UNIVERSIDAD DE PIURA

CURSO: MACHINE LEARNING Y DEEP LEARNING CON PYTHON

PROFESOR: ING. PEDRO ROTTA

GRUPO: 12

INTEGRANTES:

JIMÉNEZ FARFÁN ALEJANDRA

MONDRAGON PARDO JACKSON

PRADO CHAFLOQUE SARAI ELIZABETH

RISCO ANTON, ANITA LISBETH

SIANCAS HUERTA MANUEL REYNALDO

VÁSQUEZ LAMADRID, MARÍA ANGÉLICA

PIURA, 22 DE ENERO DE 2022

Tabla de contenido

Índice de figuras

Índice de tablas

I. Introducción

Actualmente se encuentra en desarrollo la cuarta revolución industrial conocida como industria 4.0. Entre lo que impulsa esta revolución, se encuentra la inteligencia artificial y como parte de ella, el machine learning. La IA y el machine learning permiten al usuario sacar el máximo provecho a la información generada en diferentes áreas de la sociedad. Por ejemplo, en una empresa en donde las máquinas industriales tienden a averiarse, se pueden recopilar los datos necesarios que permitan prevenir y predecir futuras fallas en las máquinas a través de algoritmos de machine learning. El machine learning específicamente consiste en entrenar un sistema entregándole un conjunto de datos para que este halle por si solo una relación entre esos datos. Es decir, el machine learning emula el razonamiento humano.

El desarrollo de nuevas tecnologías en una revolución industrial implica que algunos trabajos quedan obsoletos, por lo que es necesario mantenerse al día con alguna de esas nuevas tecnologías. En este caso, la escogida para su estudio es machine learning. En el presente proyecto, se analizan 4 algoritmos de machine learning.

II. Data frame Marathon

1. Análisis del problema

La data usada es acerca del tiempo que se demoran varios corredores en culminar una maratón.

Todo Maratonista tiene un objetivo de tiempo en mente, y este es el resultado de todo el entrenamiento realizado en meses de ejercicios. Carreras largas, Zancadas, Kilómetros y ejercicio físico, todo añade mejora al resultado. La predicción de tiempos de maratón es un arte, generalmente guiada por expertos fisiólogos que prescriben los ejercicios semanales y los hitos del maratón.

Desafortunadamente, los corredores tienen muchas distracciones mientras preparan el maratón, el trabajo, la familia, las enfermedades, y por lo tanto, cada uno de nosotros llega al maratón con su propia historia.
El enfoque "simple" es mirar los datos después de la competencia, la tabla de clasificación.

Como parte del análisis de los datos, se pasa a detallar una descripción de cada columna de la data, siendo los siguientes.

Tabla 1

Columnas del data frame Marathon

Columna	Descripción
ID	Contador.
Marathon	Nombre de la Maraton realizada en Strava.
Name	Nombre del atleta.
Category	▪ El sexo y edad del participante donde MAM Atletas Masculinos menores de 40 años ▪ WAM Mujeres menores de 40 Años

	▪ M40 Deportistas Masculinos entre 40 y 45 años
km4week	El número total de kilómetros corridos en las últimas 4 semanas antes del maratón. Ejemplo: Si el km4week es 100, el atleta ha corrido 400 km en las cuatro semanas anteriores al maratón.
sp4week	Esta es la velocidad promedio del atleta en las últimas 4 semanas de entrenamiento.
cross training	Si el corredor también es ciclista, o triatleta, ayuda a ver si el atleta también es entrenador cruzado en otras disciplinas.
Wall21:	Para reconocer una buena actuación, como maratonista. Se puede decir que la puntación según como inicio y como termino su recorrido
Marathon time	Siendo el tiempo final que duro su recorrido.
Category	Este es un campo auxiliar.

Nota. Esta tabla muestra los nombres de las columnas del data frame Marathon y sus respectivas descripciones.

Figura 1

Dataset antes de ser modificada

	identificación	Maratón	Nombre	Categoría	km4semana	sp4week	entrenamiento cruzado	pared21	MaratónTiempo	CATEGORÍA
2	1	Praga17	blair morgan	mamá	132.80	14.43478261		1.16	2.37	A
3	2	Praga17	Roberto Heczko	mamá	68.60	13.6744186		1.23	2.59	A
4	3	Praga17	michon jerome	mamá	82.70	13.52043597		1.30	2.66	A
5	4	Praga17	daniel o lek	M45	137.50	12.25854383		1.32	2.68	A
6	5	Praga17	¿Luck? señor zek	mamá	84.60	13.94505495		1.36	2.74	A
7	6	Praga17	David Pecina	M40	42.20	13.61290323		1.32	2.78	A
8	7	Praga17	tomas drabek	M40	89.00	12.59433962		1.38	2.81	A
9	8	Praga17	Jan Rada	M45	106.00	12.69461078		1.41	2.84	A
10	9	Praga17	tomas drabek	mamá	70.00	13.7704918	ciclista 1h	1.38	2.83	A
11	10	Praga17	martin ?indel?	M45	84.20	13.36507937		1.35	2.86	A
12	11	Praga17	maksim remezau	mamá	93.50	13.2		1.42	2.87	A
13	12	Praga17	Jaroslaw Marchewka	M50	65.70	13.36271186		1.40	2.87	A
14	13	Praga17	Tomás ? Romper a	M45	53.50	14.07894737	ciclista 4h	1.37	2.88	A

Sin embargo, al manejar una data con varios datos innecesarios para crear los modelos, nos hemos visto en la iniciativa de eliminar algunas columnas y hacer algunos ajustes a los features, como rellenar los espacios que no tienen valores con ceros; por otro lado, también se cambiaron de variables cualitativas únicas a variables cuantitativas únicas para el caso de las variables "Category" y "CrossTraining".

Después de realizar estos cambios, tenemos que los features y target son los que se muestran a continuación.

Features:

- Category
- km4week
- sp4week
- CrossTraining
- Wall21

Target:

- MarathonTime

Para saber qué tipo de mecanismo de machine learning usar con la data, es sustancial en primer lugar, identificar si estamos trabajando con una data de clasificación o de regresión, como mi data tiene como target MarathonTime, que es una variable que guarda datos del tiempo, por consecuencia estos datos son flotantes, entonces mediante este razonamiento me doy cuenta que se debe usar mecanismos de regresión para ajustar el modelo y que de tal manera haga predicciones lo más acertadas posibles.

