



04 DE NOVIEMBRE DE 2022

# PRÁCTICA 5

Métricas de distancia- Funciones de distancia-  
Reducción de dimensionalidad



## Objetivo de la práctica

Por Angel Damian Monroy Mendoza

No. de Cuenta: 316040707

Realizar una reducción de la dimensionalidad basada en ACD y, con base en esa selección, obtener una matriz de distancias mediante las cuatro funciones vistas en clase.



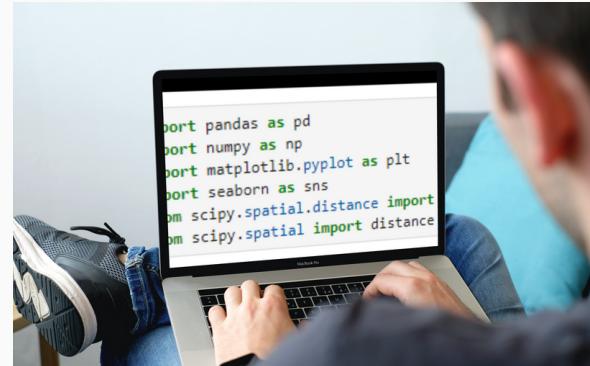
## Características:

- Por lo general, muchos algoritmos de Machine Learning utilizan medidas de distancia para identificar objetos o elementos.
- Estos elementos pueden tener características similares o no similares (disímiles).
- Así pues, estas medidas de distancia, conocidas también como "búsqueda de similitud vectorial" son (en parte) las responsables de lograr obtener información de nuestros datos.
- El siguiente conjunto de datos se obtuvo a partir de imágenes digitalizadas de pacientes con cáncer de mama de Wisconsin.
- Dentro de este conjunto contaremos con 12 características (1 de ellas es un ID del paciente) y 569 observaciones.

# Desarrollo

Para el desarrollo de esta práctica dividimos el proceso en seis secciones:

1. Importación de las bibliotecas necesarias.
2. Importación de los datos.
3. Procesamiento de datos.
4. Inspección visual.
5. Reducción de dimensiones
6. Matrices de distancia.



## 1. Bibliotecas

En este apartado se instalaron las bibliotecas necesarias para la manipulación de los datos, creación de vectores y matrices, generación de gráficas y, posteriormente, se instalaron las bibliotecas `scipy.spatial.distance` y `scipy.spatial` para las métricas de distancia.

## 2. Datos

Después, se cargaron los datos de forma tradicional en donde, en este caso, no se tuvo ningún problema como en la práctica anterior.

En este apartado nos dimos cuenta que contamos con dos variables categóricas, el ID del paciente y el Diagnóstico y las demás son variables numéricas. En este caso, eliminaremos más adelante la variable de ID y nos quedaremos con las características restantes.

IDNumber	Diagnosis	Radius	Texture	Perimeter	Area	Symmetry	Fractal Dimension
0 P-842302	M	17.99	10.38	122.88	1879.56	0.11	0.25
1 P-842517	M	20.57	17.77	132.98	2511.01	0.09	0.24
2 p-84300903	M	19.69	21.25	130.87	2575.47	0.07	0.23
3 p-84348301	M	11.42	20.38	77.31	1535.77	0.05	0.22
4 p-84358402	M	20.29	14.34	135.85	2054.19	0.04	0.21
...	...	...	...	...	...	...	...
564 P-926424	M	21.56	22.39	142.57	2638.70	0.03	0.20
565 P-926682	M	20.13	28.25	131.42	2329.03	0.02	0.19
566 P-926954	M	16.60	28.08	108.53	2054.55	0.01	0.18
567 P-927241	M	20.60	29.33	140.30	2773.74	0.00	0.17
568 P-92751	B	7.76	24.54	47.56	1062.00	-0.01	0.16

# Desarrollo

## 3. Procesamiento

En el procesamiento de los datos es importante conocer los tipos de datos que hay en tu DataSet, pues nos da una idea del procesamiento que se le tendrá que hacer para realizar los procedimientos que queremos.

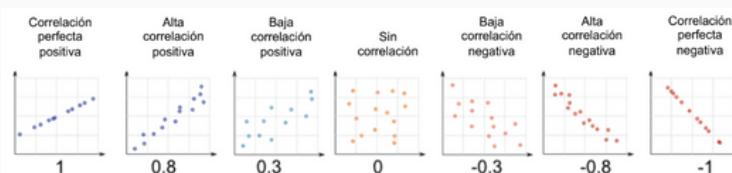
En este sentido, obtuvimos la información de nuestro conjunto de datos y corroboramos que sólo tenemos dos variables categóricas. Asimismo, realizamos una suma del método `isnull()` con la finalidad de saber si teníamos valores nulos dentro de nuestro DataSet, pues es necesario saber si contamos con éstos valores debido a que puede afectar los procesos que le haremos a nuestros datos.

Por último, con la función `get_dummies` convertimos la variable categórica del Diagnóstico en variable numérica, donde 1 corresponde a un diagnóstico de cáncer Maligno y 0 uno de cáncer Benigno y, posteriormente, se eliminó la columna de ID del paciente con la función `.drop`.

## 4. Inspección Visual y Reducción de Dimensiones

Como parte inicial es importante realizar una evaluación visual de los datos a través de gráficos de dispersión, por esta razón nos apoyamos de la biblioteca `matplotlib` para realizar graficas en donde los ejes X y Y representaran variables que nosotros creemos que están relacionadas.

