PREDICCIÓN DIABETES

Realizado por:

Manuel José López Muñoz

ÍNDICE

ENUNCIADO

Este dataset, Pima Indians Diabetes Database, procede del National Institute of Diabetes and Digestive and Kidney Diseases. Su objetivo es predecir si un paciente tiene o no diabetes, basándose en determinadas mediciones diagnósticas. Se impusieron varias restricciones a la selección de estos casos de una base de datos más amplia. En particular, todos los pacientes son mujeres de al menos 21 años y de ascendencia india Pima.

La variable diagnóstica a predecir es Outcome.

Las features (variables predictoras) del dataset son las siguientes:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function (DPF) calculates diabetes likelihood depending on the subject's age and his/her diabetic family history.
- Age (years)

La práctica consta de 4 apartados, cada uno con la ponderación de nota que se indica.

A. (20%) MODELO DE REGRESIÓN LOGÍSTICA

Presenta los resultados del modelo proporcionando los siguientes cálculos:

- Accuracy
- Matriz de confusión
- Informe de clasificación

Acompaña tu respuesta del cuaderno de Jupyter correspondiente.

En el cuaderno de Jupyter se pueden observar varias pruebas de modelos, finalmente, las variables seleccionadas para el modelo son:

- 'Glucose', 'BMI', 'Age', 'Pregnancies', 'DiabetesPedigreeFunction', 'BloodPressure'

Para este modelo hemos tomado un umbral de 30% de probabilidad, para reducir el número de falsos negativos y un buen balance de falsos positivos.

- **Accuracy**: 0.7662337662337663
- **Matriz de Confusión:**

116	**36**
18	**61**

- **Informe de Clasificación:**

	Precision	Recall	F1-Score	Support
0	**0.87**	**0.76**	**0.81**	**152**
1	**0.63**	**0.77**	**0.69**	**79**
Accuracy			**0.77**	**231**
Macro Avg	**0.75**	**0.77**	**0.75**	**231**
Weighted Avg	**0.78**	**0.77**	**0.77**	**231**

B. (20%) CÁLCULO DE MÉTRICAS

A partir de la matriz de confusión del apartado anterior, y utilizando la siguiente nomenclatura:

- TP: True Positives (VERDADEROS positivos)
- FN: False Negatives (FALSOS negativos)
- FP: False Positives (FALSOS positivos)
- TN: True Negatives (VERDADEROS negativos)

Escribe la fórmula y calcula a mano cada una de las siguientes métricas, comprobando que los resultados se corresponden con los del cuaderno de Jupyter:

- Accuracy
- Precision
- Specificity
- Sensitivity
- F1 score

Cálculo de Accuracy

(TP + TN) / (TP + TN + FP + FN) = (61+ 116) / (61 + 116 + 36 + 18) = 0.766

Cálculo de Precision

TP / (TP + FP) = 61 / (61 + 36) = 0.6288

Cálculo de Specificity

TN / (TN + FP) = 0.7631

Cálculo de Sensitivity (Recall)

TP / (TP + FN) = 0.7721

Cálculo del F1 Score

2 * (Precision * recall) / (precision + recall) = 0.6932

Comparación de métricas (Ver código en Notebook):

	Metrica	Valor Modelo	Calculo manual	Métricas iguales	Diferencia
0	Accuracy	0.766234	0.766234	True	0.0
1	Precision	0.628866	0.628866	True	0.0
2	Recall	0.772152	0.772152	True	0.0
3	Specificity	0.763158	0.763158	True	0.0
4	F1 Score	0.693182	0.693182	True	0.0

C. (30%) COMPRENSIÓN DE LOS RESULTADOS

Imagina que eres el médico al que presentan el modelo, el cual espera le sirva para realizar un pre-diagnóstico (cribado) en la consulta de cara a decidir qué pruebas médicas realizar al paciente. Ten en cuenta que si la predicción es negativa, no quiere decir necesariamente que no haya que hacerle ninguna prueba.

