**Introducción**

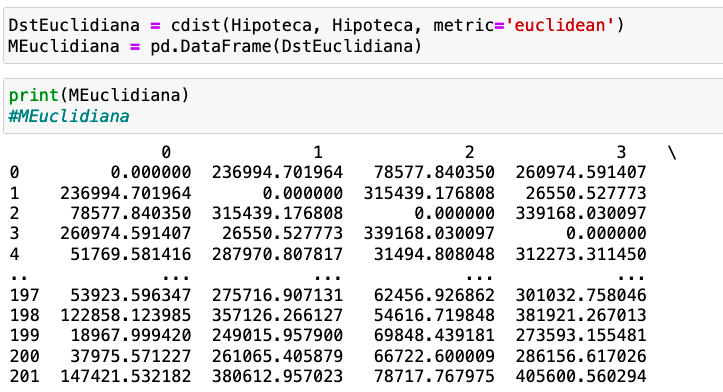
En esta práctica se utilizará el módulo cdist para calcular los diferentes tipos de métricas de distancia utilizados en el Machine Learning. Para cada una de las métricas de distancia se calculará la matriz distancia entre toda la población, un sector de la población y la distancia entre dos objetos. Se compararán dichas distancias para reforzar los conocimientos vistos en teoría sobre el cálculo que cada una de las métricas realiza.

**Objetivo**

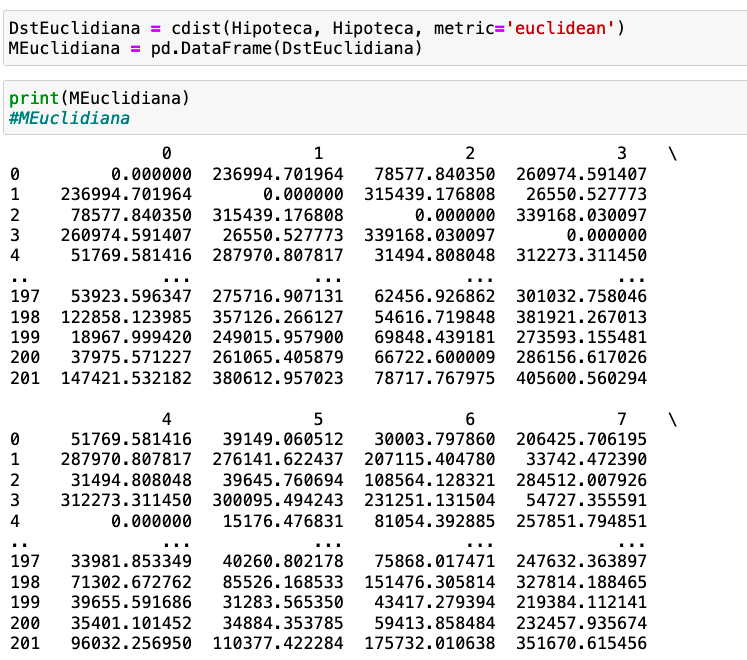
Obtener las matrices de distancia (Euclidiana, Chebyshev, Manhattan, Minkowski) a partir de una matriz de datos.

**Desarrollo**

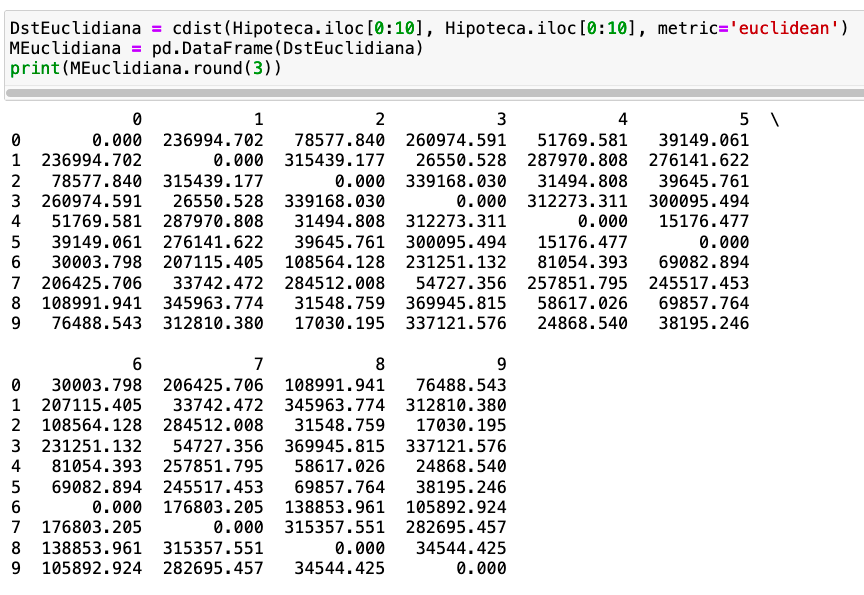
La primera matriz de distancia que se genera es Euclidiana y esto se genera con el módulo de cdist. De parámetros se pasan los datos y la opción de ‘euclidean’. A partir de estas distancias se genera un Data Frame. Dicha matriz es de dimensión de 202 columnas por 202 filas justo como se esperaba.



