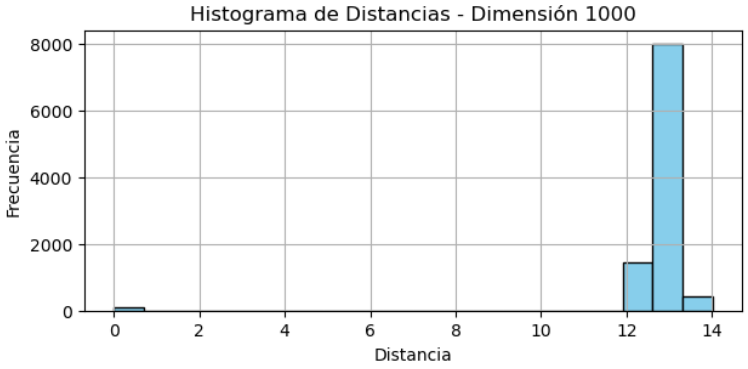
Resultados esperados:

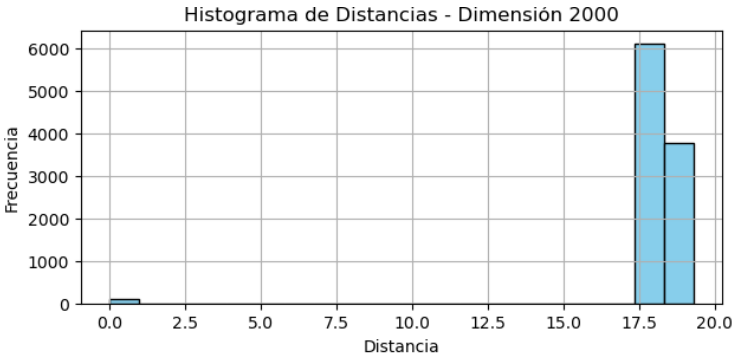


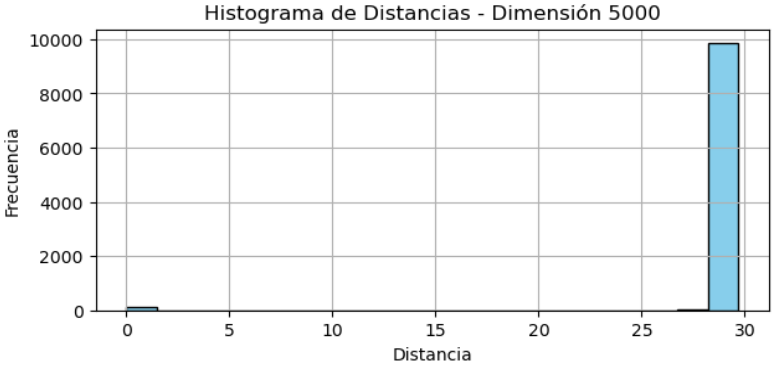












ANÁLISIS:

Se logra observar que a medida que aumenta la dimensionalidad se ven menos barras en el histograma un fenómeno relacionado con la "**maldición de la dimensionalidad".** Esto ocurre debido a que, a medida que la dimensionalidad aumenta, a medida que el número de entidades aumenta para un número dado de observaciones, el espacio de características se vuelve cada vez más escaso; Es decir, menos denso o más vacío. Por otro lado, la menor densidad de datos requiere más observaciones para mantener igual la distancia promedio entre los puntos de datos.

Cuando la distancia entre las observaciones aumenta, el aprendizaje automático supervisado se vuelve más difícil porque las predicciones para nuevas muestras tienen menos probabilidades de basarse en el aprendizaje de características de entrenamiento similares. El número de filas únicas posibles crece exponencialmente a medida que aumenta el número de entidades, lo que hace que sea mucho más difícil generalizar de manera eficiente. La varianza aumenta a medida que tienen más oportunidad de adaptarse al ruido en más dimensiones, lo que da como resultado un rendimiento de generalización deficiente.

En la práctica, las características están correlacionadas o no muestran mucha variación. Por estas razones, la reducción de la dimensionalidad ayuda a comprimir los datos sin perder gran parte de la señal, y combate la maldición a la vez que ahorra en la memoria.

References

Curse of Dimensionality. (2019, December 24). sitiobigdata.com. Retrieved September 20, 2023, from

https://sitiobigdata.com/2019/12/24/curse-of-dimensionality/#

LINK GITHUB:

https://github.com/ANTHONYCCOLQUE/EDA2023