Haz una descripción razonada de las limitaciones del modelo formulándote a ti mismo preguntas como las siguientes (se trata simplemente de ejemplos para facilitarte la construcción de tu argumentación):

- **¿Se escapan muchos pacientes que el médico criba erróneamente como NEGATIVOS, puesto que son realmente POSITIVOS?**

La sensibilidad (o recall) del modelo es de 77.21%, lo que significa que aproximadamente un 22.79% de los pacientes reales con diabetes podrían ser clasificados erróneamente como no diabéticos (falsos negativos). Esto es una preocupación significativa, ya que estos pacientes podrían no recibir el seguimiento y tratamiento oportunos.

- **¿Se escapan muchos pacientes que el médico criba erróneamente como POSITIVOS, puesto que son realmente NEGATIVOS?**

La especificidad del modelo es de 76.31%, indicando que alrededor del 23.69% de los pacientes que no tienen diabetes podrían ser identificados erróneamente como diabéticos (falsos positivos). Estos casos podrían llevar a pruebas médicas adicionales innecesarias y potencialmente a un estrés indebido para los pacientes.

- **¿Cuál de los potenciales errores de cribado de las dos preguntas anteriores es más perjudicial para el paciente? Es decir, ¿es peor diagnosticar a un paciente como diabético cuando realmente no lo es, o por el contrario concluir que no es diabético cuando realmente lo es?**

En un contexto médico, generalmente se considera más grave un falso negativo que un falso positivo. Un falso negativo (clasificar erróneamente a un paciente diabético como no diabético) podría resultar en la falta de tratamiento necesario y en la progresión de la enfermedad. En cambio, un falso positivo generalmente lleva a pruebas adicionales que pueden aclarar el diagnóstico. Aunque esto depende en mayor medida de los objetivos del negocio y del coste de estas pruebas. Sería importante hablar con la persona responsable para conocer sus costes y prioridades.

Al no conocer este dato, realizaré un análisis situacional con varias suposiciones, dependiendo la compañía interesada en el modelo:

- **Compañías de seguros de salud**: Estas compañías pueden estar interesadas en minimizar los falsos positivos para evitar costos adicionales en pruebas innecesarias. Por lo tanto, podrían preferir establecer un umbral más alto (como 0.5), lo que incrementa la precisión pero aumenta los falsos negativos. Se debe valorar el balance entre el coste generado por seguros más económicos a clientes diagnosticados como negativos pero que realmente tienen diabetes, frente al ahorro en pruebas para asegurar el diagnóstico.

- **Hospitales y clínicas públicas**: Los hospitales públicos a menudo enfrentan limitaciones presupuestarias y una alta demanda de atención médica. Su prioridad es ofrecer una atención médica de calidad a todos los pacientes, por lo que su umbral podría ser menor al de las compañías de seguros, sin embargo, este umbral no debe ser muy bajo, ya que un exceso de diagnósticos satura el sistema de salud pública, por tanto, en este caso sería más interesante ofrecer consejos de mejora de la calidad de vida a pacientes con un bajo riesgo y realizar pruebas más costosas a pacientes con un riesgo mayor.

- **Hospitales y clínicas públicas:** Debido a la naturaleza del negocio, los hospitales privados tienden a realizar un mayor número de pruebas, ya que, genera ingresos por cada procedimiento o prueba realizada. Por lo tanto, en este contexto, los hospitales privados pueden estar más inclinados a establecer umbrales de predicción más bajos para maximizar su ratio de detección de la enfermedad y a su vez maximizar las oportunidades de realizar pruebas adicionales.