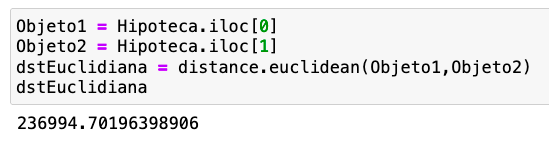
A partir de aquí se generó la distancia Euclidiana entre objetos. Se puede realizar el cálculo de toda la población, de un segmento o de dos objetos. Primero, se realizó la distancia Euclidiana de toda la población y se imprimió el resultado. El resultado completo no se puede ver porque son 202 filas y 202 columnas. Esto genera la distancia que tienen todos los vectores entre sí.



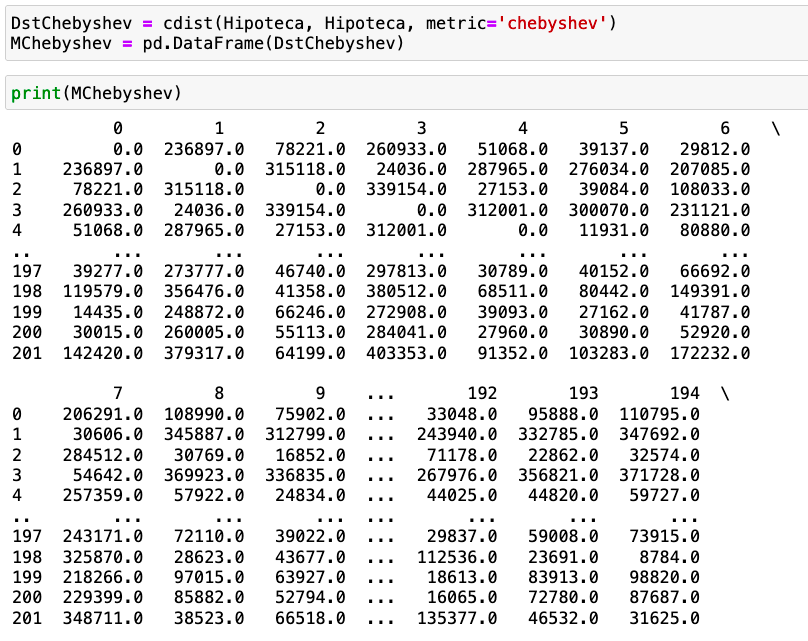
Después, se seleccionó una sola porción de los datos, 10 filas, para generar las distancias que existen entre estas filas. Para ello también se imprime una matriz y se utiliza cdist.



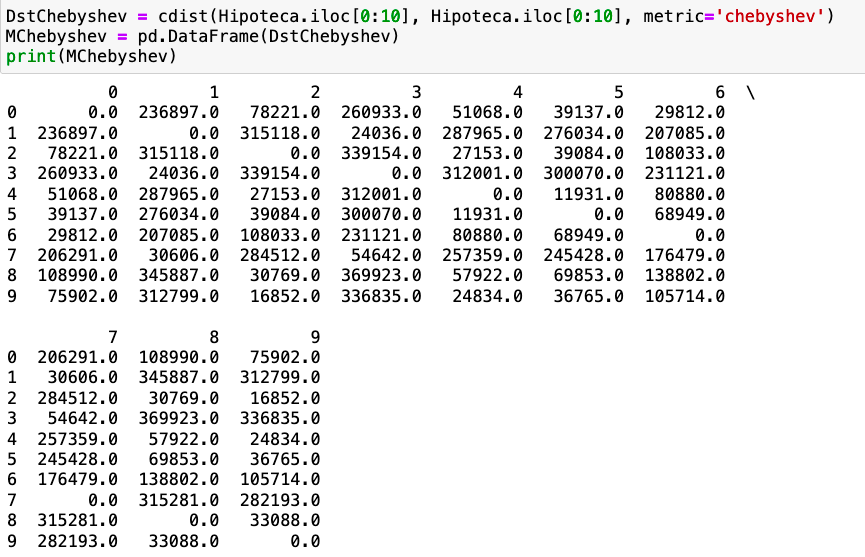
Por último, se obtuvo la distancia entre dos objetos utilizando cdist.



