

Laboratorio 1: La maldición de la dimensionalidad

Cristian Ramos Medina Ccomp6-1

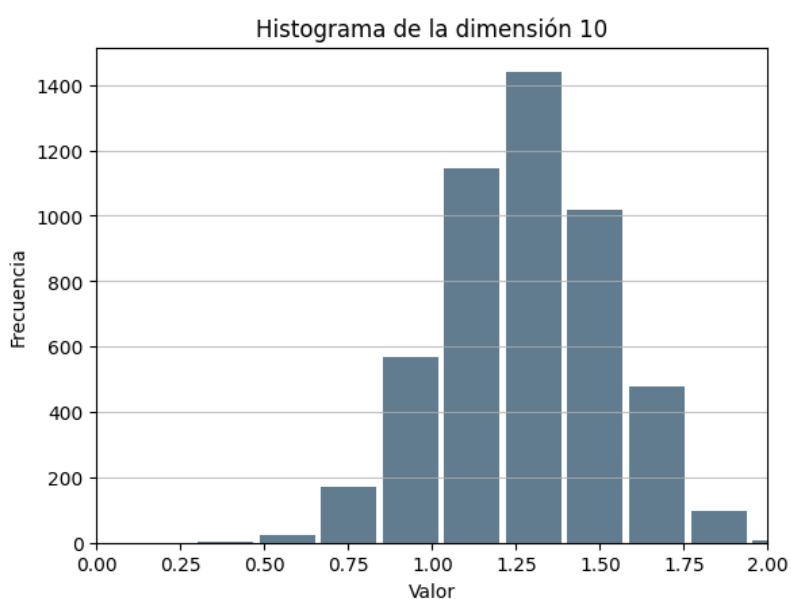
cristian.ramos@ucsp.edu.pe

Objetivo Cumplido

En este informe se mostrará la tendencia de crecimiento de espacio que se logró apreciar durante el experimento, al ir subiendo la dimensionalidad de los puntos, el espacio calculado entre los mismos se hace más y más grande; empezando en una tendencia en apenas 10 dimensiones de 0.5 a 2.00, hasta 27.00 hasta 30 en 5000 dimensiones.

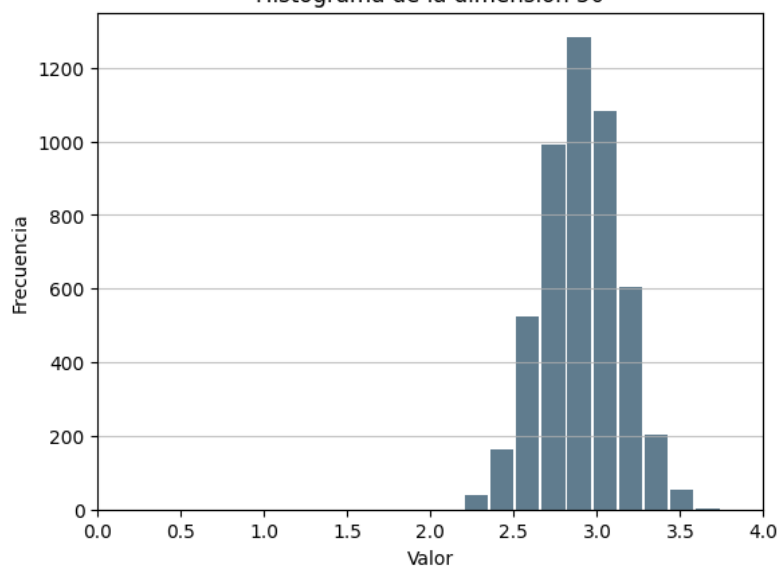
Resultados

A continuación las imágenes de los histogramas generados en python(google colab, código subido en el repositorio github) utilizando 'pandas' para su visualización adaptando el código que se nos proporcionó en clase, ajustando en algunos casos el rwidth para que no se junten nuestras líneas del histograma, todos los histogramas tendrán la limitación de rango a partir del 0 hasta un aproximado a su máximo valor.