Figura 2

Dataset modificado

	Category	km4week	sp4week	CrossTraining	Wall21	MarathonTime
0	1	132.8	14.434783	0	1.16	2.37
1	1	68.6	13.674419	0	1.23	2.59
2	1	82.7	13.520436	0	1.30	2.66
3	2	137.5	12.258544	0	1.32	2.68
4	1	84.6	13.945055	0	1.36	2.74
...	...	...	...	...	...	...
82	5	50.0	10.830325	0	2.02	3.93
83	3	33.6	10.130653	2	1.94	3.93
84	3	55.4	11.043189	0	1.94	3.94
85	2	33.2	11.066667	0	2.05	3.95
86	3	17.9	10.848485	4	2.05	3.98

87 rows × 6 columns

Para hacer los modelos de regresión, he analizado los más viables para la data, usando así Random Forest, KNeighborsRegressor, Model Ridge, Model Lasso, y por último hacer un análisis de PCA; es decir un análisis de componentes principales.

Al momento de realizar el entrenamiento de la data, se presentaron algunos errores de sintaxis que se fueron corrigiendo, pero por otra parte también se notó que métrica usada para evaluar los modelos de regresión: r2_score nos salía en valores de subajuste como es el caso de los códigos de Ridge vs Lasso, que me dan valores entre [0.0-0.5] dando prueba de que hay subajuste, pero con los otros dos modelos sucede todo lo contrario, los datos obtenidos por r2_score de validación y entrenamiento me salen en los rangos correctos que son mayores a 0.5, a continuación se puede evidenciar lo comentado.

Figura 3

R2_score dentro del rango de un modelo correcto de ajuste.

Random Forest

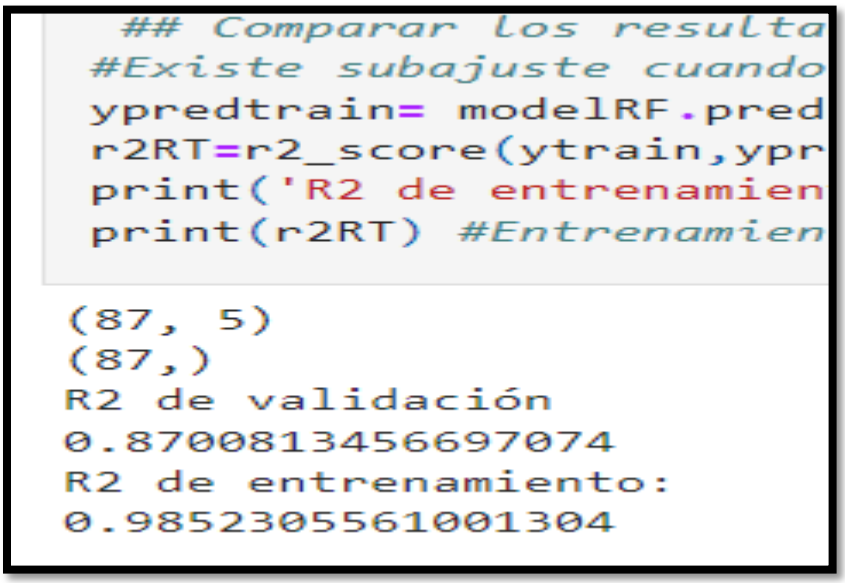


Figura 4

R2 score en el rango de un correcto ajuste. Modelo KNeighborsRegressor

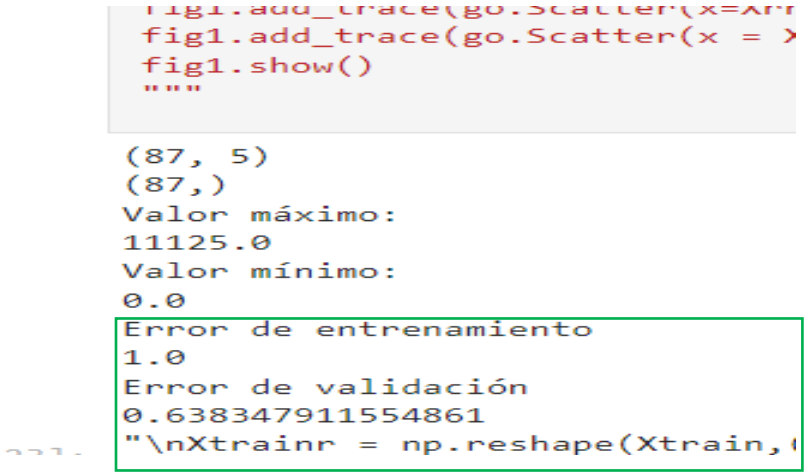


Figura 5

R2 score de Ridge y Lasso en el rango de [0.0-0.5], indicando que hay subajuste. Modelo Ridge

- Modelo Lasso

```
print(r2listpr)
```

```
(87, 5)
(87,)
Score Ridge:
[0.3959211046739298, 0.3355240735996491]
Score Lasso:
[0.3935041445146036, 0.3620268191492544]
```

2. Análisis estadístico

Figura 6

Categoría Vs. Los kilómetros recorridos en las últimas 4 semanas

En esta grafica se puede determinar que ha habido más atletas masculinos menores de 40 años (MAM) con actividad en las últimas 4 semanas, a comparación de atletas femeninas menores de 40 años. También se puede observar que son pocas las personas las que han recorrido más de 120 km por semana, equivalente a 480 km en las últimas 4 semanas.

Figura 7

Categoría Vs. Velocidad promedio del atleta en las últimas 4 semanas

Esta grafica no es tan descriptiva, ya que hay atletas que han tenido una velocidad muy parecida, por lo que hace que los puntos rojos sean cercanos. Lo que si se puede decir es que una persona masculina menor a 40 años ha tenido mayor velocidad que entre los demás concursantes, siendo esto muy buena predicción de que sea el próximo ganar ante una competencia.

Figura 8

Categoría Vs. si el atleta practica otro deporte como el ciclismo

Se puede observar que, de todos los atletas, si existe algunos que realizan ciclismo o al menos lo practican 1 hora. Siendo esto tal vez beneficioso para que su rendimiento mejore a momento de correr una maratón.

