LDA para clasificación

PRESENTA

ING. ALEJANDRO NOEL HERNÁNDEZ GUTÍERREZ

MATERIA

MODELADO PREDICTIVO

IMPARTE

DR. RIEMANN RUÍZ CRUZ

ITESO

INSTITUTO TECNOLÓGICO DE ESTUDIOS SUPERIORES DE OCCIDENTE

MAESTRÍA EN CIENCIA DE DATOS

SEPTIEMBRE 2023

Contenido

Introducción

El conjunto de datos a explorar en el presente reporte es el de "*Heart_Disease*". La base de datos cuenta con 76 atributos de los cuales, en procesos previos, se han considerado 14 de ellos para análisis de ML.

Estos mismos 14 atributos son los que están disponibles para el estudio y serán descritos a continuación.

Nombre del atributo	**Rol**	**Tipo**	**Descripción**	**Unidades**	**Missing Values**
age	Entrada	Entero	Edad del paciente	Años	Falso
sex	Entrada	Categórica	Sexo del paciente	1: Masculino 0: Femenino	Falso
cp	Entrada	Categórica	Dolor en el pecho	1: Angina típica 2: Angina atípica 3: Dolor no-anginal 4: Asintomático	Falso
trestbps	Entrada	Entero	Presión arterial en reposo	mm Hg	Falso
chol	Entrada	Entero	Colesterol sérico	mg/dl	Falso
fbs	Entrada	Categórica	Glusemia en ayunas (Azucar en sangre) > 120 mg/dl	1: Cierto 0: Falso	Falso
restecg	Entrada	Categórica	Electrocardiograma en reposo	0: Normal 1: Anomalía en onda ST-T 2: Muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes	Falso

Nombre del atributo	Rol	Tipo	Descripción	Unidades	Missing Values
thalach	Entrada	Entero	frecuencia cardíaca máxima alcanzada	BPM	Falso
exang	Entrada	Categórica	Angina inducida por ejercicio	1: Si 0: No	Falso
oldpeak	Entrada	Entero	Depresión ST inducida por ejercicio relativa al descanso.		Falso
slope	Entrada	Categórica	Pendiente del segmento ST del ejercicio máximo	1: Pendiente ascendente 2: Plano 3: Pendiente descendiente	Falso
ca	Entrada	Entero	Número de vasos principales (0-3) coloreados por fluoroscopia		Cierto
thal	Entrada	Categórica	Tipo de ritmo cardiaco	3: Normal 6: Defecto arreglado 7: Defecto reversible	Cierto
num	Objetivo	Entero	Diagnóstico de enfermedad del corazón (estado de la enfermedad angiográfica)	0: < 50% de estrechamiento del diámetro 1: > 50% de estrechamiento del diámetro	Falso

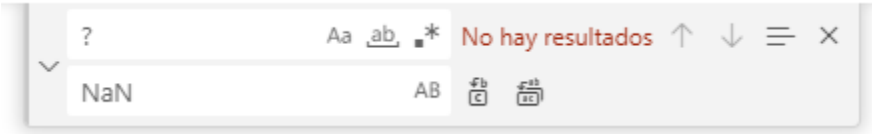
El objetivo de este análisis es clasificar la presencia de enfermedades del corazón. De la variable objetivo, la clase 0 representa ausencia de cualquier enfermedad y las clases 1,2,3 y 4 presencia.

Desarrollo

Capitulo 0. Exploración de los datos y sustitución de datos faltantes

Para facilitar el análisis se realizaron las siguientes acciones.

- Se convierte el archivo “processed.cleveland.data” a csv
- Se agrega al archivo csv el nombre de las columnas
- Se sustituye el carácter “?” por la palabra “NaN” para que Python interprete esos valores como faltantes.




```
RangeIndex: 303 entries, 0 to 302
Data columns (total 13 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    float64
 1   sex       303 non-null    float64
 2   cp        303 non-null    float64
 3   trestbps  303 non-null    float64
 4   chol      303 non-null    float64
 5   fbs       303 non-null    float64
 6   restecg   303 non-null    float64
 7   thalach   303 non-null    float64
 8   exang     303 non-null    float64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    float64
 11  ca        299 non-null    float64
 12  thal      301 non-null    float64
```

Nos damos cuenta de que ca y thal tienen valores faltantes por lo que vamos a aplicar técnicas de sustitución de estos valores. Se podría considerar despreciable el porcentaje de ellos, pero al tratarse de sustituciones rápidas, se decide proceder con estas técnicas.