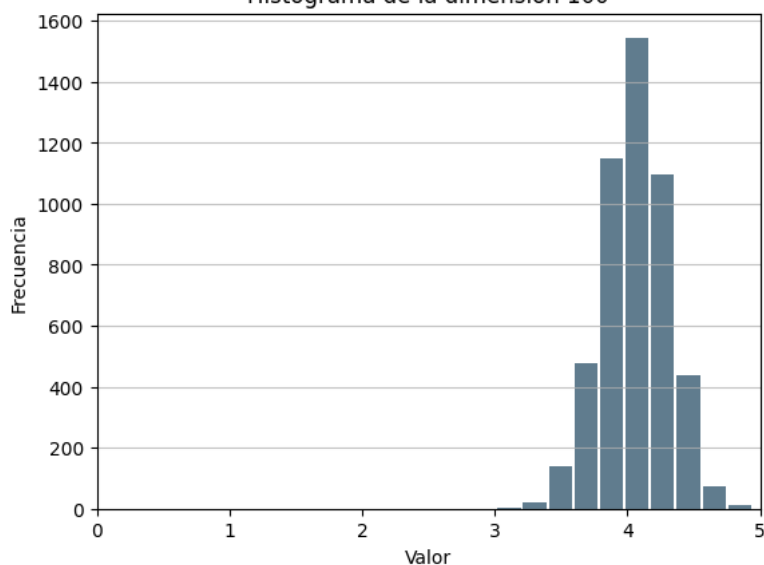
Datos de representación 10D :
cantidad de bins = 10, rwidth=0.9,
limitación de rango de 0.00 : 2.00.

Histograma de la dimensión 50



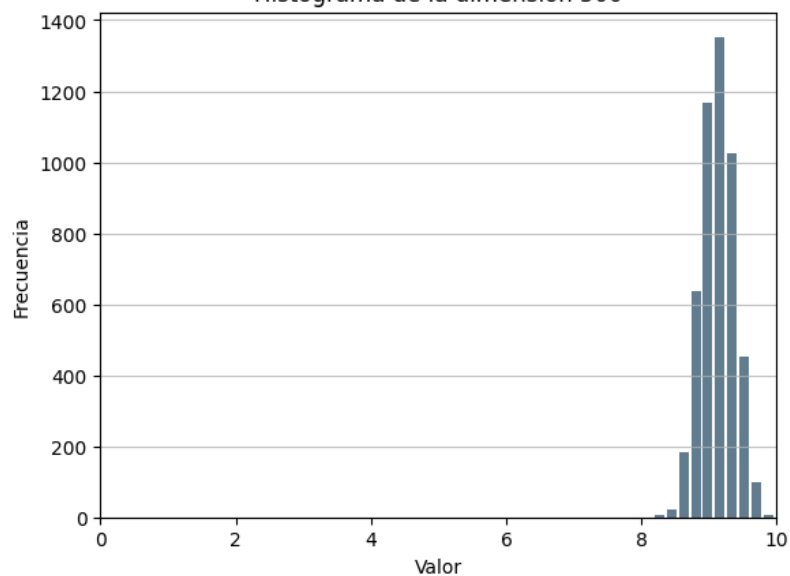
Datos de representación 50D : cantidad de bins = 10, rwidth=0.9, limitación de rango de 0.00 : 4.00.

Histograma de la dimensión 100



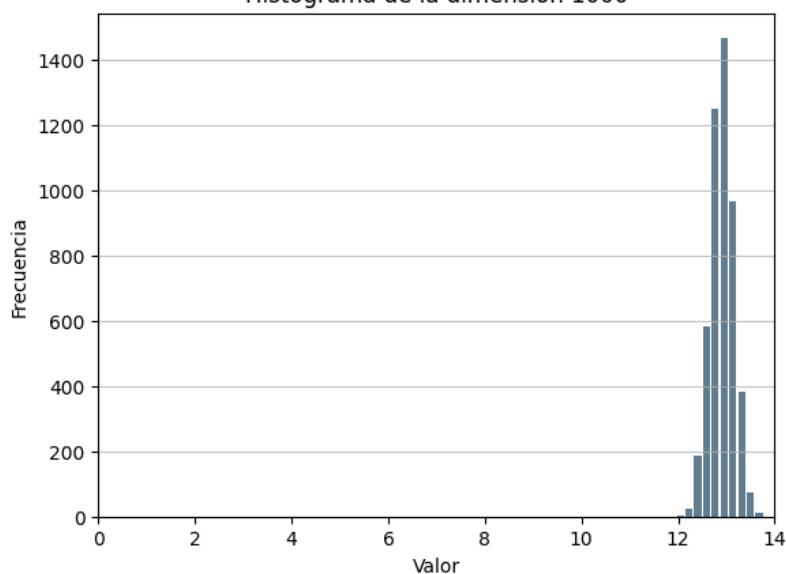
Datos de representación 100D : cantidad de bins = 10, rwidth=0.9, limitación de rango de 0.00 : 5.00.

