T3	**Minera de datos** #5 Itzel Covarrubias Vélez #19 Daniel Rodríguez Rugerio	8B	E #

¿Qué es OCR?

Es una tecnología que trata de emular la capacidad del ojo humano para reconocer objetos. Concretamente es un software que permite el reconocimiento óptico de los caracteres contenidos en una imagen, de forma que estos se vuelven comprensibles o reconocibles para una computadora, obteniendo como resultado final un archivo en un formato de texto editable.

Proceso

- Adquisición de la imagen con texto
- Binarización de la imagen
- Fragmentación o segmentación de la imagen, para separar las letras
- Procesar los segmentos para realizar características
- Reconocimiento de las letras usando una técnica de clasificación

Requerimientos

- Equipo de computo
- Carpeta de Imágenes binarias
- Spyder

Conceptos

DataSet: Un DataSet representa un conjunto completo de datos, incluyendo las tablas que contienen, ordenan y restringen los datos, así como las relaciones entre las tablas.

Imagen Binaria: La binarización de una imagen consiste en un proceso de reducción de la información de la misma, en la que sólo persisten dos valores: verdadero y falso. En una imagen digital, estos valores, verdadero y falso, pueden representarse por los valores 0 y 1 o, más frecuentemente, por los colores negro (valor de gris 0) y blanco.

Knn de clasificación: Clasifica nuevas instancias como la clase mayoritaria de entre los k vecinos más cercanos de entre los datos de entrenamiento

Procedimiento:

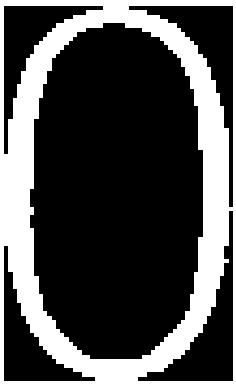
Generación de dataset

1. Leer las carpetas que contenían las imágenes segmentadas.
2. Por cada carpeta se lee cada una de las imágenes
3. Por cada imagen se obtienen las siguientes características, las cuales son almacenadas en un documento csv:

Característica 1

Columnas

Filas



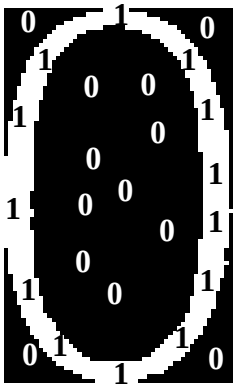
Razón = columnas/ filas

Obtener la razón de filas y columnas de la imagen.