- **Programas de prevención de la Diabetes:** Estos programas tienen un enfoque más preventivo, buscando detectar personas en riesgo de desarrollar la enfermedad, para darles recomendaciones de estilos de vida saludables. Por tanto, en este caso, sería conveniente establecer un umbral de 0.1, el cual, reduce al mínimo el número de falsos negativos y el impacto perjudicial sobre los falsos positivos es mínimo,

- **¿Qué métrica es la que permite cuantificar el error identificado como más perjudicial?**

Desde un punto de vista médico y el caso particular de la diabetes, el error más perjudicial para los pacientes es el de los falsos negativos, ya que, no diagnosticar la enfermedad puede ser muy perjudicial para el paciente. En este caso, la métrica que cuantifica el error es el recall (sensibilidad), es decir, la proporción de casos positivos correctamente identificados sobre el total de casos positivos reales.

Por otro lado, desde un punto de vista económico, el error más perjudicial es el de los falsos positivos, por tanto, la métrica más importante en este caso es la especificidad, es decir, la proporción de negativos correctamente predichos, frente al total de casos negativos.

- **¿Cuál es la tasa de falsos positivos y falsos negativos del modelo? Es decir, ¿con qué frecuencia el modelo diagnostica erróneamente a un paciente como positivo o negativo?**

La tasa de falsos positivos es de 0.2368, es decir, el modelo diagnostica a un paciente sin diabetes incorrectamente un 23% de las veces. Por otro lado, la tasa de falsos negativos es de 0.2278, por lo tanto, el modelo diagnostica a un paciente con diabetes de forma incorrecta el 22% de las veces.

- **¿Existen limitaciones en los datos utilizados para entrenar el modelo? ¿Se ha entrenado con datos de pacientes de diferentes edades, géneros y grupos étnicos para evitar sesgos?**

El modelo se ha entrenado con un dataset de mujeres de al menos 21 años y de ascendencia india Pima. Por lo que su contenido está sesgado y puede no funcionar correctamente con pacientes de otras edades, géneros o grupos étnicos.

- **¿Se puede concluir que el modelo es objetivamente bueno o malo? ¿Hay alguna métrica de las estudiadas que permita obtener dicha conclusión objetiva?**

Podríamos establecer que el modelo tiene buena capacidad predictiva, teniendo en cuenta que todas sus métricas son de aproximadamente el 70% con un umbral de 0.3, dando un buen balance entre falsos negativos y falsos positivos (recall y specifity). Sin embargo, se debe tener en cuenta que existen varias limitaciones:

- **Datos y representatividad:** El modelo se ha entrenado en un conjunto de datos específico (mujeres Pima).Esto podría limitar su aplicabilidad a otros grupos demográficos o poblaciones. Además, los valores atípicos o los datos incorrectamente codificados (como 0 en algunas variables) pueden afectar la precisión del modelo.

- **Variables consideradas:** El modelo solo incluye variables disponibles en el conjunto de datos. Existen factores importantes no incluidos (como el estilo de vida o la historia familiar detallada) que podrían influir en la precisión del diagnóstico.

- **Errores del modelo:**

    - **Falsos Positivos y Falsos Negativos**: En un contexto médico, ambos tipos de errores tienen consecuencias significativas. Un falso positivo puede causar ansiedad y pruebas innecesarias, mientras que un falso negativo podría retrasar un tratamiento necesario.

    - **Sobreajuste y Generalización:** Existe el riesgo de sobreajuste, esto puede afectar su capacidad para generalizar a nuevos datos.

Por tanto, el modelo es objetivamente bueno como herramienta de cribado, teniendo en cuenta que deben de realizarse pruebas adicionales y tomar la decisión siguiendo un criterio médico.