Histograma de la dimensión 500



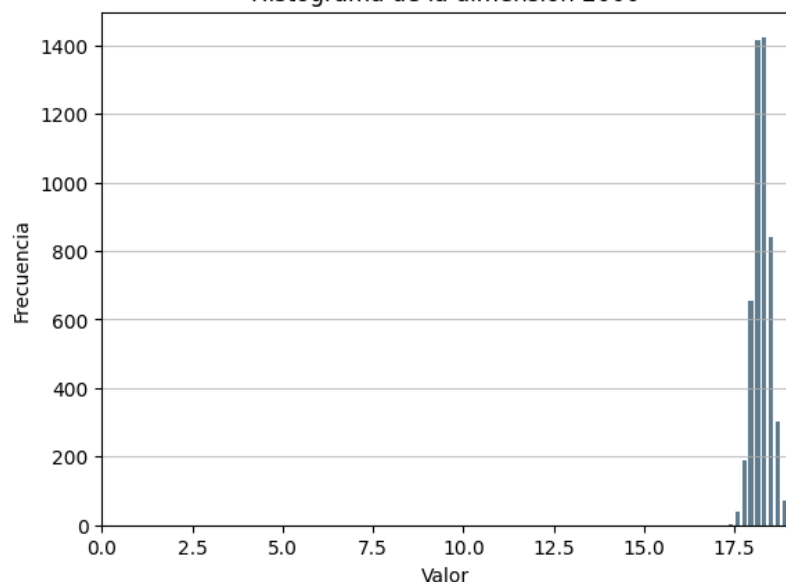
Datos de representación 500D : cantidad de bins = 10, rwidth=0.9, limitación de rango de 0.00 : 10.00.

Histograma de la dimensión 1000



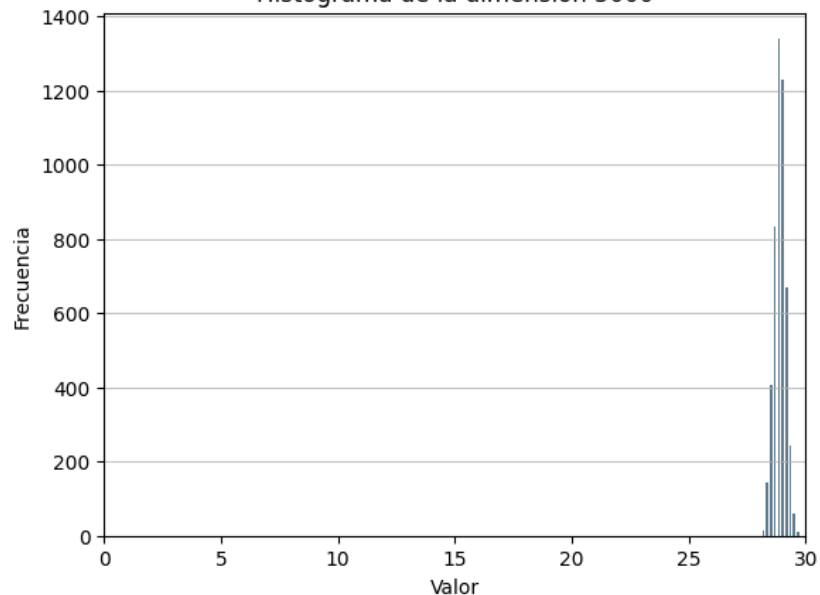
Datos de representación 1000D :
cantidad de bins = 10, rwidth=0.8,
limitación de rango de 0.00 : 14.00.

Histograma de la dimensión 2000



Datos de representación 2000D :
cantidad de bins = 10, rwidth=0.7,
limitación de rango de 0.00 : 19.00.

Histograma de la dimensión 5000



Datos de representación 5000D :
cantidad de bins = 10, rwidth=0.6,
limitación de rango de 0.00 : 30.00.

Análisis y conclusión


Podemos apreciar cómo a medida que aumentamos la dimensionalidad de los datos, la distancia entre estos crece, esto se debe a que crece exponencialmente con cada dimensión que sube, el espacio literalmente se infla, esto se ve representado por las mismas graficas, y vemos como simples puntos entre 1 y 0 generan distancias enormes, ocasionando una mayor dispersión de los datos, un notable sesgo a la izquierda porque las distancias solo suben si la dimensionalidad sube, incluso aplicando la distribución uniforme como hicimos en el código principal (contenido del github).

En conclusión, el aumento de la dimensionalidad de los datos, incluso si estos están distribuidos uniformemente, lleva a la dispersión más amplia de los mismos puesto a que el hiperespacio utilizado crece de forma exponencial conforme aumenta su dimensión, con este experimento sencillo podemos notar que para poder trabajar con datos de tantas dimensiones necesitamos de técnicas especiales o modos de pensar alternativos para poder reducir esta inmensa carga sobre el factor de la maldición de la dimensionalidad.

Herramientas para el experimento

Visual Studio 2022(C++)

Google colab(Python) :

 `Histogramas EDA.ipynb`

Colab adjunto en caso de que se necesite revisar o verificar(archivos y ejemplos en github)

Github (Repositorio)

https://github.com/CristianRamosMedina/Histogramas_EDA.git