Característica 2

Columnas

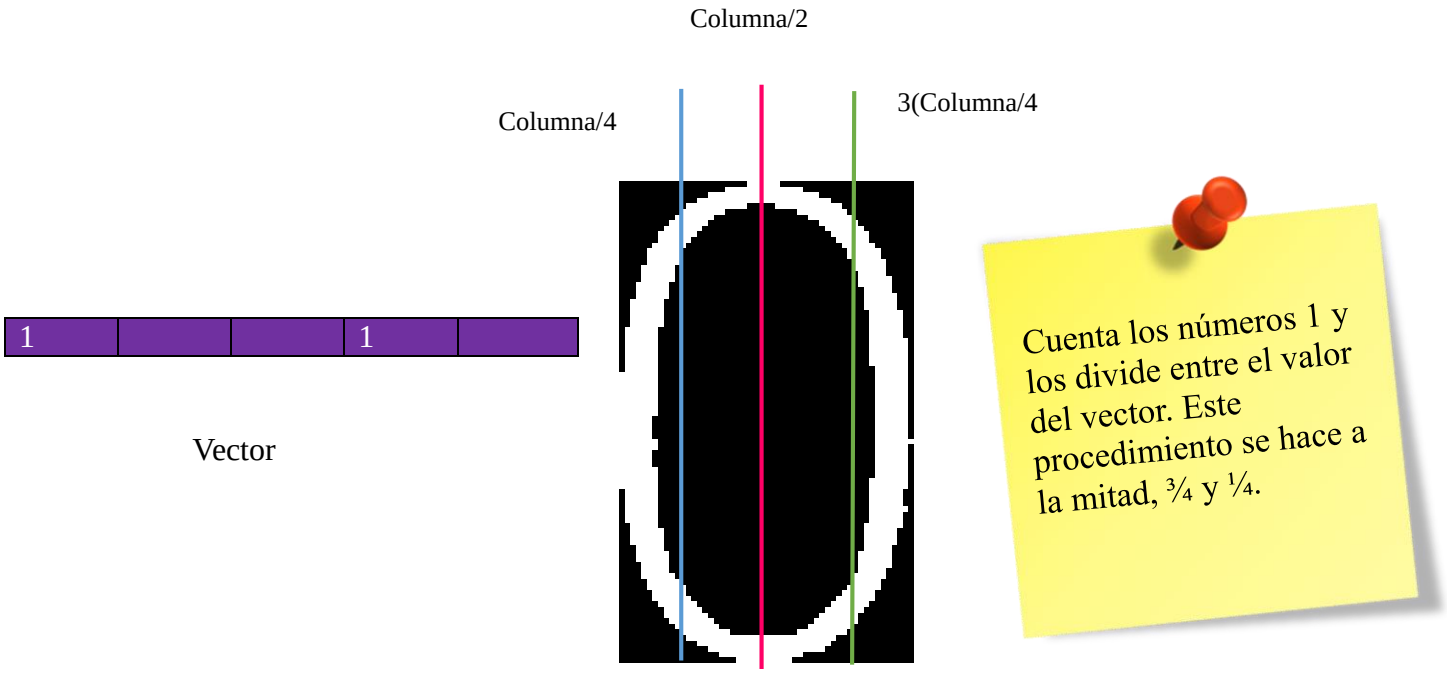
Filas



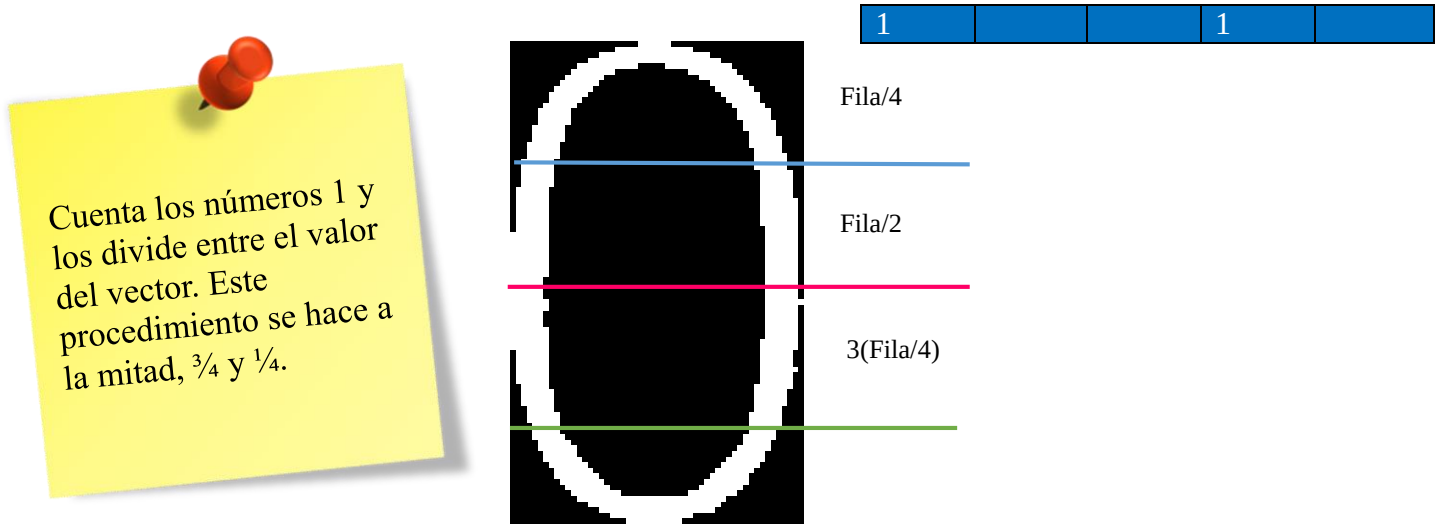
Se deben recorrer las filas y columnas, cuando se encuentra un uno, estos serán contabilizados.

En esta característica se saca el área de la imagen y se cuentan los unos que existen en la imagen.