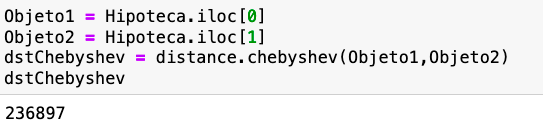
Para Chebyshev se tiene un proceso similar, solo cambia un parámetro en cdist para indicar que se quiere este tipo de distancia. Primero se obtiene la distancia de toda la población.



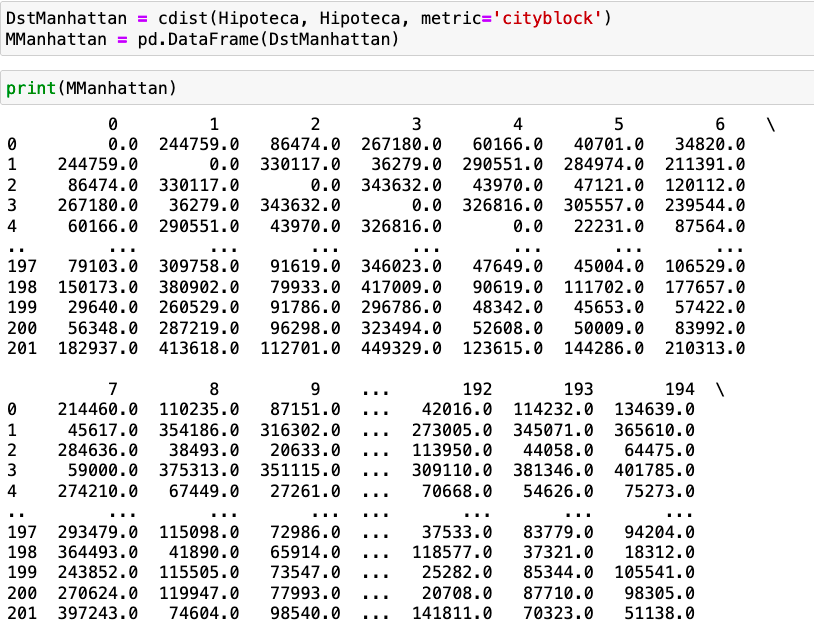
Después se obtuvo la distancia Chebyshev de diez filas de la población.



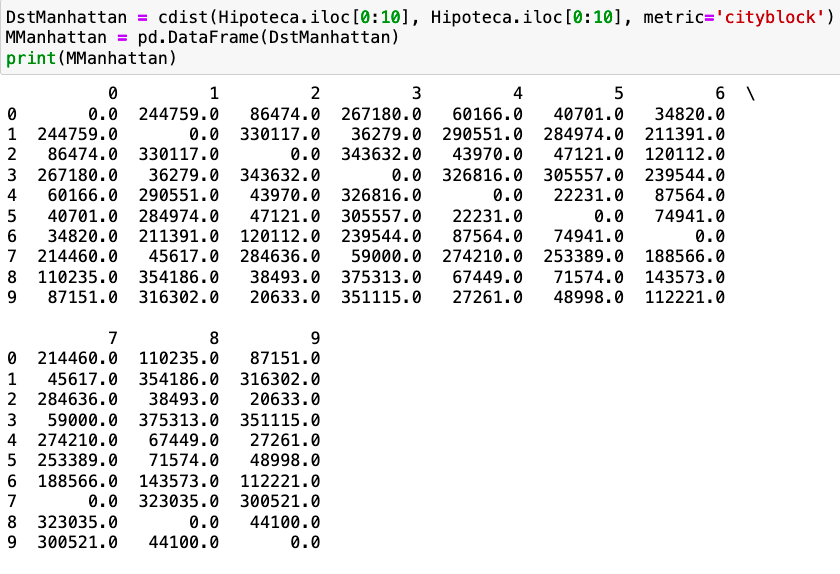
Por último se obtuvo la distancia Chebyshev entre dos objetos.



La siguiente distancia que se obtuvo de manera análoga a las enteriores fue la de Manhattan. Para esta distancia se define el parámetro de cdist como ‘cityblock’. Primero se obtuvo la distancia Manhattan entre toda la población de los datos.