Figura 9

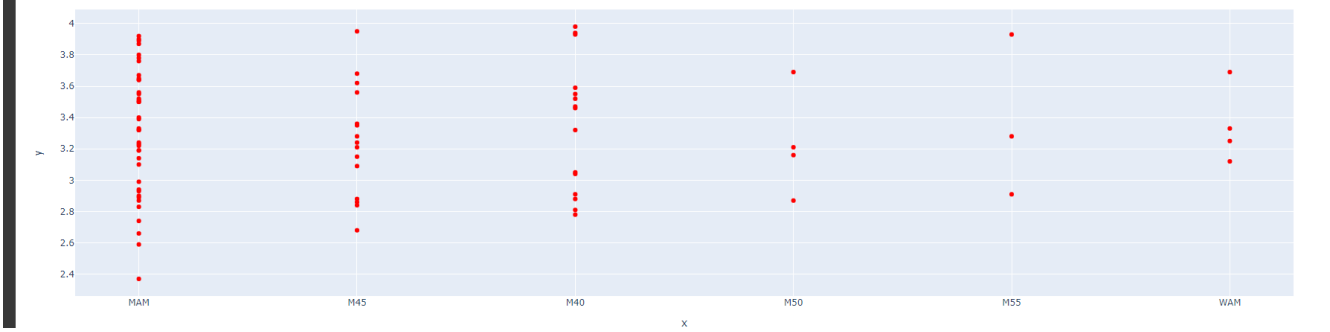
Categoría Vs. Puntuación como maratonista

Existe un participante con mejor participación siendo este un maratonista masculino de 50 años, diferente caso de la gráfica anterior de la velocidad, ya que fue uno menor de 40 años que obtuvo mejor velocidad. Esto quiere decir que la velocidad y como se desempeña el maratonista en la carrera no guarda relación. Siendo una característica que lo puede hacer ganar teniendo esta mejor puntuación.

Figura 10

Categoría Vs. tiempo de maratón

```
dataset = pandas.read_csv("https://raw.githubusercontent.com/Jacksonpardo/datos/main/MarathonData.csv")
dataset.head()
x_edad = list(dataset['Category'])
y_4km = list(dataset['MarathonTime'])
fig1 = px.scatter(x = x_edad, y= y_4km , color_discrete_sequence=['rgb(255,0,0)'])
fig1.show()
```

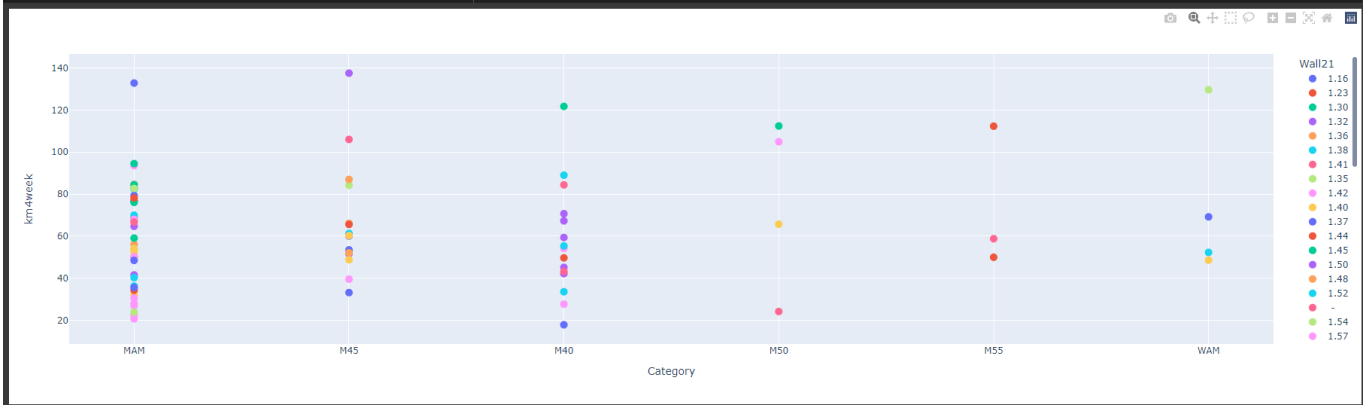


Desde aquí se puede ver que solo una persona alcanzo un tiempo de 2.4, siendo este un valor muy menor y que probablemente fue el resultado que lo puede hacer ganar la competencia. No se puede asegurar que es la misma persona que obtuvo la mejor puntacion en la grafica anterior

Figura 11

Categoria Vs. los kilómetros realizados en las ultimas 4 semanas

```
import plotly.express as px
fig1 = px.scatter(dataset , x = 'Category', y = 'km4week', color = 'Wall21')
fig1.update_traces(marker_size = 10)
```



Nota. La puntuación está representada por colores.

Aquí se puede decir que la persona con una puntuacion de 1.5 y con los mayores kilometros recorridos es un atleta masculino de 45 años de edad, esto dependera de su condicion fisica y de la experiencia que tiene en realizar este deporte.

Después de las observaciones que se han hecho a las graficas anteriores, se puede dar un resumen estadístico en donde se puede analizar de forma general los datos del dataset observando cuanto se puede recorrer un maratonista en 4 semanas y cuánto puede durar en una maratón.

Figura 12

Resumen estadístico del dataset Marathon

```
dataset.describe()
```

	id	km4week	sp4week	MarathonTime
count	87.000000	87.000000	87.000000	87.000000
mean	44.000000	62.347126	139.840706	3.319080
std	25.258662	26.956019	1191.427864	0.376923
min	1.000000	17.900000	8.031414	2.370000
25%	22.500000	44.200000	11.498168	3.045000
50%	44.000000	58.800000	12.163424	3.320000
75%	65.500000	77.500000	12.854036	3.605000
max	87.000000	137.500000	11125.000000	3.980000

3. Análisis del resultado

Luego de haber hecho el modelamiento, y de haber analizado el subajuste y/o sobreajuste de los modelos, se obtuvieron resultados esperados.

a. Modelo 1: Random Forest

Para el caso del Random Forest que me ofreció un R2_score muy aproximado a uno, tanto en validación como en entrenamiento; al ingresar nuevos datos, me dio una salida, que es la variable MarathonTime con valor igual a 2.7503, quedando en validez de que el modela sí funciona.

