Universidad Nacional de San Agustín

Unidad de Postgrado de la Facultad de Producción y Servicios

Maestría en Ciencias de la Computación

TRABAJO FINAL: KNN para la Clasificación y Determinación de Diabetes

Docente:
Ph.D. Vicente Enrique Machaca Arceda

Grupo 10: Christian Néstor Barriga Marcapura
Weimar Ccapatinta Huamani
Roger Gutierrez Espinoza

2022

1 Introducción

Los conjuntos de datos consisten en varias variables predictoras médicas (independientes) y una variable objetivo (dependiente), resultado.
Las variables independientes incluyen el número de embarazos que ha tenido la paciente, su IMC, nivel de insulina, edad, entre otros

2 Desarrollo del Proyecto

Se envía el siguiente enlace en Github, donde se encuentran los codigos elaborados y con las animaciones respectivas.
`https://github.com/weicap/MCC_TrabajoFinal`.
Para el desarrollo del proyecto, se ha realizado el uso de una base de datos en extensión CSV, para poder predecir la probabilidad de contraer Diabetes, utilizando el algoritmo KNN, en el cual teniendo las inputs de:

- Embarazos
- Glucosa en la sangre y presión
- Grosor de la piel
- Insulina
- IMC
- Resultado de la edad funcional

Realizaremos un entrenamiento de los datos, teniendo en cuenta que tenemos multiples entradas utilizaremos un cambio de dimensiones por PCA Se tiene un analisis de la data tal como muestra la siguiente figura.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figura 1 Descripción de la data - Elaboración propia

Asi tambien la gráfica de los inputs que permiten tener un analisis de la data si es o no adecuada antes de realizar el entrenamiento.