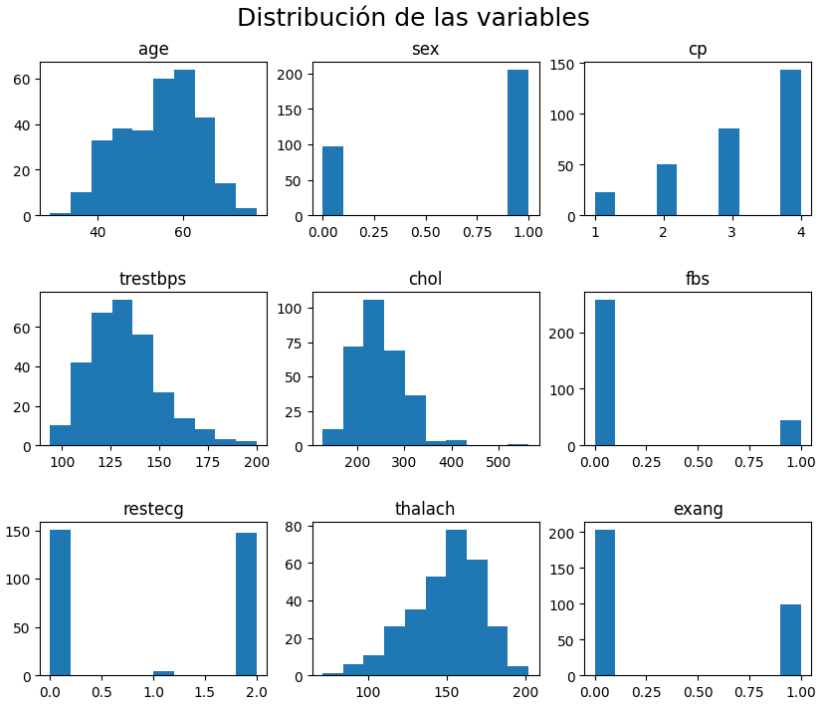
- Como ca y thal son variables categóricas, la sustitución se realiza por su moda, que es “1.0” y “3.0” respectivamente. Una vez echa la sustitución, se comprueba que no hay datos nulos.

```
RangeIndex: 303 entries, 0 to 302
Data columns (total 13 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    float64
 1   sex       303 non-null    float64
 2   cp        303 non-null    float64
 3   trestbps  303 non-null    float64
 4   chol      303 non-null    float64
 5   fbs       303 non-null    float64
 6   restecg   303 non-null    float64
 7   thalach   303 non-null    float64
 8   exang     303 non-null    float64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    float64
 11  ca        303 non-null    float64
 12  thal      303 non-null    float64
```

- Se realiza un análisis de outliers para detectar posibles errores de medición y se decide no eliminar ningún valor al considerar que todos son posibles y pueden aportar información al modelo.

Distribución de las variables:

En la distribución es posible ver claramente las variables categóricas y el comportamiento de las variables numéricas.



Capítulo 1. División de los datos en entrenamiento y prueba

El conjunto de datos seleccionado es el obtenido por la *V.A. Medical Center, Long Beach and Cleveland Clinic Foundation*: processed.cleveland.csv. Este se divide (en el código en conjuntos de datos de entrenamiento y prueba para X e Y respectivamente.

Capítulo 2. Regresión logística sin tratamiento previo (reg_log)

En este capítulo se aborda el entrenamiento de un modelo de regresión logística para estimar la variable objetivo "num" con las 13 variables de entrada. Se obtiene el accuracy en entrenamiento y prueba para evaluar la efectividad del modelo.

Los resultados son:

- Entrenamiento accuracy score: 0.67
- Prueba accuracy score: 0.54

Luego de haber corrido el reporte completo con la clase 0 (no enfermo) y las clases 1,2,3 y 4 como distintos niveles de enfermad del corazón. El Accuracy máximo presentado fue alrededor de 68.

En este estudio la parte más importante es predecir si el paciente está enfermo del corazón o no de forma exitosa, por tanto, se reduce a un problema de clasificación binario para mejorar la predicción.

Con las clases binarias (0: no enfermo y 1: enfermo) se obtienen los siguientes resultados:

Entrenamiento accuracy score: 0.83

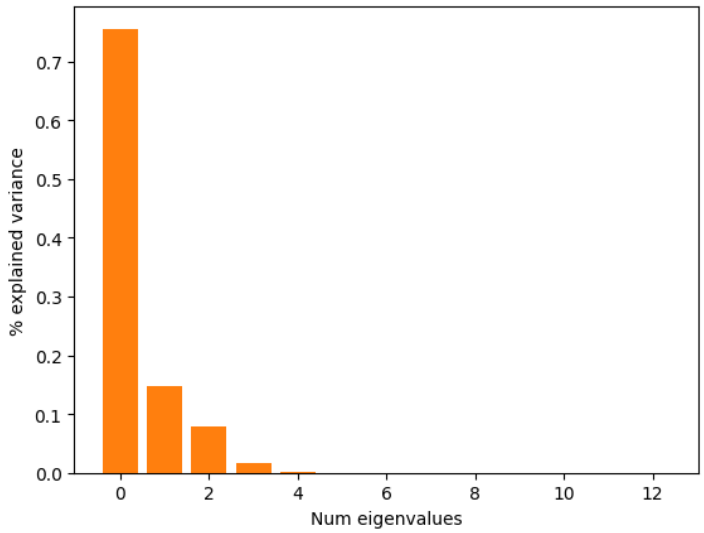
Prueba accuracy score: 0.89

Capítulo 3. Regresión logística con reducción de variables por PCA (reg_log_rdv_PCA)

Considerando el conjunto de datos usado en el capítulo 1, se ejecuta una reducción de variables con PCA. Con el conjunto de datos reducido, se entrena un nuevo modelo de regresión logística y se calcula el accuracy para comparar el desempeño.