Figura 13

Resultado obtenido para el primer modelo Random Forest

```
Podrías intentar con 1-400-20-0-1.4
Ingresa valor para Category:1
Ingresa valor para km4week:400
Ingresa valor para sp4week:20
Ingresa valor para CrossTraining:0
Ingresa valor para Wall21:1.4
```

	Category	km4week	sp4week	CrossTraining	Wall21
0	1	400	20	0	1.4

```
La predicción es:
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning:

X has feature names, but RandomForestRegressor was fitted without feature names

array([2.7503])
```

b. Modelo 2: KNeighborsRegressor

Para el caso del KNeighborsRegressor que me ofreció un R2_score igual y aproximado a uno, tanto en validación como en entrenamiento respectivamente; cuando se ingresan nuevos datos, me dio una salida, que es la variable MarathonTime con valor igual a 2.68, que a su vez es muy próximo al valor obtenido con el modelo anterior que tiene mayor precisión aún, en consecuencia, se prueba la validez del correcto funcionamiento del modelo.

Figura 14

Resultado obtenido para el segundo modelo KNeighborsRegressor

```
Podrías intentar con 1-400-20-0-1.4
Ingresa valor para Category:1
Ingresa valor para km4week:400
Ingresa valor para sp4week:20
Ingresa valor para CrossTraining:0
Ingresa valor para Wall21:1.4
```

	Category	km4week	sp4week	CrossTraining	Wall21
0	1	400	20	0	1.4

```
La predicción es:
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning:

X has feature names, but KNeighborsRegressor was fitted without feature names

array([2.68])
```

c. Modelos 3: Comparación entre Modelo Ridge y Modelo Lasso

Para el caso del Modelo Ridge vs Modelo Lasso, que me ofreció un R2_score muy alejado a uno, tanto en validación como en entrenamiento respectivamente para cada modelo; cuando se ingresan nuevos datos, me dieron 2 salidas, que son la variable MarathonTime con valores iguales a 1.2897 y 1.0396 para Ridge y Lasso. Datos que al ser compararlos con las anteriores predicciones de los modelos no son tan precisos, y por tanto queda en evidencia que estos modelos no se han ajustado tan bien a la data, por ende, la precisión es variante respecto a otros modelos.

Figura 15

Resultado obtenido para la comparación Modelo Ridge – Modelo Laso

```
Podrías intentar con 1-400-20-0-1.4
Ingresa valor para Category:1
Ingresa valor para km4week:400
Ingresa valor para sp4week:20
Ingresa valor para CrossTraining:0
Ingresa valor para Wall21:1.4
```

	Category	km4week	sp4week	CrossTraining	Wall21
0	1	400	20	0	1.4

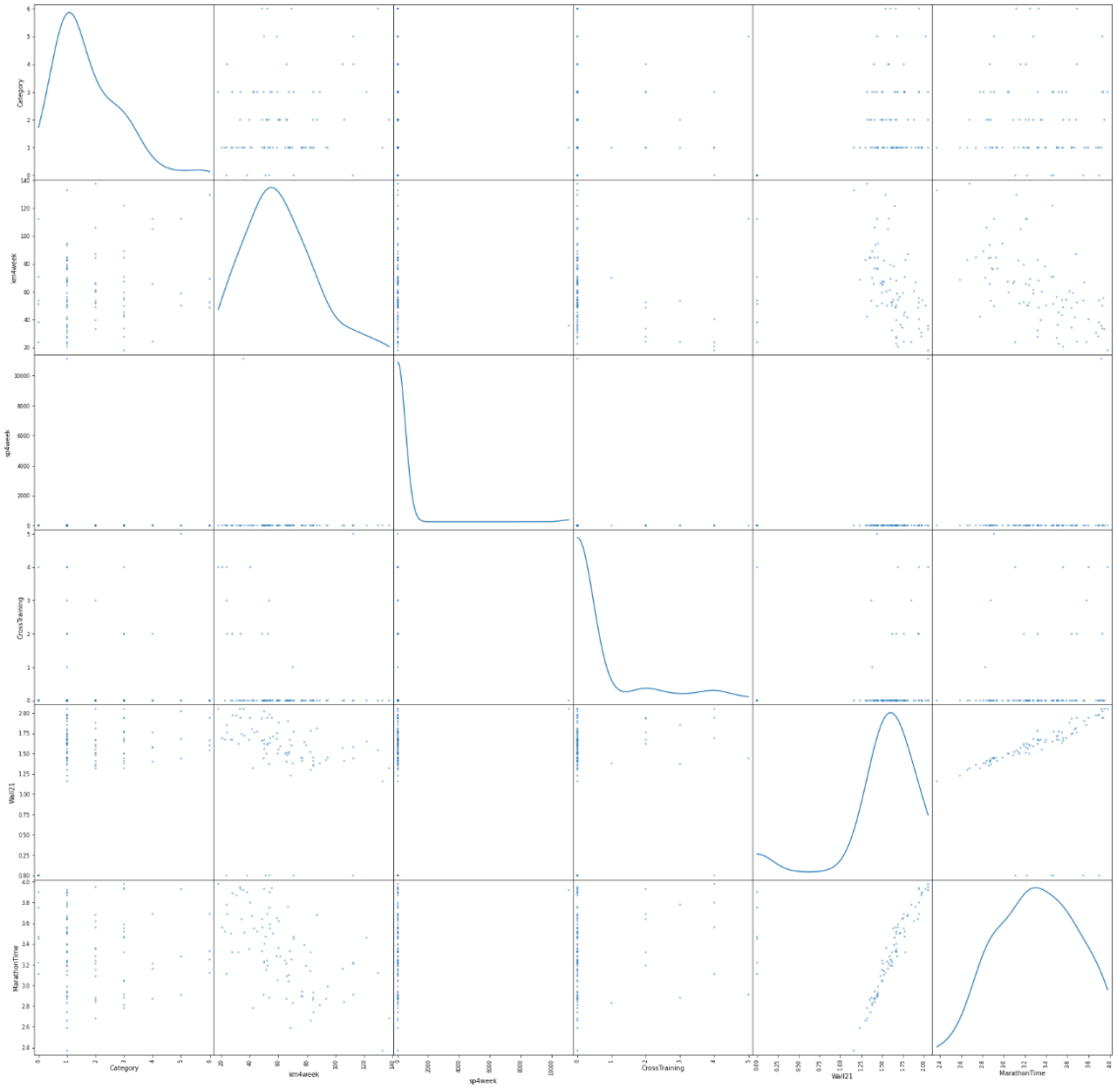
```
Las predicciones son:
[1.28974008]
[[1.03964247]]
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning:
```

d. Modelo 4: PCA

Cuando se hizo el análisis con PCA, se notó que los datos se escalaron y entonces tuvimos un súper gráfico donde me aparece la comparación de todas las variables contra todas las variables, en otras palabras, este algoritmo me permitió el análisis de componentes principales en otras palabras, se hizo un proceso que analiza los datos e intenta encontrar la estructura subyacente, la esencia de la información contenida en ellos. Esta estructura se define por las direcciones donde la varianza de los datos es mayor, es decir, donde hay una mayor dispersión de éstos. La forma más sencilla de comprender este concepto es mediante una visualización como en la siguiente ilustración.

Figura 16

Gráfica con escalamiento de modelo de PCA



III. Data frame Heart

1. Análisis del problema

En la actualidad la medicina se ha desarrollado notablemente en base a los avances científicos en los diferentes campos de investigación existentes que favorecen la creación de nuevos fármacos y tratamientos contra las enfermedades que afectan al ser humano. Por otro lado, si bien es cierto, gracias a la existencia de equipos electrónicos como por ejemplo el equipo de resonancia magnética, o el equipo de tomografía permiten obtener imágenes que favorecen al diagnóstico de una enfermedad, lo cierto es que sigue siendo el médico de turno quien finalmente

realiza dicho diagnóstico, y la probabilidad de éxito está directamente ligada a la experiencia y la formación de dicho profesional de la salud, de lo cual se deduce que la distribución de médico calificados para realizar un diagnóstico correcto de una determinada enfermedad: no es uniforme, lo cual representa un problema para las personas que requieren de un diagnóstico efectivo para poder tratar la patología que adolecen.

Una solución que se podría plantear ante dicha no uniformidad, es crear herramientas que emitan un diagnóstico en base a una información extensa y confiable de casos (lo cual emularía en cierta medida la experiencia de un médico), es decir una herramienta que realice una predicción o clasificación en base a una data set de calidad: Ante ello, salta a la vista que el Deep Learning encaja bien en esta definición y por lo tanto constituye una solución de vanguardia ante este problema (Expósito et al., 2008).

En este trabajo se ha decidido implementar un algoritmo de machine learning que permita determinar el diagnóstico de enfermedad al corazón en una base datos de pacientes UCI, y que se basa en cuatro modelos: k-nearest neighbors (KNN), Regresión Lineal, Ridge y Lasso. Finalmente se evaluará la eficiencia de esos modelos a través de una métrica apropiada: El coeficiente r^2 y el coeficiente de precisión según corresponda.

La data se trata sobre un estudio a pacientes que han estado en Unidades de Cuidado Intensivo consecuencia del covid, para lo cual han presentado una cardiopatía. En esta data se ha medido la intensidad de la enfermedad en niveles del 1 al 4 al haber presencia de la cardiopatía y como 0 al no presentar este.

La base de datos ha sido hecha por estudios en Cleveland utilizando únicamente a los investigadores ML. La creación de la base de datos se dio gracias las investigaciones de:

1. Instituto Húngaro de Cardiología. Budapest: Andras Janosi, MD

2. Hospital Universitario, Zúrich, Suiza: William Steinbrunn, MD

3. Hospital Universitario, Basilea, Suiza: Matthias Pfisterer, MD

4. VA Medical Center, Fundación de Long Beach y Cleveland Clinic: Robert Detrano, MD, Ph.D.

La data contiene 76 atributos con 14 subconjuntos descritos en la tabla anterior.

Como parte del análisis de los datos, se pasa a detallar una descripción de cada columna de la data, siendo los siguientes:

Tabla 2

Columnas del data frame Marathon

Columna	Descripción
age	Edad del paciente
sex	Genero del paciente definido como 1 = masculino, 0 = femenino
cp	Tipo de dolor de pecho definido en un nivel de intensidad de 0, 1, 2 y 3
trestbps	Presión arterial en reposo
chol	Medida de colesterol del paciente en mg/dl
fbs	Glucemia en ayunas
restecg	Resultados electrocardiográficos en reposo
thalach	Frecuencia cardiaca máxima alcanzada
exang	Angina inducida por el ejercicio
oldpeak	Depresión del ST inducida por el ejercicio en relación con el reposo
slope	La pendiente del segmento ST de ejercicio máximo
ca	Numero de vasos principales coloreados por fluorescencia (0 -3)
thal	Defectos (3= Normal, 6= Defecto fijo, 7= Defectos reversibles)
target	Objetivos

Nota. Esta tabla muestra los nombres de las columnas del data frame Heart y sus respectivas descripciones.

