



UNIVERSIDAD NACIONAL
DE SAN AGUSTÍN



FACULTAD DE INGENIERÍA, PRODUCCIÓN Y SERVICIOS

ESCUELA PROFESIONAL DE CIENCIA DE LA
COMPUTACIÓN

La Maldición de la Dimensionalidad

Alumno:

Torres Quispe Erick Jesus

Docente:

Paccotacya Yanque Rosa Yuliana Gabriela

20 de septiembre de 2023

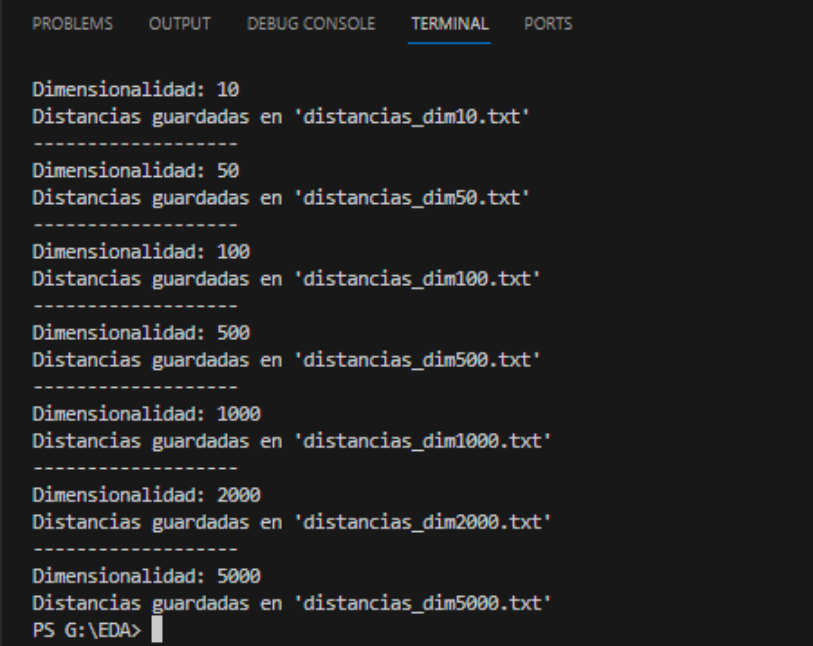
1. Introducción

La maldición de la dimensionalidad es un fenómeno que se manifiesta cuando trabajamos con datos en espacios de alta dimensionalidad. Se caracteriza por la disminución de la utilidad de las medidas de distancia y similitud, como la distancia euclidiana, a medida que aumenta la dimensionalidad de los datos. En este informe, se presentarán los resultados del análisis realizado a través de dos códigos interconectados: uno en C++ para generar datos y calcular distancias, y otro en Google Colab (Python) para visualizar las distribuciones de estas distancias en diferentes dimensionalidades.

2. Generación de Datos y Cálculo de Distancias en C++

El código en C++ se encargó de la generación de datos y el cálculo de las distancias euclidianas entre los puntos en diferentes dimensionalidades. A continuación, se resumen los principales componentes del código y sus resultados:

1. Generación de Datos Aleatorios: El código generó 100 puntos aleatorios en cada dimensión especificada, con valores en el rango de 0 a 1.
2. Cálculo de Distancias:
Se calculó la distancia euclidiana entre todos los pares únicos de puntos generados. Se utilizaron fórmulas matemáticas para calcular el número total de distancias a calcular.
3. Escritura en Archivos de Texto:
Las distancias resultantes se guardaron en archivos de texto separados, uno para cada dimensión.
4. Resultados del Código en C++:
 - Se generaron conjuntos de datos en diferentes dimensionalidades (10, 50, 100, 500, 1000, 2000, 5000).
 - Se calcularon y guardaron las distancias euclidianas entre todos los pares de puntos en cada conjunto de datos.
 - Los archivos de texto resultantes se utilizaron como entrada para el análisis en Google Colab.
5. Captura de pantalla del proceso de compilación en C++:



```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

Dimensionalidad: 10
Distancias guardadas en 'distancias_dim10.txt'
-----
Dimensionalidad: 50
Distancias guardadas en 'distancias_dim50.txt'
-----
Dimensionalidad: 100
Distancias guardadas en 'distancias_dim100.txt'
-----
Dimensionalidad: 500
Distancias guardadas en 'distancias_dim500.txt'
-----
Dimensionalidad: 1000
Distancias guardadas en 'distancias_dim1000.txt'
-----
Dimensionalidad: 2000
Distancias guardadas en 'distancias_dim2000.txt'
-----
Dimensionalidad: 5000
Distancias guardadas en 'distancias_dim5000.txt'
PS G:\EDA>
```

Figura 1: Proceso de Compilación

3. Visualización de Histogramas en Google Colab

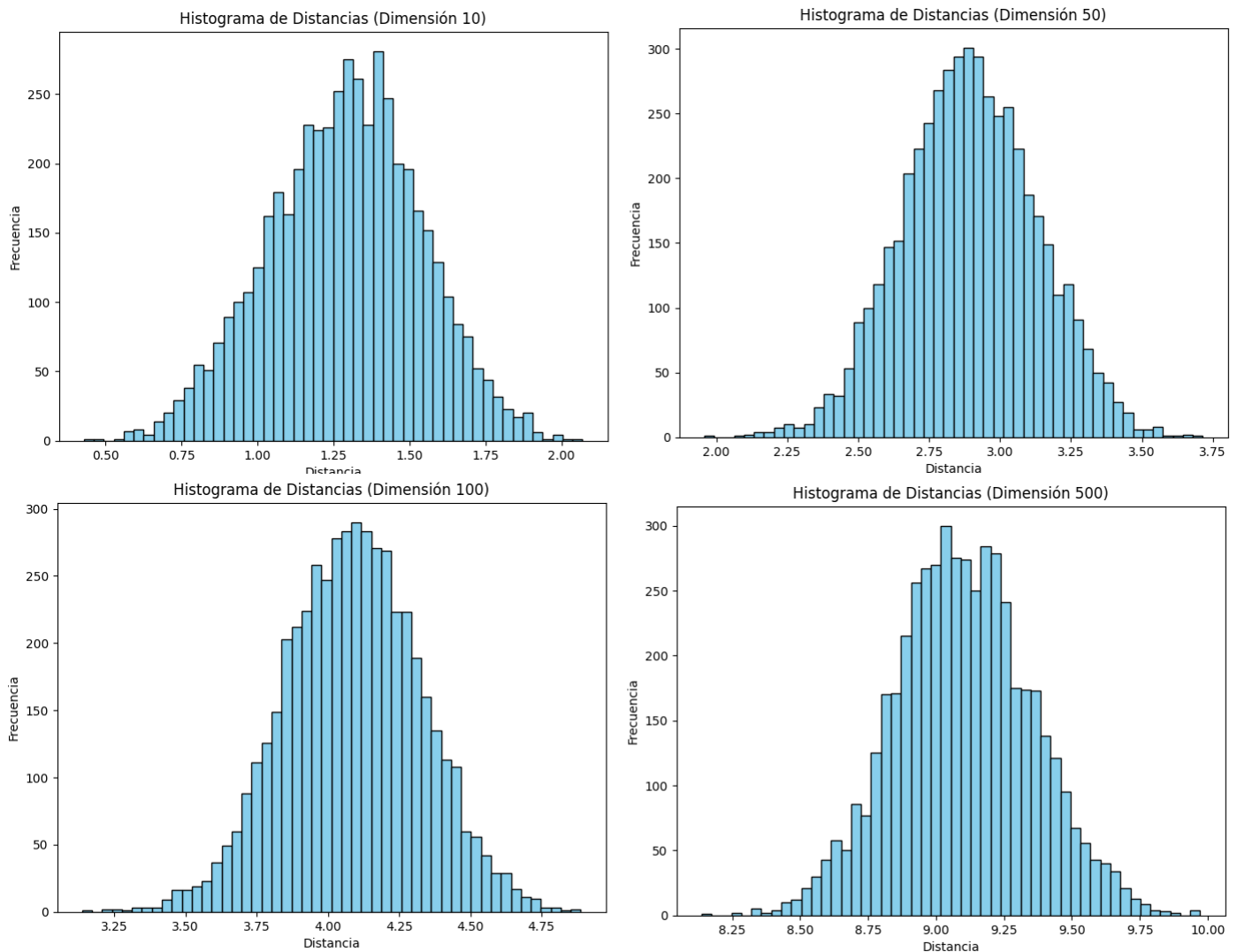
El código en Google Colab se centró en cargar los archivos de distancias generados en C++ y en crear histogramas para visualizar las distribuciones de estas distancias. A continuación, se resumen los componentes clave del código y sus resultados:

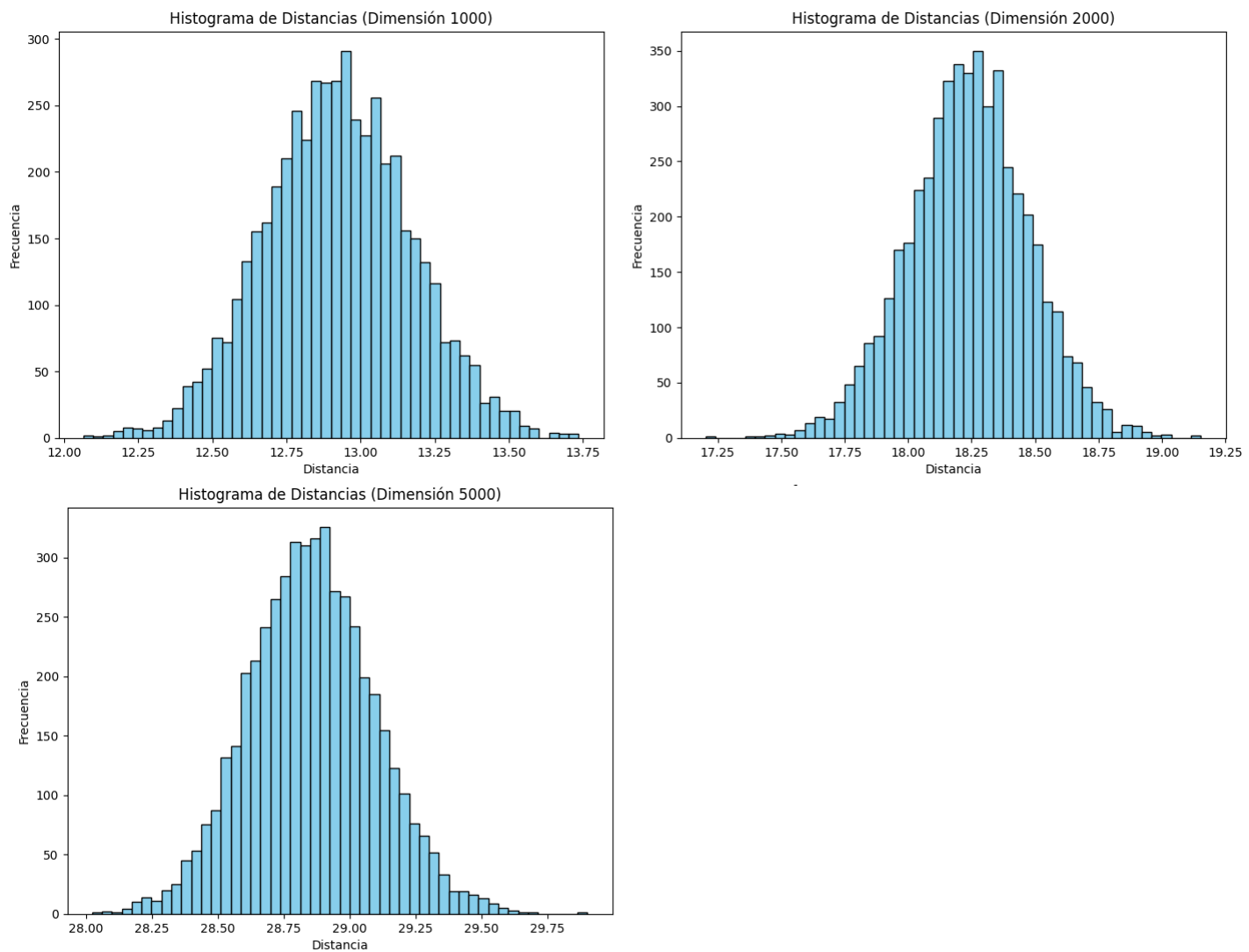
- **Carga de Archivos de Distancias:** Los archivos de distancias generados en C++ se cargaron en el entorno de Google Colab para su procesamiento.
- **Creación de Histogramas:** Se generaron histogramas para cada dimensión de datos utilizando la biblioteca *matplotlib* en Python.

Cada histograma representó la distribución de las distancias calculadas en esa dimensión.

- **Resultados del Código en Colab:** Se visualizaron las distribuciones de distancias en forma de histogramas para cada dimensionalidad.

Cada histograma proporcionó información sobre cómo cambian las distancias a medida que aumenta la dimensionalidad.





4. Análisis de Resultados

4.1. Distribución de Distancias

En las dimensiones más bajas, como 10 y 50, las distancias parecen seguir una distribución más cercana a una campana de Gauss (normal) con un pico en torno a un valor central. A medida que la dimensionalidad aumenta, las distancias tienden a extenderse y volverse más uniformes en su distribución. Esto significa que las distancias entre puntos tienden a ser similares y menos discriminativas.

4.2. Valores en los Ejes

En el eje "x" de los histogramas, se encuentran los valores de las distancias calculadas. A medida que la dimensionalidad aumenta, es posible observar una mayor dispersión de los valores en el eje x.

En el eje "y" de los histogramas, se representa la frecuencia con la que aparecen las distancias en cada intervalo. En dimensiones más bajas, se observan picos más marcados, lo que indica una mayor concentración de distancias alrededor de ciertos valores. En dimensiones más altas, la frecuencia tiende a dispersarse, lo que sugiere que las distancias tienen menos valores comunes.

5. Conclusiones

El análisis de la maldición de la dimensionalidad a través de estos códigos revela que a medida que aumenta la dimensionalidad de los datos, las distancias euclidianas tienden a volverse menos distintivas. Esto significa que en espacios de alta dimensionalidad, las distancias entre puntos se vuelven más uniformes, lo que puede dificultar la identificación de similitudes y la separación de datos.

Estos resultados tienen importantes implicaciones en campos como el aprendizaje automático y la minería de datos, donde la selección y el procesamiento de características pueden ser críticos para abordar el problema de la dimensionalidad y evitar la maldición de la dimensionalidad.

En resumen, el análisis realizado a través de estos códigos proporciona una comprensión más profunda de cómo cambian las distancias en función de la dimensionalidad de los datos y destaca los desafíos que surgen al trabajar en espacios de alta dimensionalidad.