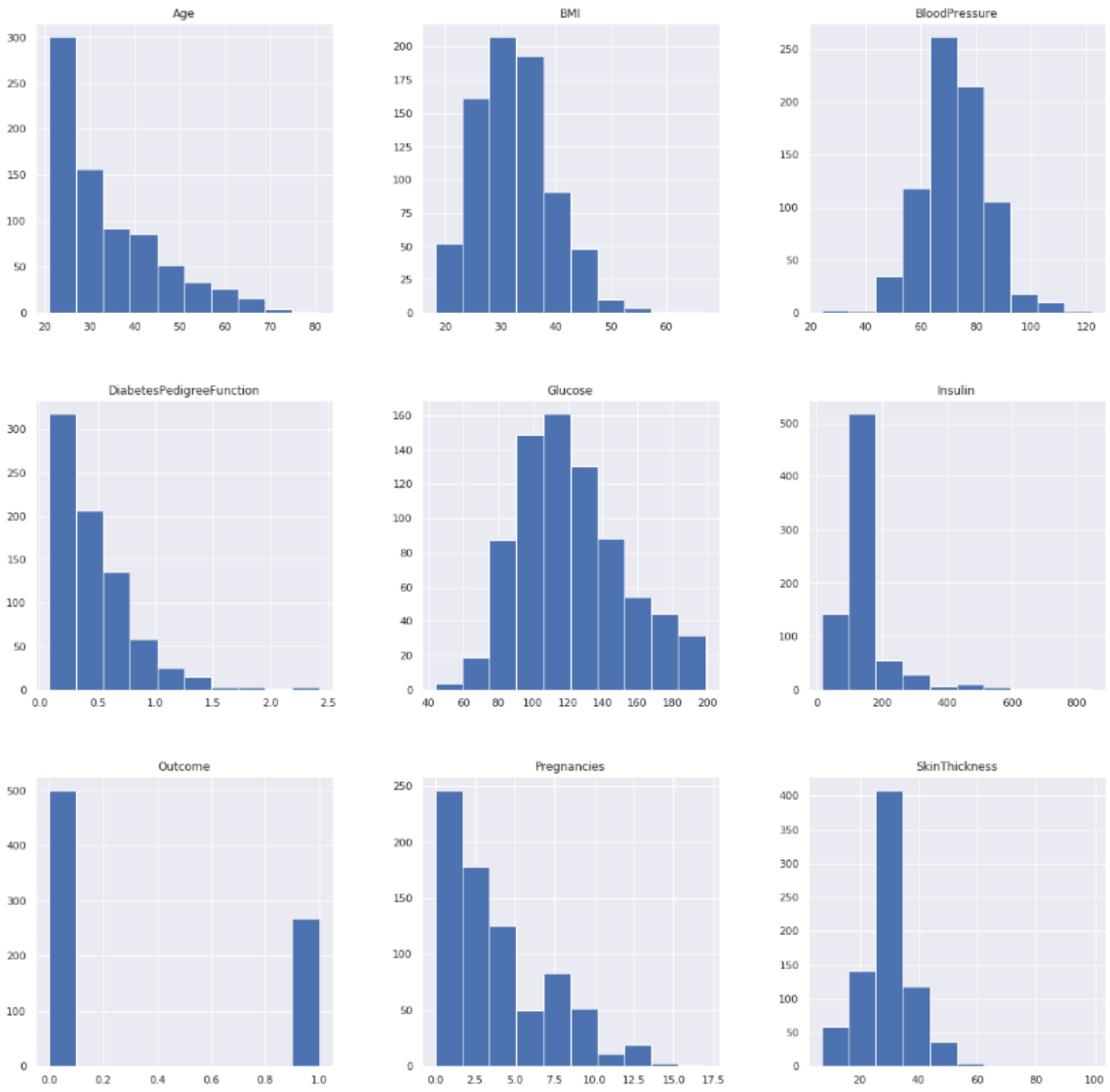

Figura 2 Datos de entrada - Elaboración propia

Luego de tener un analisis previo de la data debemos trabajar y reducir las dimensiones a dos utilizando PCA, de acuerdo a la base de datos todas las salidas estan asociadas a analisis previos en las que se determinan bajo estos factores si la persona tuvo o no diabetes, por lo que tambien se coloca etiquetas binarizando la salida a 1 y 0 para determinar si tuvo o no diabetes respectivamente, para luego poder cuantificar la cantidad de personas con diabetes tal como muestra la siguiente figura.

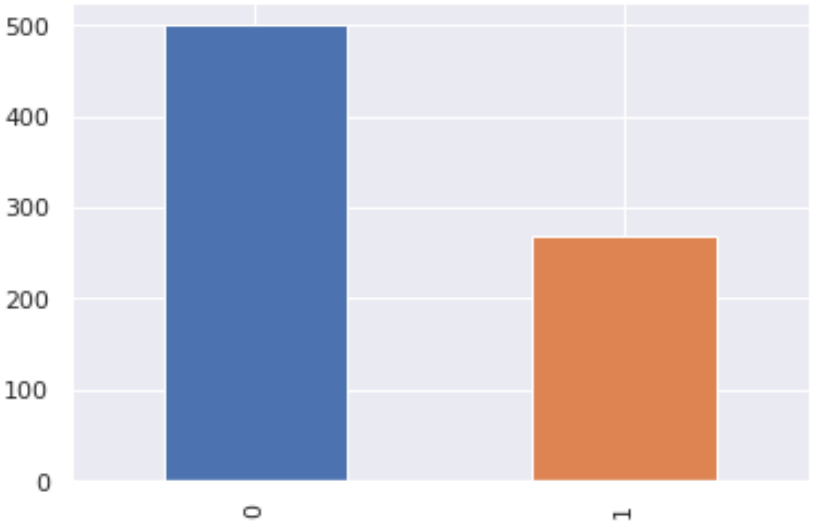

Figura 3 Cantidad de personas que tuvieron diabetes y no - Elaboración propia.

Cargamos al modelo al KNN y generamos nuestra grafica de dispersión en dos dimensiones, de tal manera que podemos clasificar los datos.