Durante el análisis de PCA, para seleccionar los componentes principales (los que aportan más información), se revisan los eigenvalores, estos determinan la importancia de cada componente principal. Cuanto mayor sea el eigenvalor de una componente principal, más varianza del conjunto de datos explica esa componente. Las componentes principales se ordenan en función de la magnitud de sus eigenvalores, de manera que la primera componente principal (la que tiene el eigenvalor más grande) explica la mayor cantidad de varianza, seguida de la segunda componente principal y así sucesivamente.

En este caso podemos ver que tomando la componente 0, 1 y 2 incluyen la mayor parte de la varianza acumulada. Estos componentes son los que se seleccionan.



Después de entrenar el modelo estos son los resultados:

```
Entrenamiento accuracy score: 0.55
Prueba accuracy score: 0.49
```

Con las clases binarias (0: no enfermo y 1: enfermo) se obtienen los siguientes resultados:

```
Entrenamiento accuracy score: 0.71
Prueba accuracy score: 0.75
```

Capítulo 4. Regresión logística con reducción de variables por LDA (reg_log_rdv_LDA)

Considerando el conjunto de datos usado en el capítulo 1, se ejecuta una reducción de variables con LDA. Con el conjunto de datos reducido, se entrena un nuevo modelo de regresión logística y se calcula el accuracy para comparar el desempeño.

Después de entrenar el modelo estos son los resultados:

Entrenamiento accuracy score: 0.65

Prueba accuracy score: 0.54

Con las clases binarias (0: no enfermo y 1: enfermo) se obtienen los siguientes resultados:

Entrenamiento accuracy score: 0.83

Prueba accuracy score: 0.90

Capítulo 5. LDA

Considerando el conjunto de datos usado en el capítulo 1, se entrena un modelo LDA y se calcula el accuracy para comparar el desempeño.

Después de entrenar el modelo estos son los resultados:

Entrenamiento accuracy score: 0.68

Prueba accuracy score: 0.49

Con las clases binarias (0: no enfermo y 1: enfermo) se obtienen los siguientes resultados:

Entrenamiento accuracy score: 0.83

Prueba accuracy score: 0.90

Capítulo 6. Tabla comparativa de los modelos

Modelo	**Accuracy clases (0,1,2,3,4)**		**Accuracy clases (0: enfermo, 1: no enfermo)**	
	Entrenamiento	Prueba	Entrenamiento	Prueba
reg_log_mod	0.67	0.54	0.83	0.89
reg_log_rdv_PCA_mod	0.55	0.49	0.71	0.75
reg_log_rdv_LDA_mod	0.65	0.54	0.83	0.90
LDA_mod	0.68	0.49	0.83	0.90

Conclusiones

Para este problema específico y el conjunto de datos disponible. La mejor opción es simplificar la variable objetivo para predecir si un paciente está enfermo o no, porque es más exacto que tratar de clasificar los diferentes tipos de posibles enfermedades y el modelo que arrojó mejores resultados fue el LDA en sus dos versiones, aprovechando la transformación y aplicando un modelo de regresión para predecir o simplemente aplicando el LDA para entrenamiento y predicción.

Como profesional de la ciencia de datos, sugeriría para casos reales, que se entrenen dos modelos que expongo a continuación:

Primero entrenar el modelo de clasificación LDA con las clases (1,2,3 y 4) unificadas como (1: Enfermo del corazón) y aplicar el modelo a los nuevos pacientes para que, en caso de salir positivo a esta prueba, aplicar el segundo modelo.

Segundo entrenar un modelo de regresión logística para la fase en la que se quiera categorizar la enfermedad de un paciente que ya se aceptó como posible enfermo del corazón por el primer modelo. Sugiero este modelo porque presentó un accuracy aceptable en la etapa de prueba. Esto será útil para tener una mejor aproximación antes de continuar con el diagnóstico. Sugiero también que se contemple la posibilidad de entrenar el modelo sin la clase 0, es decir utilizando solo los niveles de enfermedad del corazón.

Bibliografía

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, y Detrano,Robert. (1988*). Heart Disease. UCI Machine Learning Repository*. https://doi.org/10.24432/C52P4X.

Amat Rodrigo, Joaquín (2016). *Regresión logística simple y múltiple.* https://cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple