UD02_01 Práctica Predicción enfermedades cardíacas

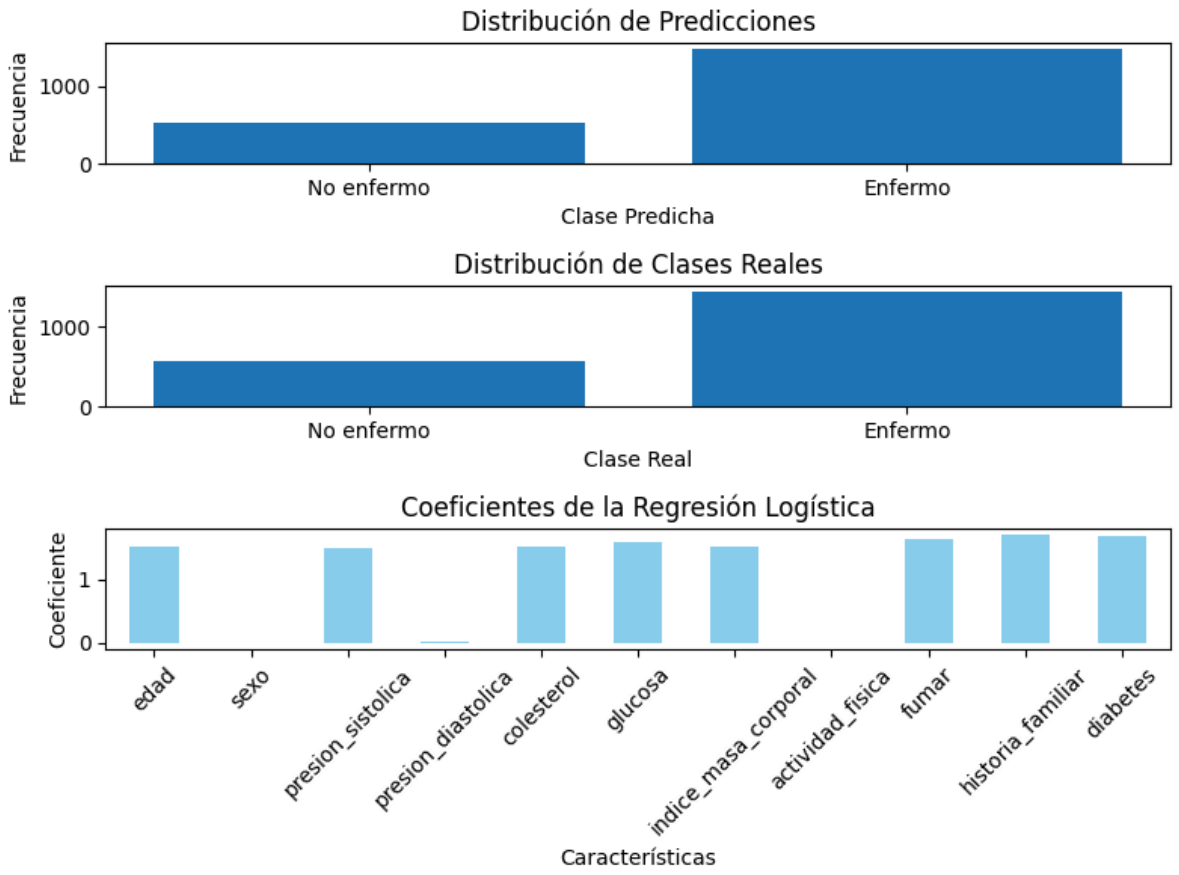
Sistemas de Aprendizaje Automático.
Álvaro Martínez Lineros

Análisis del problema.

El problema planteado sugiere la creación de un modelo de predicción para pacientes que pueden sufrir enfermedades cardíacas en función de otras métricas de salud. Se deberán analizar los factores de riesgo para encontrar aquellas métricas que tengan más peso en la aparición de enfermedades cardíacas.

Regresión Logística.