La forma de saber si gráficamente tienen dependencia o no es observando si cumple algún comportamiento como el mostrado a continuación:



# Desarrollo

## 5. Reducción de Dimensiones

Ahora bien, una vez realizada la inspección visual, se llevó a cabo la matriz de correlación con la función `.corr()` en donde observamos el coeficiente de correlación que nos indica que tanto dependiente es una variable con otra, respecto al siguiente criterio:

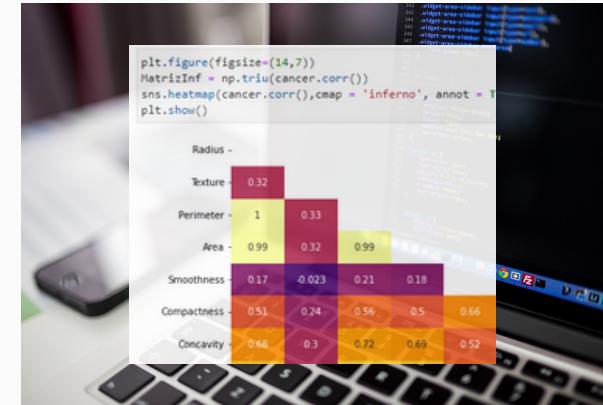
- **Coeficiente de correlación:**

- $|+1.00 \text{ a } +0.67| \rightarrow$  Fuerte o Alta
- $|+0.66 \text{ a } +0.34| \rightarrow$  Moderada o Media
- $|+0.33 \text{ a } +0.00| \rightarrow$  Débil o Baja

No obstante, para poder obtener una conclusión más visual se realizó un mapa de calor en donde se obtuvieron las siguientes conclusiones:

• Perímetro - Radio	(1.00)
• Área - Radio	(0.99)
• Concavidad - Radio	(0.68)
• Puntos Cóncavos - Radio	(0.82)
• Área - Perímetro	(0.99)
• Concavidad - Perímetro	(0.72)
• Puntos Cóncavos - Perímetro	(0.85)
• Concavidad - Área	(0.69)
• Puntos Cóncavos - Área	(0.82)
• Suavidad - Compactibilidad	(0.66)
• Concavidad - Compactibilidad	(0.88)
• Puntos Cóncavos - Compactibilidad	(0.85)
• Puntos Cóncavos - Concavidad	(0.92)

En seguida se decidió eliminar las características de: Perímetro, Área, Concavidad, Puntos Cóncavos y cabe recalcar que no se redujo la dimensión de la relación **Suavidad - Compactibilidad** debido a que el mapa de calor redondeo la puntuación que sacó, pues estrictamente cuenta con un valor de 0.659123.



```
state={ products: storeProducts
}
render() {
  return (
    <React.Fragment>
      <div className="py-5">
        <div className="container">
          <Title name="our" title="products: storeProducts" />
          <div className="row">
            {value) =>
              <ProductConsumer>
                <div>
                  <h2>Product Name</h2>
                  <p>Product Description</p>
                  <img alt="Placeholder image" />
                  <button>Buy Now</button>
                </div>
              </ProductConsumer>
            }
          </div>
        </div>
      </div>
    </React.Fragment>
  )
}
```

# Desarrollo



```
cancer = pd.DataFrame(MEstandarizada)
cancer

      0      1      2      3      4
0  1.097064 -2.073335  1.568466  3.283515  2.217515  2.255
1  1.829821 -0.353632 -0.826962 -0.487072  0.001392 -0.868
2  1.579888  0.456187  0.942210  1.052926  0.939685 -0.398
3 -0.768909  0.253732  3.283553  3.402909  2.867383  4.910

print(MEuclidiana)
#MEuclidiana
```

	0	1	2	3	4
0	1.097064	-2.073335	1.568466	3.283515	2.217515
1	1.829821	-0.353632	-0.826962	-0.487072	0.001392
2	1.579888	0.456187	0.942210	1.052926	0.939685
3	-0.768909	0.253732	3.283553	3.402909	2.867383

	0	1	2	3
0	0.000000	6.174365	4.546932	4.396037
1	6.174365	0.000000	2.705871	8.987098
2	4.546932	2.705871	0.000000	6.961528
3	4.396037	8.987098	6.961528	0.000000
4	4.834014	1.736955	2.060248	7.985967
...	...	...	...	...

## 6. Métricas de distancia

Por último, en los algoritmos basados en distancias es fundamental escalar o normalizar los datos para que cada una de las variables contribuyan por igual en el análisis de los datos, por lo que se procedió a estandarizar los datos y obtener las matrices de distancia Euclíadiana, de Chebyshev, de Manhattan y de Minkowski con una lambda igual a 1.5

# Conclusiones

De esta práctica podemos concluir que es una parte importante el tratado de los datos antes de calcular métricas de distancia, pues al estar basados en cálculos de los propios datos, se vuelve indispensable normalizar o estandarizar los datos, así como cuidar que no hayan datos nulos o variables categóricas.

Asimismo, se apreció la importancia de una buena ingeniería de características, pues la elección de características relevantes para tu conjunto de datos es primordial para no caer en la *maldición de la dimensionalidad* y evitar el sesgo en tus datos o conclusiones erróneas.

Por último, obtuvimos una alta reducción de dimensiones, pues nos quedamos con seis características sin contar el diagnóstico, por lo que podemos decir que se logra explicar la información de los datos con seis características en lugar de once, lo cual tiene sentido porque las variables que se eliminaron estaban relacionadas al radio de la muestra del cáncer, por lo que se volvían dependientes de ella.

