**Universidad de Medellín**

**Énfasis II – Aprendizaje de Máquina**

**Proyecto final**

Usando el conjunto de datos que se le suministra, implemente un modelo que prediga si un paciente tiene o no enfermedades del corazón. El dataset contiene las siguientes variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Número de columna | Nombre de variable | Tipo de variable | Explicación |
| 1 | Edad | Numérica | Edad del paciente en años |
| 2 | Sexo | Categórica | Sexo del paciente. Puede tomar 2 valores:  1 = Hombre  0 = Mujer |
| 3 | DP | Categórica | Dolor de pecho. Puede tomar 4 valores:  1 = angina típica  2 = angina atípica  3 = dolor no debdo a angina  4 = asintomático |
| 4 | PA | Numérica | Presión arterial en mm Hg |
| 5 | COL | Numérica | Colesterol total en mg/dl |
| 6 | AZ | Categórica | Prueba de azúcar en la sangre mayor a 120 mg/dl. Puede tomar 2 valores:  1 = verdadero  0 = falso |
| 7 | ECG | Categórica | Resultado de electrocardiograma en descanso. Puede tomar 3 valores:  0 = normal  1 = tiene anormalidades en la forma de onda  2 = exhibe hipertrofia en el ventrículo izquierdo |
| 8 | MAXFC | Numérica | Máxima frecuencia cardiaca alcanzada |
| 9 | EJANG | Categórica | Angina inducida por ejercicio. Puede tomar 2 valores:  1 = si  0 = no |
| 10 | ST | Numérica | Depresión del segmento ST del ECG inducida por el ejercicio |
| 11 | Pendiente | Categórica | Pendiente del segmento ST. Puede tomar 3 valores:  1 = Hacia arriba  2 = Plana  3 = Hacia abajo |
| 12 | HR | Categórica | Puede tomar 3 valores:  3 = normal  6 = defectos fijos  7 =defectos reversibles |
| 13 | DIAG | Categórica | Diagnóstico de enfermedad cardíaca. Puede tomar 2 valores:  0 = diagnóstico negativo (no tiene enfermedad)  1 = diagnóstico positivo (si tiene enfermedad)  Esta es la variable a predecir |

Asegúrese que cada variable sea del tipo correcto, de acuerdo con la tabla de arriba; en caso de que no lo sea tiene que cambiar el tipo de variable.

Si más del 80% de los valores de una variable son datos nulos, elimine las columnas correspondientes. En los demás casos, haga las siguientes imputaciones:

* si la variable es categórica impute el valor más frecuente a los datos
* si la variable es numérica impute el valor promedio

Separe las variables en salida (la variable DIAG), y entradas (el resto de las variables).

Estandarice las variables de entrada numéricas, y codifique las variables de entrada categóricas.

Cada estudiante debe implementar solo una de las opciones referidas abajo.

**Opción A: (Cristian)**

Implemente un modelo de bosque aleatorio, y ajuste el modelo optimizando los siguientes hiperparámetros: número de estimadores, factor de complejidad (ccp\_alpha). Debe indicar cuales fueron los mejores valores hallados de estos hiperparámetros, y cuál fue el valor obtenido de la métrica utilizada con estos hiperparámetros.

**Opción B: (Daniel)**

Implemente un modelo de gradient boosting con parada temprana (n\_iter\_no\_change=5, tol=0.001), y ajuste el modelo optimizando los siguientes hiperparámetros: número de estimadores, factor de complejidad (ccp\_alpha). Debe indicar cuales fueron los mejores valores hallados de estos hiperparámetros y del número de estimadores, y cuál fue el valor obtenido de la métrica utilizada con estos hiperparámetros.

**Opción C: (Fayder)**

Implemente un modelo de máquinas de vectores de soporte y ajuste el modelo optimizando los siguientes parámetros: kernel (‘rbf’, ‘poly’, ‘sigmoid’), factor de complejidad (C), gamma. Debe indicar cuales fueron los mejores valores hallados de estos hiperparámetros, y cuál fue el valor obtenido de la métrica utilizada con estos hiperparámetros.

**Opción D: (Romario)**

Implemente un modelo de regresión logística con solver ‘saga’ y max\_iter=10000, y ajuste el modelo optimizando los siguientes hiperparámetros: tipo de regularización (l1, l2, elasticnet), coeficiente de regularización (C), relación regularización l1 – l2 (li\_ratio). Debe indicar cuales fueron los mejores valores hallados de estos hiperparámetros, y cuál fue el valor obtenido de la métrica utilizada con estos hiperparámetros.

**Opción E: (Sebastian)**

Implemente un modelo de vecinos más cercanos y ajuste el modelo optimizando los siguientes hiperparámetros: número de vecinos, potencia de Minkowski (p), pesos (weights). Debe indicar cuales fueron los mejores valores hallados de estos hiperparámetros, y cuál fue el valor obtenido de la métrica utilizada con estos hiperparámetros.

En todos los casos debe presentar la matriz de confusión obtenida.

**Debe enviar el notebook, asegurándose que cada celda de código se ejecute correctamente en el orden que están. Documente cada celda de código, indicando qué está haciendo y para que lo está haciendo.**