2. Análisis estadístico

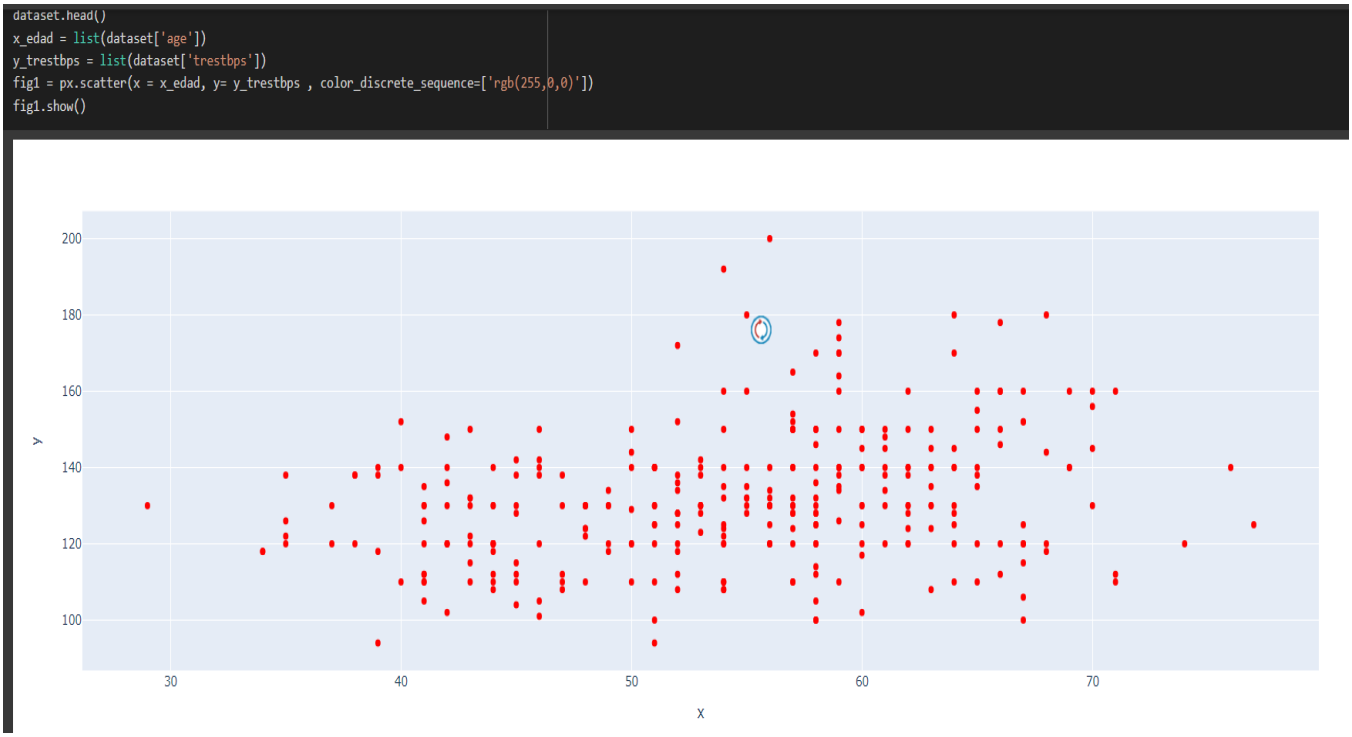
Para el análisis estadístico se ha planteado que una grafica distribuida que ayudara a observar en donde habrá más datos acumulados y en donde hay pocos datos acumulados, por lo que mas adelante se pueda realizar predicciones y cálculos estadístico a base de esta observación.

a. Edad del paciente Vs. Presión arterial en reposo

Teniendo en cuenta los valores normales de la presión sistólica de 120 mm hg, se traza una línea que limite este valor, para observar que paciente esta sufriendo una presión arterial anormal y elevado o no elevada. Se interpreta que pacientes fuera de este limite, ya están sufriendo una cardio patio ya que los valores de muestras que su presión arterial en reposo está siendo anormal, para lo que se acerca a un 90 % de estos pacientes.

Figura 17

Edad del paciente Vs. Presión arterial en reposo

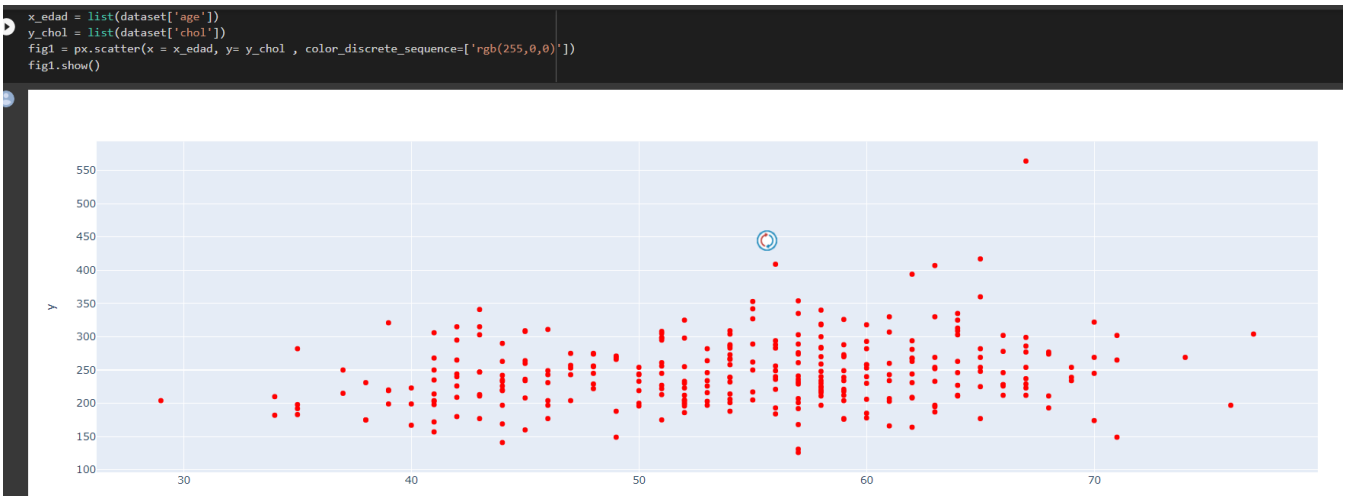


b. Edad del paciente Vs. Niveles de colesterol

De acuerdo con los valores normales de colesterol en la sangre en un adulto (menor a 200 mg/dl) observamos que cerca del 70% sufren de colesterol alto, para lo que significaría bastante que estos valores están dando a conocer ya una enfermedad cardiaca y en donde se debería empezar un tratamiento para evitar las consecuencias graves que esto involucra.

Figura 18

Edad del paciente Vs. Niveles de colesterol



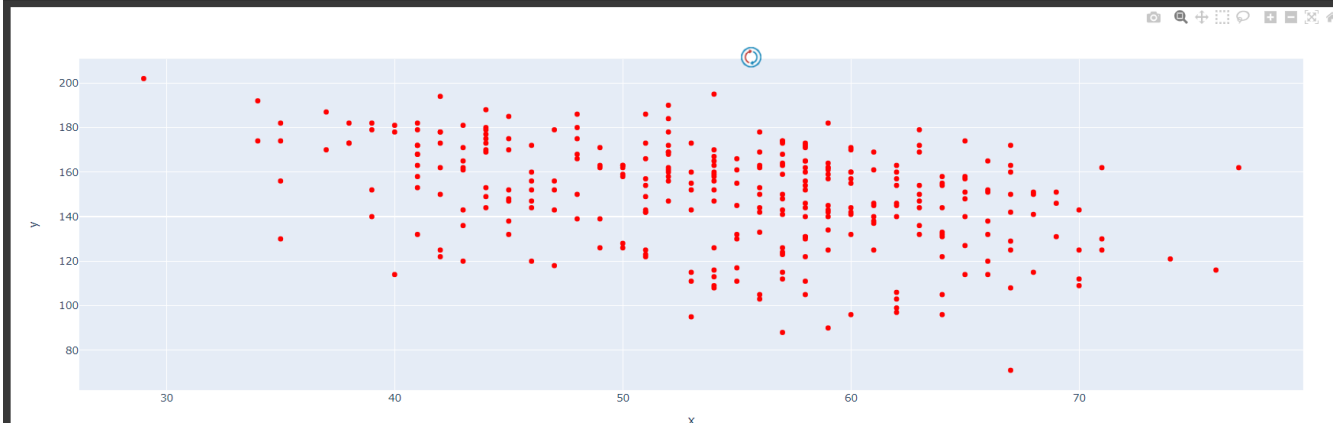
c. Edad del paciente Vs. Frecuencia cardiaca máxima alcanzada