Se entrenó un modelo utilizando el algoritmo de regresión logística con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:



Matriz de confusión:

Verdaderos negativos	452	Falsos positivos	112
Falsos negativos	71	Verdaderos positivos	1365

Reporte de clasificación:

	precision	recall	f1-score	support
No enfermo	0.86	0.80	0.83	564
Enfermo	0.92	0.95	0.94	1436
accuracy			0.91	2000
macro avg				
weighted avg	0.91	0.91	0.91	2000

Coeficientes de las características:

Característica	Coeficiente
edad	1.503978
sexo	-0.003410
presión sistólica	1.481234
presión diastólica	0.010884
colesterol	1.502532
glucosa	1.577629
indice_masa_corporal	1.513781
actividad_fisica	-0.015763
fumar	1.638051
historia_familiar	1.707357
diabetes	1.669886

Interpretación de los resultados.

Se obtuvo una precisión del 91% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 71 personas

con enfermedades cardíacas que no se previeron correctamente, adicionalmente hubo 112 personas no enfermas diagnosticadas como enfermas.

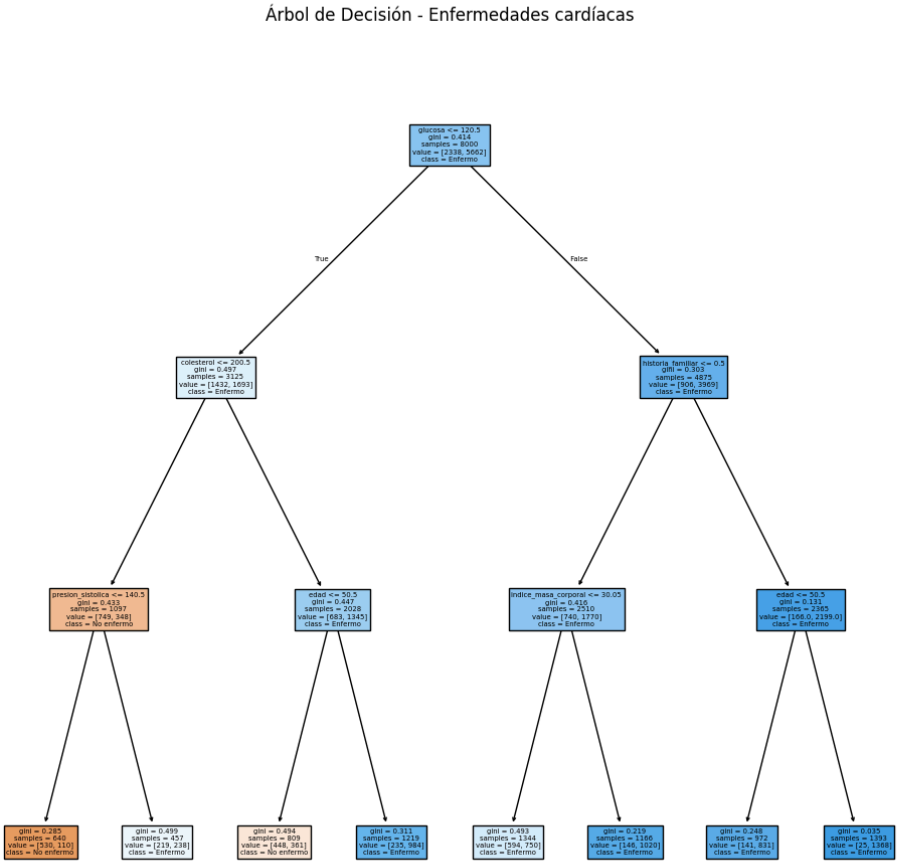
El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
    - Se acertó el 86% de predicciones como no enfermo.
    - Se acertó el 92% de predicciones como enfermo.
- Recall:
    - El 80% de los no enfermos fueron identificados correctamente.
    - El 95% de los enfermos fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para no enfermos podría mejorarse, pero muestra un buen rendimiento.
    - El balance entre precisión y recall para enfermos muestra un rendimiento muy alto.
- Support:
    - La muestra cuenta con 564 pacientes no enfermos.
    - La muestra cuenta con 1436 pacientes diagnosticados.
- Accuracy:
    - La precisión total del modelo es del 91%
- Weighted avg:
    - El promedio ponderado de las métricas es del 0.91

Finalmente, se puede observar el peso de las características con su coeficiente. La característica que más influye es la historia familiar seguido de la diabetes. Por otro lado el sexo tiene un impacto muy bajo en la probabilidad de desarrollar una enfermedad cardíaca. La actividad física sugiere que si somos más activos disminuye la probabilidad de sufrir una enfermedad cardíaca.

Árboles de decisión.

Se entrenó un modelo utilizando el algoritmo de árboles de decisión de 3 niveles con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:



Matriz de confusión:

Verdaderos negativos 240	Falsos positivos 324
Falsos negativos 146	Verdaderos positivos 1290

Reporte de clasificación:

	precision	recall	f1-score	support
No enfermo	0.62	0.43	0.51	564
Enfermo	0.80	0.90	0.85	1436

accuracy			0.77	2000
macro avg	0.71	0.66	0.68	2000
weighted avg	0.75	0.77	0.75	2000

Importancias de las características:

Característica	Importancia
edad	0.159366
sexo	0.000000
presión sistólica	0.071231
presión diastólica	0.000000
colesterol	0.187019
glucosa	0.310079
indice_masa_corporal	0.137468
actividad_fisica	0.000000
fumar	0.000000
historia_familiar	0.134838
diabetes	0.000000

Interpretación de los resultados.

Se obtuvo una precisión del 77% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo 146 personas con enfermedades cardíacas que no se previeron correctamente, adicionalmente hubo 324 personas no enfermas diagnosticadas como enfermas.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
    - Se acertó el 62% de predicciones como no enfermo.
    - Se acertó el 80% de predicciones como enfermo.
- Recall:
    - El 43% de los no enfermos fueron identificados correctamente.
    - El 90% de los enfermos fueron identificados correctamente.
- F1-score:
    - El balance entre precisión y recall para no enfermos es del 0.51, es decir, es un rendimiento muy bajo para un modelo predictivo.

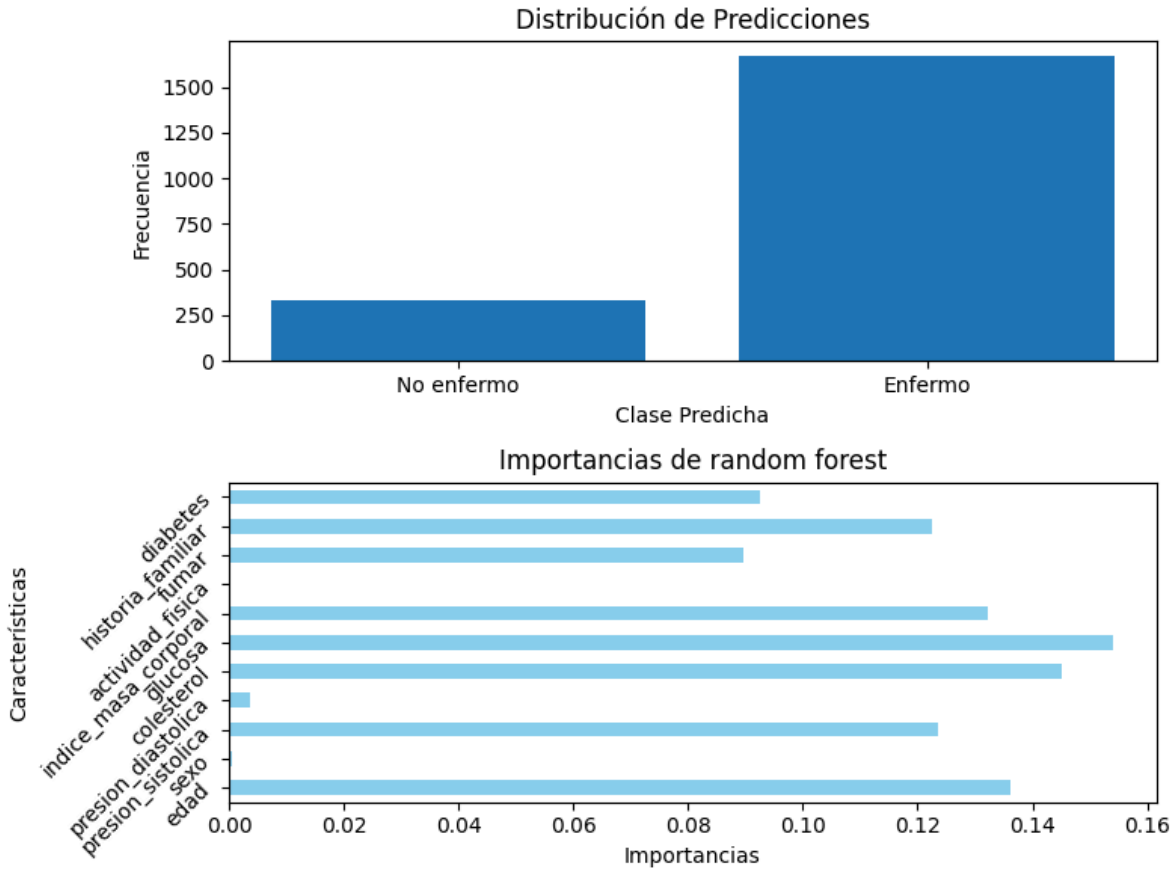
    - El balance entre precisión y recall para enfermos muestra un rendimiento del 0.85, que podría mejorarse.
- Support:
    - La muestra cuenta con 564 pacientes no enfermos.
    - La muestra cuenta con 1436 pacientes diagnosticados.
- Accuracy:
    - La precisión total del modelo es del 77%
- Weighted avg:
    - El promedio ponderado de las métricas es del 0.75 para la precisión, 0.77 para el recall y 0.75 para la f1-score.