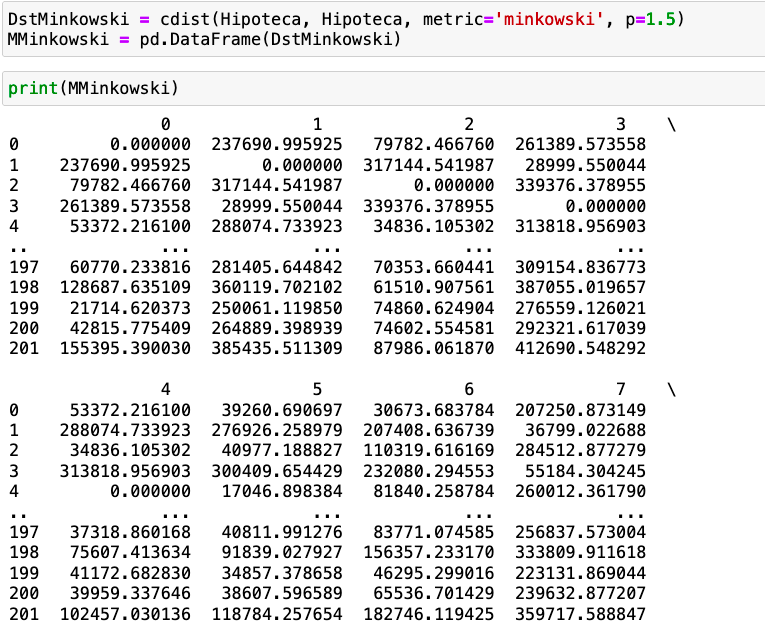
Se obtuvo la distancia Manhattan entre diez filas de los datos.



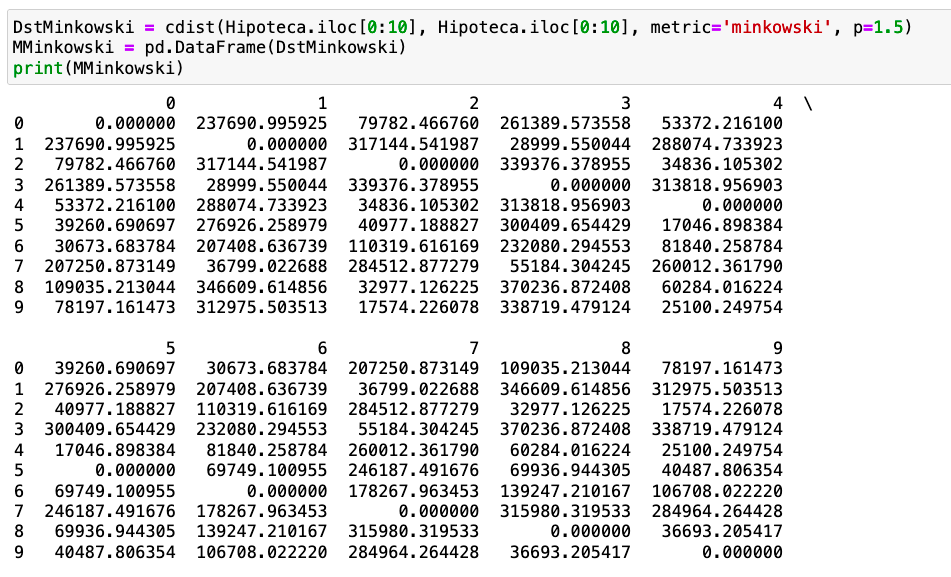
Por último se calculó la distancia Manhattan entre dos objetos.



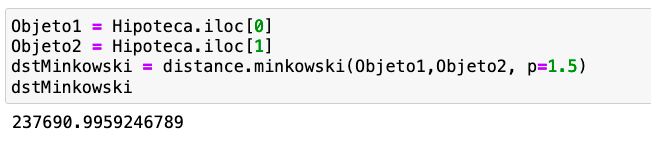
Para Minkowski se define un Lambda como parámetro de lo recto o curvo que debe ser la ruta para la distancia. Para esta práctica se define a Lambda como 1.5 en el algoritmo, este parámetro corresponde a p. Se realizan los mismos pasos que en el cálculo de distancias anteriores. Se calculó la distancia entre todos los datos de la población.