D. (30%) ANÁLISIS DE VARIABLES PREDICTORAS

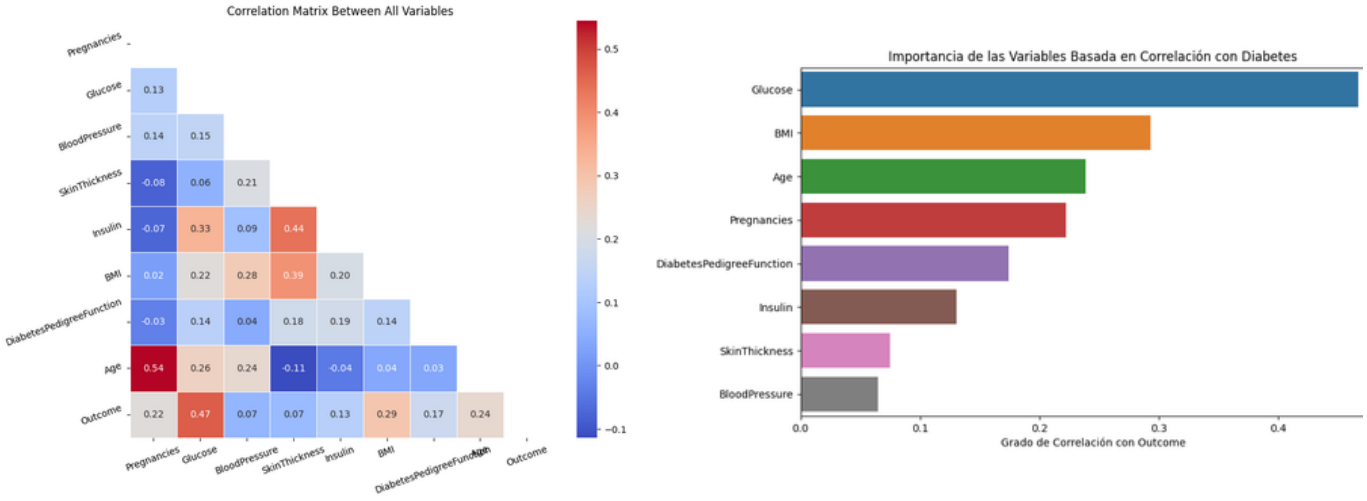
Evaluar la importancia de cada variable en lo que se refiere a su capacidad para predecir la diabetes. Utilizar para ello la matriz de correlaciones. Proporciona el orden de relevancia de mayor a menor en función del grado de correlación.

La matriz de correlaciones es un método aproximado para evaluar la influencia de cada variable, y además tiene la limitación de ser cualitativo, es decir, no permite evaluar la contribución numérica a la predicción. Vamos a emplear un método qué sí permite hacer dicha cuantificación, y que es comúnmente utilizado. Este método se basa en el modelo Random Forest, para el que te proporcionamos el cuaderno de Jupyter con el desarrollo completo: FEAT_BASIC predicting-diabetes.ipynb

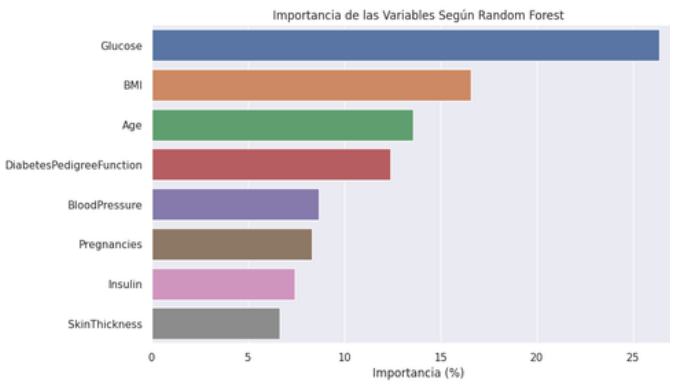
---

Para evaluar las variables se utilizaron los dos métodos comentados, el mapa de correlaciones y el Feature importance de Random Forest.

Matriz de correlaciones:



Para evaluar las variables se utilizaron los dos métodos comentados, el mapa de correlaciones y el Feature importance de Random Forest.



La matriz de correlaciones muestra la relación lineal entre cada variable y el resto de variables, pero no tiene en cuenta la colinealidad, las relaciones no lineales ni los efectos de la interacción entre variables. Sin embargo, la importancia de característica de Random Forest muestran la contribución de cada variable a la capacidad predictiva del modelo, además tiene en cuenta relaciones no lineales y el efecto interacción entre las variables.