Se puede observar la importancia de las características con su índice. Para la predicción de este modelo, la glucosa, el colesterol y la edad son los factores más determinantes a la hora de predecir si una persona padece o no una enfermedad cardíaca. Por otro lado, la actividad física, fumar y la diabetes no tienen ningún peso en esta decisión.

Se ha de aclarar que añadiendo más niveles de decisión en el árbol se puede mejorar la precisión del modelo.

Random Forest.

Se entrenó un modelo utilizando el algoritmo de random forest (con 100 árboles de decisión y una profundidad máxima de 5 niveles en cada árbol) con el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:



Matriz de confusión:

Verdaderos negativos	334	Falsos positivos	230
Falsos negativos	0	Verdaderos positivos	1436

Reporte de clasificación:

	precision	recall	f1-score	support
No enfermo	1.00	0.59	0.74	564
Enfermo	0.86	1.00	0.93	1436
accuracy			0.89	2000
macro avg				
weighted avg	0.90	0.89	0.87	2000

Importancias de las características:

Característica	Importancia
edad	0.136056
sexo	0.000445
presión sistólica	0.123617
presión diastólica	0.003660
colesterol	0.145054
glucosa	0.153947
indice_masa_corporal	0.132187
actividad_fisica	0.000283
fumar	0.089642
historia_familiar	0.122635
diabetes	0.092475

Interpretación de los resultados.

Se obtuvo una precisión del 89% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. No hubo falsos negativos, es decir, no hubo personas con enfermedades cardíacas que no se

diagnosticaron. Adicionalmente hubo 230 personas no enfermas diagnosticadas como enfermas.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
  - Se acertó el 100% de predicciones como no enfermo.
  - Se acertó el 86% de predicciones como enfermo.
- Recall:
  - El 59% de los no enfermos fueron identificados correctamente.
  - El 100% de los enfermos fueron identificados correctamente.
- F1-score:
  - El balance entre precisión y recall para no enfermos es del 0.71, es un desempeño aceptable pero muy mejorable..
  - El balance entre precisión y recall para enfermos muestra un rendimiento del 0.93, mostrando una alta predictibilidad del modelo.
- Support:
  - La muestra cuenta con 564 pacientes no enfermos.
  - La muestra cuenta con 1436 pacientes diagnosticados.
- Accuracy:
  - La precisión total del modelo es del 89%
- Weighted avg:
  - El promedio ponderado de las métricas es de 0.90 para la precisión, 0.89 para el recall y 0.87 para la f1-score.

Se puede observar la importancia de las características con su índice. Para la predicción de este modelo, la glucosa, el colesterol y la edad son los factores más determinantes a la hora de predecir si una persona padece o no una enfermedad cardíaca. Por otro lado, la actividad física, el sexo, fumar y la diabetes no tienen prácticamente ningún peso en esta decisión.

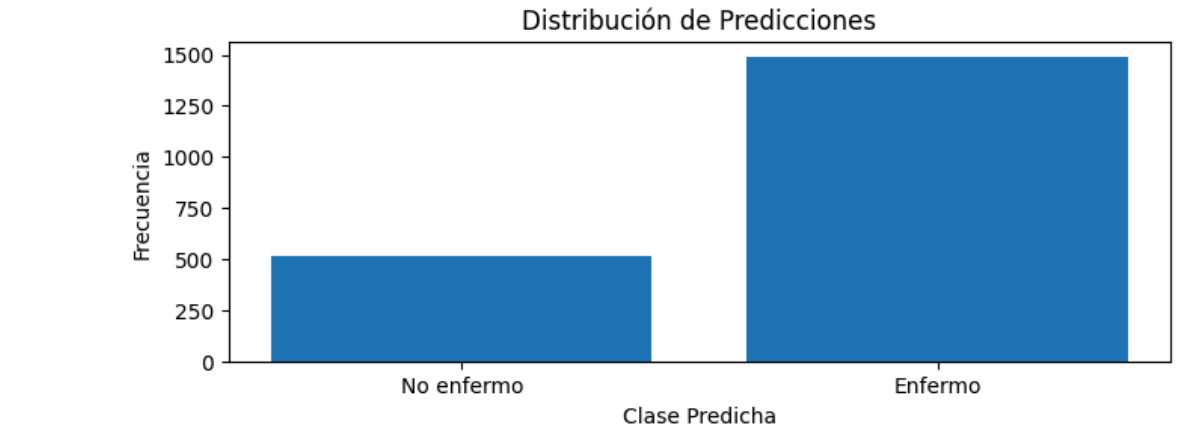
Se ha de aclarar que añadiendo más niveles de profundidad en el algoritmo se puede mejorar notablemente el desempeño.

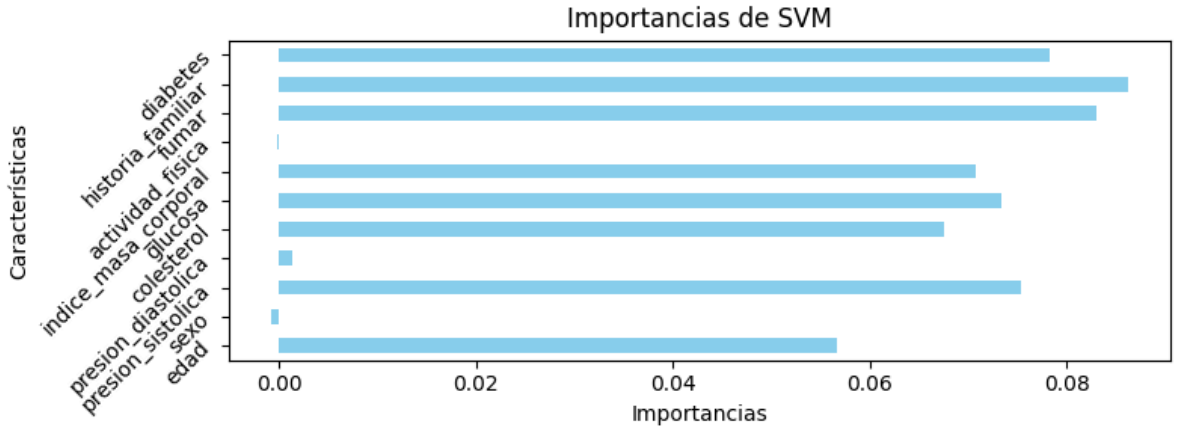
Support Vector Machines (SVM).