Características 3,4 y 5



Características 6,7 y 8



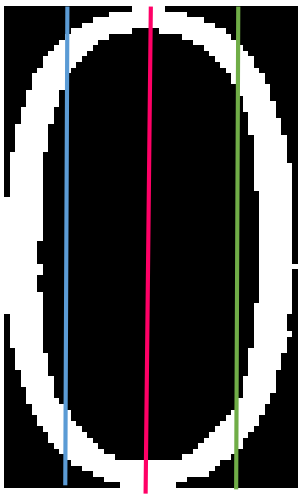
Características 9,10 y 11

1			1		1

Se cuentan los cambios del 1 a 0 en las columnas mencionadas, después divide los cambios entre 2 para determinar los cortes en este vector

Columnas

Filas



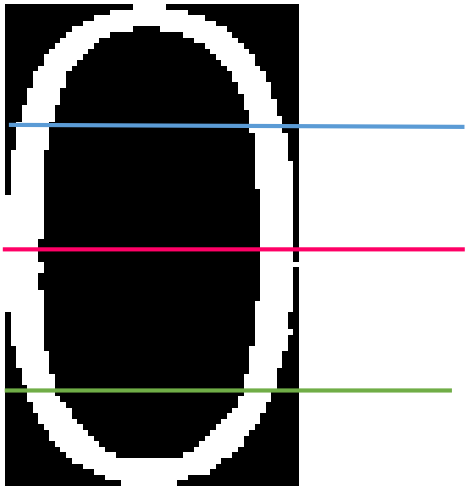
Características 12,13 y 14

	1		1			1

Se cuentan los cambios del 1 a 0 en las columnas mencionadas, después divide los cambios entre 2 para determinar los cortes en este vector

Columnas

Filas



Paso 4:

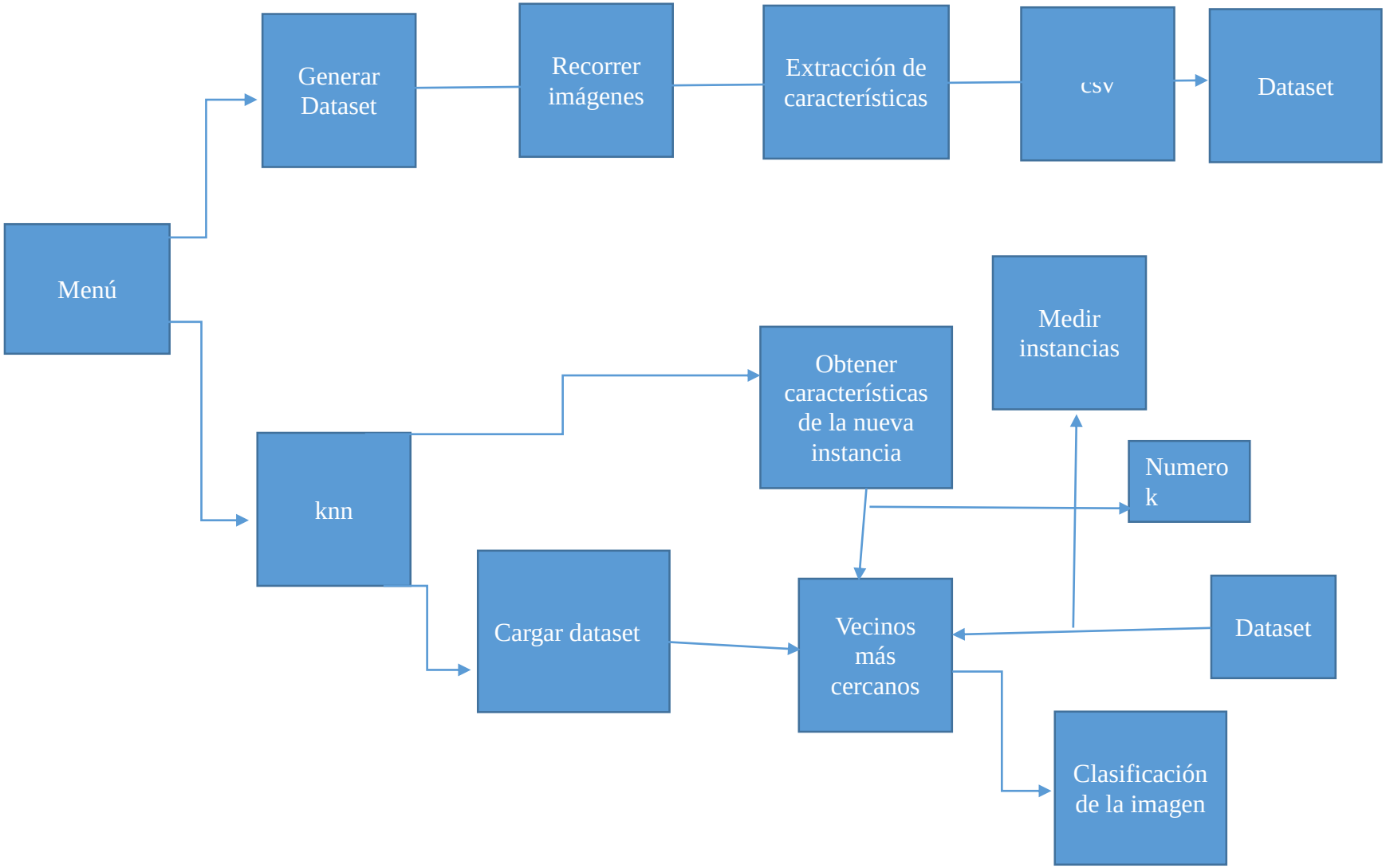
Una vez Generado dataset se aplicara el método de clasificación Knn.