Como se puede observar, si trazamos un limite entre 60 a 100 latidos por minutos, siendo estos valores normales de pulsación, se tiene cerca del 15 % que tienen los valores normales de pulsaciones. Por lo que demuestra que hay muchas más personas con problemas cardiovasculares.

Figura 19

Edad del paciente Vs. Frecuencia cardiaca máxima alcanzada

```
dataset = pandas.read_csv("https://raw.githubusercontent.com/manuelrey94/PYTHON-TRABAJO/ac9ff7a33b76166d39b6dfd6e4fc0c66289b3c3b/heart.csv")
dataset.head()
x_edad = list(dataset['age'])
y_thalach = list(dataset['thalach'])
fig1 = px.scatter(x = x_edad, y= y_thalach , color_discrete_sequence=['rgb(255,0,0)'])
fig1.show()
```



d. Edad del paciente Vs. Presión arterial en reposo, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo).

En esta gráfica se ha intentado detallar el nivel de dolor que puedenn sentir al sufrir una presión arterial elevada (mayor a 120 mm hg), para lo cual hay más acumulación de datos en el dolor de pecho de nivel cp=0. Esto significa que los pacientes que están sufriendo este desperfecto pueden no notar que están sufriendo de presión alta, haciendo esto aún más peligroso para quien lo sufra. En el nivel más alto de dolor, es decir en cp=3, se tiene un rango de valores en donde los hombres están sintiendo este nivel de dolor en pecho, más que la mayoría de la mujeres, ya que los datos de mujeres son pocos para este nivel de dolor.

Figura 20

Edad del paciente Vs. Presión arterial en reposo, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)

e. Edad del paciente Vs. Nivel de colesterol en la sangre, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)

Es prudente decir que al tener colesterol alto los dolores de pecho no vienen ha ser un síntoma notable antes de designar que se tiene un colesterol elevado, por la observación en el nivel de cp=0. Pero si se puede interpretar en este caso, que hay personas que están padeciendo de dolencia que den indicio de que se tiene este desperfecto, esto demostrado ya que hay una acumulación de datos en el rango de cp=3 y cp=2 en donde no se siente el dolor de pecho.

Figura 21

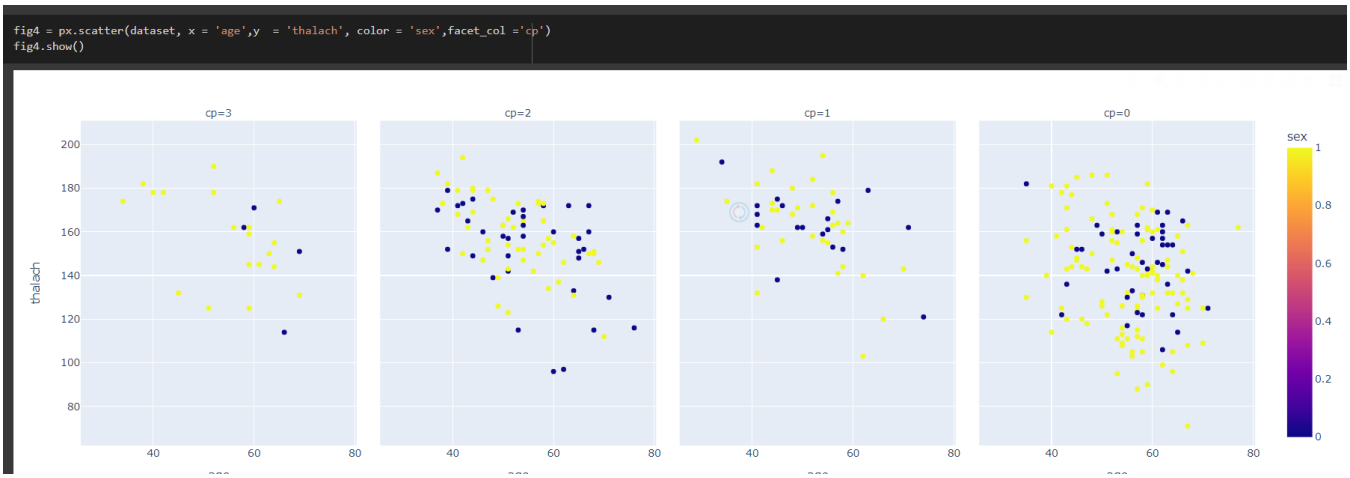
Edad del paciente Vs. Nivel de colesterol en la sangre, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)

f. Edad del paciente Vs. Frecuencia cardiaca máxima alcanzada, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)

Aquí si hay una notable variación de datos, entre el rango de cp = 3 y cp= 1 se tiene una gran cantidad de personas que están sufriendo dolor de pecho ante una frecuencia cardiaca elevada (mas de 100 pulsaciones/minuto), lo que demuestra que el dolor de pecho es un factor importante para sospecha de una patología. Se dice sospecha porque también en cp=0 no siente dolor a pesar de que su frecuencia cardíaca está siendo elevada.