Se entrenó un modelo utilizando el algoritmo de SVM con:

- kernel rbf: un kernel no lineal que mide la similitud entre los puntos de datos y crea una curvatura.
- C con valor de 1.0: valor por defecto para la tolerancia a errores
- Gamma de tipo scale: ajusta gamma según la fórmula $\gamma = 1 / n \cdot \text{varianza}(X)$. Gamma es el parámetro que controla la influencia de cada punto en la función de decisión.

Se entrenó el dataset proporcionado y se puso a prueba con una parte de los mismos (80% datos de entrenamiento, 20% datos para prueba). Se obtuvieron los siguientes resultados:





Matriz de confusión:

Verdaderos negativos	455	Falsos positivos	109
Falsos negativos	59	Verdaderos positivos	1377

Reporte de clasificación:

	precision	recall	f1-score	support
No enfermo	0.89	0.81	0.84	564
Enfermo	0.93	0.96	0.94	1436
accuracy			0.92	2000
macro avg	0.91	0.88	0.89	2000
weighted avg	0.91	0.92	0.91	2000

Importancias de las características:

Característica	Importancia
edad	0.05670
sexo	-0.00075
presión sistólica	0.07535
presión diastólica	0.00130
colesterol	0.06760
glucosa	0.07335
indice_masa_corporal	0.07085
actividad_fisica	-0.00025
fumar	0.08305
historia_familiar	0.08620
diabetes	0.07825

Interpretación de los resultados.

Se obtuvo una precisión del 92% siguiendo este algoritmo. Si se observa la matriz de confusión podemos ver la cantidad de valores en los que falló el modelo. Hubo un total de 59 personas con enfermedades cardíacas no identificadas correctamente. Adicionalmente hubo 109 personas no enfermas diagnosticadas como enfermas.

El reporte de clasificación informa en detalle del desempeño del modelo. Este modelo entrenado con regresión logística indica que:

- Precisión:
  - Se acertó el 89% de predicciones como no enfermo.
  - Se acertó el 93% de predicciones como enfermo.
- Recall:
  - El 81% de los no enfermos fueron identificados correctamente.
  - El 96% de los enfermos fueron identificados correctamente.
- F1-score:
  - El balance entre precisión y recall para no enfermos es del 0.84, es un desempeño aceptable pero mejorable..
  - El balance entre precisión y recall para enfermos muestra un rendimiento del 0.94, mostrando una alta predictibilidad del modelo.
- Support:
  - La muestra cuenta con 564 pacientes no enfermos.
  - La muestra cuenta con 1436 pacientes diagnosticados.
- Accuracy:
  - La precisión total del modelo es del 92%
- Weighted avg:

- El promedio ponderado de las métricas es de 0.91 para la precisión, 0.92 para el recall y 0.91 para la f1-score.

Se puede observar la importancia de las características con su índice. Para la predicción de este modelo, la historia familiar, fumar y la diabetes son los factores más determinantes a la hora de predecir si una persona padece o no una enfermedad cardíaca. Por otro lado, la actividad física, el sexo, la actividad física y la presión distólica no tienen prácticamente ningún peso en esta decisión.

Se ha de aclarar que añadiendo más niveles de profundidad en el algoritmo se puede mejorar notablemente el desempeño.

Comparación de los resultados.

El modelo más preciso es el entrenado con Support Vector Machines, sin embargo, tratándose de un diagnóstico de enfermedades cardíacas el modelo entrenado con Random Forest parece el más adecuado. Aunque su precisión sea menor (89%) este modelo no presenta falsos negativos, es decir, no diagnostica erróneamente a personas enfermas como no enfermas. Adicionalmente, podemos añadir densidad a los árboles del modelo para hacerlo más preciso (con una densidad máxima de 7 se obtiene un 97% de precisión).