UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PRACTICA 3: Modelo de Clasificación.

APRENDIZAJE AUTOMATICO

CÉSAR ALEJANDRO HERNÁNDEZ OROZCO

Matricula: 1990010

GRUPO 003

MAESTRO: JOSÉ ANASTASIO HERNÁNDEZ SALDAÑA

INTRODUCCIÓN:

En esta tarea, se enfoca en la selección y comparación de diferentes modelos de clasificación para predecir una variable de interés en un conjunto de datos específico. El objetivo principal es identificar el modelo que mejor se desempeña en la predicción de clases, utilizando una serie de enfoques y técnicas para evaluar su rendimiento.

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 clasificación. La clasificación es una técnica crucial en el análisis de datos que permite asignar una etiqueta o clase a observaciones basadas en sus características. En esta tarea, se examinan y comparan varios modelos de clasificación para determinar cuál ofrece el mejor desempeño en función de diferentes criterios.

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.

Las demas variables con las cuales trabajaremos para obtener un buen agrupamiento son las siguientes:

- Las variables con las que trabajaremos seran sex, cp, fbs y exang.

  Sex =Sex (1=male,0=female)

  cp = Chest Pain (1=typical, 2=atypical,3=non-anginal pain, 4=asymptomatic)

  fbs= Fasting Blood Sugar (1=true,0=false)

  exang=exercise induced angina (1=yes,0=no)

Para mejorar el rendimiento de nuestro codigo decidimos eliminar los demas datos no relevantes y se aseguro que este interpretara todas estas variables como categoricas.

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 70% de los datos para entrenar y el 30% para probar. Los modelos con los que se trabajo fueron los siguientes:

- Clasificación K-Nearest Neighbors.
- Arbol de decisión.
- Support vector classification.
- Regresión Logistica.

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

Modelo	Precisión
Clasificación K-Nearest Neighbors.	0.5705426356589147
Arbol de decisión.	0.5707641196013288
Support vector classification	0.5708748615725359
Arbol de Decisión.	0.5610188261351052

Al ver esta tabla podemos concluir que cualquiera de los metodos puede darte un buen de modelo de clasificación. Aun asi el que mejor rendimiento tiene por muy poco es el metodo SVC.

Una vez logramos ver esto ponemos este modelo a prueba con los datos que nos sobrarón para poner el modelo a prueba y llegamos podemos ver que este termina con una precisión de 0.5494. Muy cercana a la obtenida anteriormente usando el conjunto de entrenamiento. Podemos ver que es un buen modelo.

Conclusión:

Tras aplicar los modelos y evaluar su desempeño utilizando su precisión se selecciona el modelo de SVC como el que mejor precisión tiene. Es interesante ver como cada una de las técnicas te ofrece un resultado diferente y además me sorprende lo fácil que son de implementar al usar toda la tecnología que nos proporcionan las distintas librerías de Python

Bibliografia:

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

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