UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PRACTICA 2: Modelo de Regresión.

APRENDIZAJE AUTOMATICO

CÉSAR ALEJANDRO HERNÁNDEZ OROZCO

Matricula: 1990010

GRUPO 003

MAESTRO: JOSÉ ANASTASIO HERNÁNDEZ SALDAÑA

INTRODUCCIÓN:

En esta tarea, se lleva a cabo un análisis exhaustivo de diferentes modelos de regresión para predecir una variable de interés en un conjunto de datos específico. El objetivo es identificar el modelo que mejor se adapta a los datos en términos de precisión y capacidad predictiva, utilizando para ello diversos enfoques y técnicas.

Este proyecto se realizo usando el conjunto de datos proporcinado por varios hospitales de distintas partes del mundo. Este es usado para tratar de encontrar si existe alguna enfermedad del corazon en los pacientes de estos hospitales.

En este conjunto de datos podemos encontrar bastante información de los distintos pacientes como lo son su edad, su sexo, si han tenido problemas cardiacos y como variable de respuesta se tiene si cuentan con un diagnostico de enfermedad del corazon.

En esta práctica se hizo uso de regresión. La regresión es una técnica fundamental en el análisis de datos que permite modelar y comprender las relaciones entre una variable dependiente y una o más variables independientes. En esta tarea, se seleccionan y comparan varios modelos de regresión para determinar cuál ofrece el mejor rendimiento en función de criterios específicos.

DESARROLLO:

Lo primero que hicimos para realizar esta practica fue empezar a conocer un poco mas como esta estructurado este conjunto de datos:

	name	role	type	demographic	description	units	missing_values
0	age	Feature	Integer	Age	None	years	no
1	sex	Feature	Categorical	Sex	None	None	no
2	cp	Feature	Categorical	None	None	None	no
3	trestbps	Feature	Integer	None	resting blood pressure (on admission to the ho...	mm Hg	no
4	chol	Feature	Integer	None	serum cholestoral	mg/dl	no
5	fbs	Feature	Categorical	None	fasting blood sugar > 120 mg/dl	None	no
6	restecg	Feature	Categorical	None	None	None	no
7	thalach	Feature	Integer	None	maximum heart rate achieved	None	no
8	exang	Feature	Categorical	None	exercise induced angina	None	no
9	oldpeak	Feature	Integer	None	ST depression induced by exercise relative to ...	None	no
10	slope	Feature	Categorical	None	None	None	no
11	ca	Feature	Integer	None	number of major vessels (0-3) colored by flour...	None	yes
12	thal	Feature	Categorical	None	None	None	yes
13	num	Target	Integer	None	diagnosis of heart disease	None	no

Podemos ver que este conjunto de datos esta bastante bien estructurado y nos presenta incluso como podemos trabajar con estos datos.

La variable con la trabajaremos principalmente es “num”. Esta variable toma el valor 1 si se diagnostica a esta persona con una enfermedad del corazon y y 0 si esta no tiene una enfermedad del corazon diagnosticada.

A continuación exploraremos un poco de los primeros datos que encontramos al ver el conjunto de datos:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	\
0	63	1	1	145	233	1	2	150	0	2.3	3	
1	67	1	4	160	286	0	2	108	1	1.5	2	
2	67	1	4	120	229	0	2	129	1	2.6	2	
3	37	1	3	130	250	0	0	187	0	3.5	3	
4	41	0	2	130	204	0	2	172	0	1.4	1	

	ca	thal
0	0.0	6.0
1	3.0	3.0
2	2.0	7.0
3	0.0	3.0
4	0.0	3.0

Como podemos ver la mayoria de los datos estan completos y cuentan con numeros como variables categoricas.

Para trabajar adecuadamente con este regresión nos apoyaremos en la variable “thalach” que hace referencia a el mayor pulso cardiaco que han tenido las personas

Durante el desarrollo de esta practica vimos que la gran mayoria de los datos estaban completos, aun asi se eliminaron 6 registros que estaban vacios.

Una vez limpiamos los datos vacios se dividio el conjunto de datos en una de entrenamiento y uno de pruebas dejando el 80% de los datos para entrenar y el 20% para probar. Los modelos con los que se trabajo fueron los siguientes:

- Regresión Lineal.
- Regresión Polinomial.
- Regresión KNN,
- Arbol de Decisión.
- Random Forest.

Usando las librerias que nos proporciona sklearn pudimos crear estos modelos rapidamente al compararlos unos con los otros obtenemos lo siguiente:

Modelo	MSE	R2
Regresión Lineal.	298.085	0.3336530
Regresión Polinomial.	458.654717	-0.020859
Regresión KNN	288.778	0.357246
Arbol de Decisión.	530.0000	-0.179657
Random Forest.	285.921640	0.363605

Criterios de Evaluación: Para seleccionar el mejor modelo, se consideran métricas de rendimiento como el R2 (coeficiente de determinación), RMSE (raíz del error cuadrático medio) y MSE (error cuadrático medio).

Al ver esta tabla podemos concluir facilmente que el mejor modelo de Regresión para predecir si hay posibilidad de que haya algun problema del corazón esta entre Random Forest y la regresión KNN.

Conclusión:

Tras aplicar los modelos y evaluar su desempeño utilizando las métricas mencionadas, se selecciona el modelo de Random Forest como el que mejor cumple con los requisitos de precisión y generalización. Este análisis no solo resalta el modelo más efectivo para el conjunto de datos en cuestión, sino que también proporciona insights valiosos sobre la naturaleza de los datos y las relaciones entre variables.

Ademas, esta practica nos ayudo mucho a crecer en nuestras habilidades como cientificos de datos. Pero aun asi los modelos contaron con poca precisión asi que es importante mejorar esto en las siguientes practicas si queremos mejorar nuestra habilidad de predecir cosas

Bibliografia:

UCI Machine Learning Repository. (2019). Uci.edu.

https://archive.ics.uci.edu/dataset/45/heart+disease