No.Instancia	Filas/Colum	NumUnos	UnoCol1/2	UnoCol ¼	UnoCol3/4	UnosFila1/2	UnoFila1/4	UnoFila3/4	CorteFil1/2	CorteFil1/4	CorteFil3/4	CorteCol1/2	CorteCol1/4	CorteCol3/4	Clase
1	1.622642	0.252084	0.000219	0.000219	0.000219	0.000219	0.000219	0.000219	4	4	4	4	4	4	0
2	1.603774	0.247947	0.000222	0.000222	0.000222	0.000222	0.000222	0.000222	4	4	4	4	4	4	0
3	1.641509	0.245717	0.000217	0.000217	0.000217	0.000217	0.000217	0.000217	4	4	4	4	4	4	0
4	1.634615	0.245928	0.000226	0.000226	0.000226	0.000226	0.000226	0.000226	4	4	4	4	4	4	0
5	1.634615	0.245928	0.000226	0.000226	0.000226	0.000226	0.000226	0.000226	4	4	4	4	4	4	0
6	1.634615	0.249321	0.000226	0.000226	0.000226	0.000226	0.000226	0.000226	4	4	4	4	4	4	0
7	1.622642	0.251207	0.000219	0.000219	0.000219	0.000219	0.000219	0.000219	4	4	4	4	4	4	0
8	1.653846	0.254249	0.000224	0.000224	0.000224	0.000224	0.000224	0.000224	4	4	4	4	4	4	0
9	1.622642	0.25011	0.000219	0.000219	0.000219	0.000219	0.000219	0.000219	4	4	4	4	4	4	0
10	1.603774	0.246837	0.000222	0.000222	0.000222	0.000222	0.000222	0.000222	4	4	4	4	4	4	0
11	1.603774	0.250166	0.000222	0.000222	0.000222	0.000222	0.000222	0.000222	4	4	4	4	4	4	0
12	1.603774	0.245949	0.000222	0.000222	0.000222	0.000222	0.000222	0.000222	4	4	4	4	4	4	0
13	1.641509	0.242897	0.000217	0.000217	0.000217	0.000217	0.000217	0.000217	4	4	4	4	4	4	0

```
Menu de opciones
1 Cargar Archivos
2 Reconocer imagen
>>>>>>>>>  2
Ingrese la ruta de la imagen que desea reconocer
 3_001
Ingrese el valor de K
>>>>>  3
Numero de Intancias 14264
Numero de CLases 36
Numero de Caracteristicas 14

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>

Valor de K =3
No Instancia     distancias           Clase
 674         0.001316366827555926    3
 690         0.00153576129881524 7   3
 701         0.001316366827555953 7   3
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>
Realizar otra operaion
1 >>>> Si
2 >>>> No
>>>>>>
```


Menú
Generar Dataset
Recorrer imágenes
Extracción de características
csv
Dataset
knn
Obtener características de la nueva instancia
Medir instancias
Numero k
Cargar dataset
Vecinos más cercanos
Dataset
Clasificación de la imagen