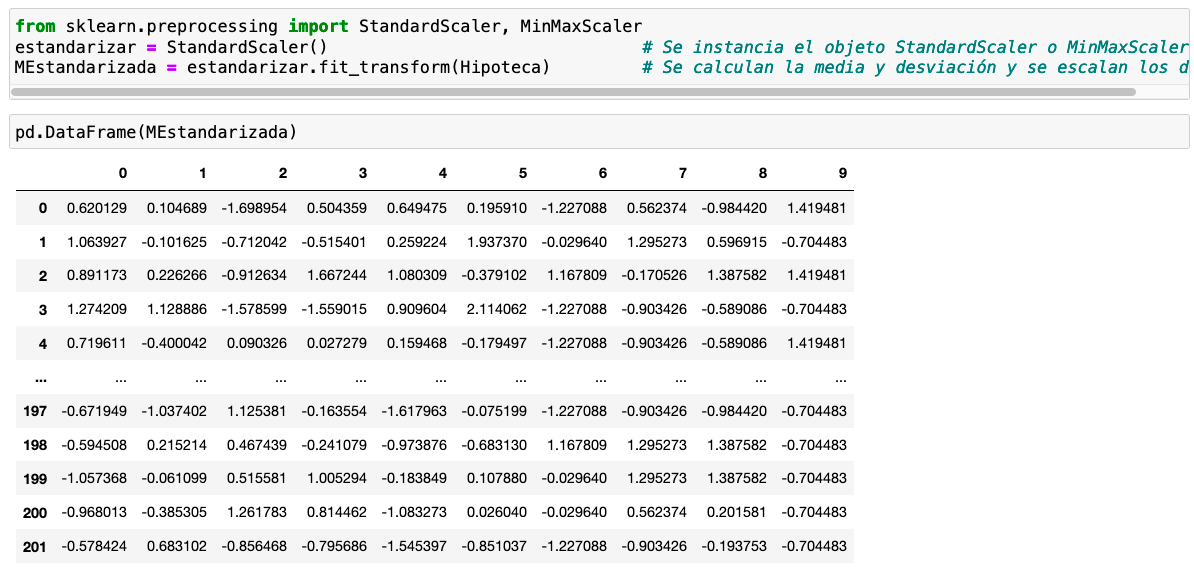
Se generó una matriz de distancias Minkowski de 10 filas de los datos.



Por último, se calculó la distancia Minkowski entre dos objetos.



Se generó una versión normalizada de esta práctica para tener facilidad para manejar los datos. La normalización consiste en estandarizar la escala que tienen los datos. De esta manera, cuando un algoritmo opera sobre dichos datos se tiene una mayor facilidad de las operaciones. Se le da a todas las variables el mismo peso para cuando se deba trabajar con un algoritmo que tome dos o más variables. Se normalizó utilizando la biblioteca de sklearn donde solo se pasan los datos a trabajar y se obtiene una matriz estandarizada.



Se compararon las cuatro métricas generadas entre la distancia de dos objetos.

|  |  |
| --- | --- |
| **Tipo de distancia** | **Distancia Calculada** |
| Euclidiana | 236994.70196398906 |
| Chebyshev | 236897 |
| Manhattan | 244759 |
| Minkowski | 237690.9959246789 |

La distancia de Manhattan es la mayor de todas y esto tiene mucho sentido ya que genera una distancia calculando el espacio como si fueran cuadras de una ciudad, no se genera una distancia en línea recta como la Euclidiana. Minkowski es mayor que la Euclidiana y menor que la de Manhattan porque es una combinación de ambas. Se genera una curva que hace que no sea directa como la Euclidiana pero tampoco se como una cuadrícula. La distancia de Chebyshev es ligeramente menor que la Euclidiana y esto se puede explicar gracias a que no se toman todos los valores de las coordenadas que se tienen. Se toma el valor máximo pero esto puede generar que la distancia sea menor o mayor.

**Conclusión**

En esta práctica se pudo visualizar la manera en que se calculan distancias en un conjunto de datos utilizando el módulo de cdist. También, se pudo visualizar la manera en que se pueden seleccionar datos para generar una matriz de distancias o cómo se puede calcular la distancia entre dos objetos. Esta práctica será de gran utilidad para futuros algoritmos porque se utilizan las métricas de distancia para posteriores algoritmos. Se observaron las variaciones que existen entre las distancias. Para la distancia Euclidiana se calcula la distancia directa hacia el otro objeto. Para la distancia Manhattan se calcula como si fuera la cuadra de una ciudad. Para Minkowski se genera una curva que es el punto medio entre las dos métricas anteriores. Para Chebyshev se seleccionan solo las coordenadas máximas por lo que su tamaño puede variar. Con esta práctica se generó un dominio sobre el cálculo de distancias con diferentes métricas en Python.