Figura 22

Edad del paciente VS Frecuencia cardiaca máxima alcanzada, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)



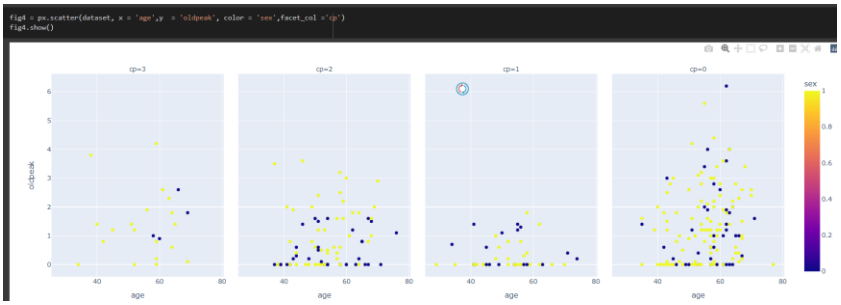
g. Edad del paciente Vs. Depresión del ST inducida por el ejercicio en relación con el reposo, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)

En esta grafia hay un dato mas clínico que podría predicir que personas están propensaa a sufrir un paro cardiaco, con los valores de depresión del ST inducida por el ejercicio. Con esto se puede decir que los valores son potentes predictores de enfermedad coronaria, siendo importante conocer: tiempo de comienzo, magnitud, extensión, duración, síntomas acompañantes (especialmente la aparición de dolor de pecho), ya que mejoran la utilidad diagnóstica de la prueba. Los valores normales de depresión del ST son menores a 1.5 y como se puede observar sobre ese

valor normal se encuentra en un aproximado del 50 % de pacientes que están propenso a sufrir un paro cardiaco, por lo que los datos de colesterol alto, frecuencia cardiaca y dolor de pecho observados anteriormente si están prediciendo ya una patología con consecuencias muy graves como la muerte. Esto siempre y cuando se evaluado por profesional de salud.

Figura 23

Edad del paciente VS Depresión del ST inducida por el ejercicio en relación con el reposo, asociado con el dolor de pecho, entre mujeres (azul) y hombres (amarillo)



Después de las observaciones que se han hecho a las graficas anteriores, se puede dar un resumen estadístico en donde se puede analizar de forma general los datos de la Dataset prediciendo una cardiopatía en los pacientes analizados.

Figura 24

Resumen estadístico del dataset Heart

```
dataset.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

3. Análisis de resultados

a. Modelo 1: KNN

Los parámetros del modelo fueron: n_neighbors=1.

Los resultados coeficientes r^2 obtenidos fueron:

0.5402730241616209
0.3850579299969694

De lo cual se infiere que el modelo tiene una precisión medianamente aceptable, dado que no existe una diferencia excesiva entre las predicciones hechas entre el set de validación y el set de entrenamiento. Aún así, tampoco se puede decir que el modelo tiene una muy buena precisión.

b. Modelo 2: Comparación entre modelo Lasso y Regresión lineal

Los parámetros del modelo fueron: Alpha=0.5.

Los coeficientes r^2 obtenidos fueron:

[0.8108569108839528, 0.8109030793591738]
[0.7923955906704134, 0.8014471104298373]

De los resultados obtenido se encuentra que este modelo presenta una mejor precisión e n base al modelo anterior.

c. Modelo 3: Comparación entre modelo Ridge y Regresión lineal.

Los parámetros del modelo fueron: Alpha=0.5.

Los coeficientes r^2 obtenidos fueron:

[0.809673632111269, 0.8109030793591738]
[-0.7670553187232201, -0.8014471104298373]

De los resultados obtenido se encuentra que este modelo presenta una mejor precisión e n base al modelo anterior. Siendo entonces más efectivo para este tipo de problemas.

d. Modelo 4: Clasificación binaria-Regresión logística

Los parámetros del modelo fueron: Número máximo de iteraciones = 200.

El coeficiente r^2 obtenido fue: 0.75.

Este modelo presentó un desempeño sustancialmente mejor que el primer modelo, pero inferior al desempeño de ridge y Lasso.

IV. Conclusiones

Los mecanismos usados en Machine Learning son de gran ayuda, ya que nos permiten la predicción de datos con bastante precisión si se logra ajustar de la manera correcta el modelo a analizar, en nuestro trabajo se analizaron 2 datas con diferentes tipos de clasificación, una de ellas fue la data de Maratón de tipo regresión y por tanto se debieron usar modelos de regresión para ajustar y generar modelos que predigan nuevos datos sin mucho error. Además, el aspecto iterativo del machine learning es importante porque a medida que los modelos son expuestos a nuevos datos, éstos pueden adaptarse de forma independiente. Aprenden de cálculos previos para producir decisiones y resultados confiables y repetibles.

Los modelos de ridge y lasso son los más adecuados para plantear una solución al problema planteado del diagnóstico de enfermedad al corazón en pacientes UCI.

La aplicación de la inteligencia artificial y machine learning al diagnóstico constituyen una opción muy interesante e importante que tiene la capacidad de cambiar el campo actual de la medicina.

El machine learning abarca gran parte de la industria 4.0, la cual se encuentra en desarrollo en la actualidad, por lo que es necesario que los jovenes aprendan a utilizar modelos de machine

learning para resolver problemas en sus distintas futuras áreas de trabajo, de forma que sus puestos no queden obsoletos y puedan abrirse paso en el mundo laboral.

A partir de lo trabajado, se puede decir que la programación con Python permite observar valores mediante análisis estadísticos y gráficas, con los cuales se puede estudiar una variable llevándolo a diferentes parámetros y sacar conclusiones de estos. Además que al compararlo con otros programas, este es otra forma de analizar datos, usando un correcto código y realizar los diferentes observaciones y dar con veracidad resultados que demuestran el estudio a realizar.

Python es muy útil para realizar graficas que nos permitan un mejor panorama con respecto a los datos a utilizar. Por ejemplo, en este trabajo hemos utilizado dos datas en donde una utilizamos un conjunto de datos sacados de un archivo .csv, y el mismo contiene información sobre los datos de tiempo de unos maratonistas en los cuales podemos sacar distintos gráficos como un historiograma o unos gráficos de dispersión en donde podemos ver las distintas correlaciones entre 2 variables.

V. Referencias

¿Qué es la Industria 4.0 y cómo funciona?. IBM. Recuperado de https://www.ibm.com/es-es/topics/industry-4-0.

Expósito Gallardo, María del Carmen, & Ávila Ávila, Rafael. (2008). Aplicaciones de la inteligencia artificial en la Medicina: perspectivas y problemas. *ACIMED*, *17*(5) Recuperado en 06 de febrero de 2022, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352008000500005&lng=es&tlng=es.

Rotta, P. (2022). Lecturas 1-9. En P. Rotta (Comp.), *Machine learning y deep learning con python*. Universidad de Piura.