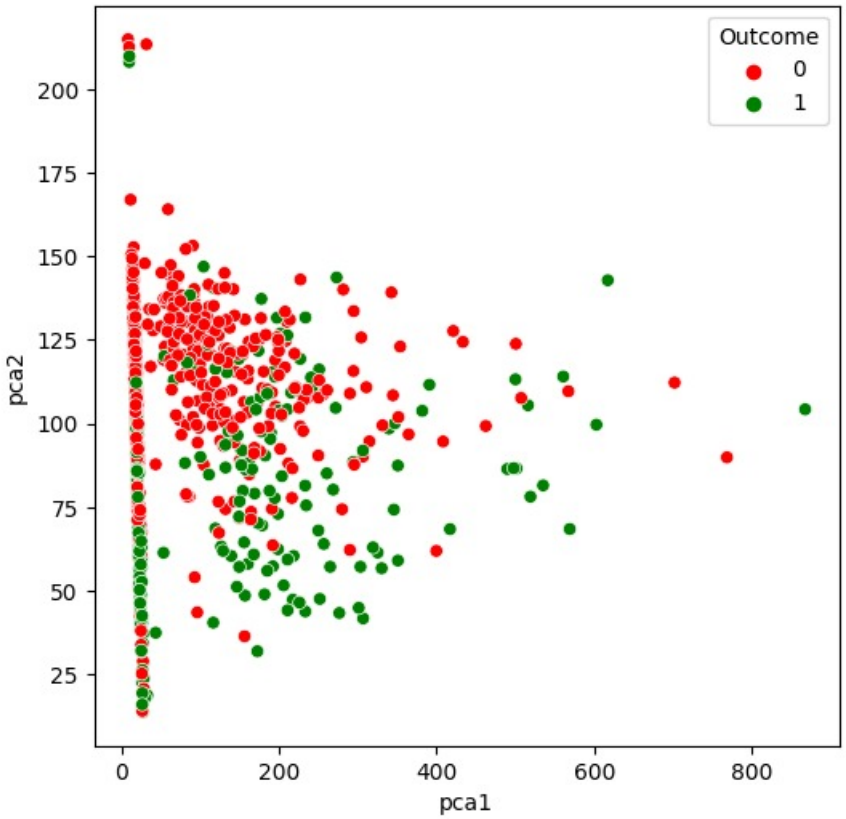

Figura 4 Gráfica de dispersión - Elaboración propia.

Realiza la busqueda de los mas vecinos cercanos, considerando entradas para poder predecir.

Maestría en Ciencias de la Computación

Grupo 10 ::> Trabajo Final

Predicción de la Diabetes

# de embarazos = 3
Glucosa = 84
Presion arterial = 78 (mm Hg)
Grosor de piel = 32 (mm)
Insulina sérica = 0 (mu U/ml)
masa corporal = 37.2 (peso en kg/(altura en m)^2)
Función de pedigrí de diabetes = 0.267
Edad = 28 (mu U/ml)

Predicción

El paciente es No Diabetico

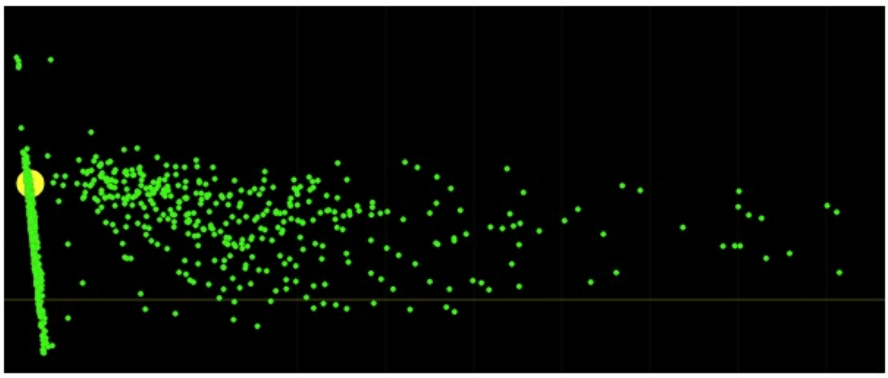

Figura 5 Predicción - Elaboración propia.

Se realiza la predicción de los puntos las cercanos y obteniendo la probabilidad de tener o no diabetes, asi tambien se puede observar la distancia entre los puntos desde la consola.

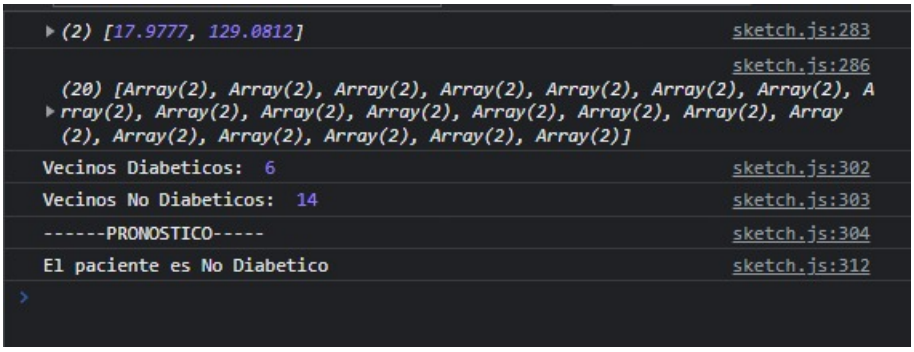

Figura 6 Consola - Elaboración propia.

3 Conclusiones

- A medida que se agregan nuevas muestras de entrenamiento, el algoritmo se ajusta para tener en cuenta cualquier dato nuevo, ya que todos los datos de entrenamiento se almacenan.
- KNN también es más propenso al sobreajuste. Si bien se aprovechan las técnicas de selección de características y reducción de dimensionalidad para evitar que esto ocurra, el valor de k también puede afectar el comportamiento del modelo.
- Utilizando este algoritmo, podemos tener una predicción de si puedes o no tener diabetes, sin embargo, depende mucho de la cantidad de datos que se testeen, mientras mas data mas precisa puede lograr a ser, pero tambien mas lento el algoritmo puede llegar a ser.