- **Compara el orden de variables obtenido con este método en comparación con el orden proporcionado por la matriz de correlaciones.**

Al analizar la relevancia de las variables predictoras en la detección de la diabetes, observamos diferencias entre los resultados arrojados por la matriz de correlaciones y el modelo de Random Forest.

La matriz de correlaciones sugiere que la 'Glucose' tiene la correlación más fuerte con la presencia de diabetes, seguida de 'BMI', 'Age', y 'DiabetesPedigreeFunction'. Por su parte, el modelo de Random Forest confirma la importancia de 'Glucose', BMI y Age.

Asimismo, RandomForest asigna mayor importancia a BloodPressure y a DiabetesPedigreeFunction de lo que lo hace la matriz de correlaciones.

En cuanto a la eliminación de 'Insulin' y 'SkinThickness' del modelo, la decisión se basa en la información obtenida a través de la importancia de características del modelo de Random Forest, así como la existencia de correlació moderada entre Insulin y Glucose, así como de SkinThickness y BMI.

El método de Random Forest proporciona una medida de la importancia de cada característica basándose en cuánto contribuye cada una a la mejora de la precisión del modelo. Si 'Insulin' y 'SkinThickness' muestran una importancia relativa baja, esto sugiere que no aportan significativamente a la capacidad predictiva. Por lo tanto, su eliminación puede simplificar el modelo, incluso mejorar su precisión al reducir la complejidad.

Respecto a las pruebas realizadas eliminando 'SkinThickness' e 'Insulin' en un caso y 'SkinThickness' y 'BloodPressure' en otro, los resultados indican que 'BloodPressure', ofrece una contribución relevante al poder predictivo del modelo, su eliminación resultó en un modelo con mejor rendimiento en comparación con la eliminación de ‘BloodPressure’.

- **¿Tiene validez la comparación de correlaciones para decidir que una variable tiene más influencia que otra en la predicción?**

La correlación puede ser un indicativo inicial de la importancia, pero tiene limitaciones. No captura relaciones no lineales ni la importancia de las interacciones entre variables. Además, una alta correlación no implica causalidad ni una fuerte influencia predictiva de manera aislada.

Si dos variables tienen un nivel similar de correlación con la variable a predecir, pero están fuertemente correlacionadas entre sí, pueden estar proporcionando información redundante al modelo. Esto se conoce como multicolinealidad y puede afectar la interpretación de los coeficientes en modelos de regresión lineal y logística.

En el caso de que las variables sean independientes, cada una aporta información única al modelo, lo cual es preferible.

- **¿Se te ocurre alguna forma alternativa para hacer el estudio de importancia, usando el modelo de Regresión Logística?**

Se me ocurren varias alternativas para lograr esto:

- **Coeficientes estandarizados**: Al tratarse de regresión logística, podemos estandarizar las variables antes de entrenar el modelo, de esta forma, una vez entrenado, un coeficiente más alto en valor absoluto indica un efecto mayor en la probabilidad de tener diabetes.

- **Análisis de sensibilidad:** Podemos cambiar aleatoriamente los valores de una característica y medir como afecta este cambio al modelo. Si la modificación de la característica empeora significativamente el resultado, esto indica que esa variable es importante para el modelo.

- **Lasso**: Esta añade una penalización a los coeficientes del modelo, con ello puede forzar algunos coeficientes a cero, eliminando variables poco importantes.

- **Observar la matriz de confusión**: Si tras eliminar una variable, los falsos positivos o negativos aumentan, esto indica que esa variable, o la interacción de esta con otra de las variables son importantes para detectar estos falsos positivos o negativos.

- **Algoritmos de selección hacia atras y hacia delante**: Podemos usar algoritmos como el FeatureSelector de Scikit learn para ajustar el número de variables, comenzando con todas y eliminandolas una a una o viceversa.