**Tópico:** Aplicación de Modelo de Clasificación de Regresión Logística para el diagnóstico de enfermedades cardiovasculares.

**Dataset:** <https://www.kaggle.com/code/dkurbatovv/heart-disease-regression/input>

**Proceso:**

* **Objetivo de Investigación**

El propósito de esta investigación consiste en analizar el dataset heart disease y crear un modelo predictivo que permita determinar si una persona con una condición de salud específica tiene o no probabilidades de presentar una enfermedad cardiovascular.

* **Descripción de las variables presentes en el Dataset**

El conjunto de datos queda integrado por 319795 filas, contiene 18 columnas con variables comprendidas de 9 booleanas, 5 cadenas y 4 decimales.

* **Enfermedad de Corazón**

La característica "Enfermedad de Corazón" consta de dos valores posibles: "Sí" y "No", representando una variable booleana.

* **IMC**

La variable IMC tiene un total de 3604 valores, proporciona una estimación del nivel de grasa corporal de una persona, los cuales son datos numéricos con decimales.

* **Fuma**

La característica Fumar consta de dos valores posibles, "Sí" y "No", que representando una variable booleana.

* **Bebe Alcohol**

La característica Beber Alcohol consta de dos valores posibles, "Sí" y "No", que representando una variable booleana.

* **Accidente Cerebro Vascular**

La característica Accidente Cerebrovascular consta de dos valores posibles, "Sí" y "No", que representando una variable booleana.

* **Salud Física**

La variable de Salud Física cuenta con un total de 31 valores que representan los últimos 31 días en los cuales tuvo una enfermedad o lesión, los cuales son datos numéricos en formato decimal.

* **Salud Mental**

La variable de Salud Mental cuenta con un total de 31 valores que representan los 31 últimos días en los cuales su salud mental no fue buena, los cuales son datos numéricos en formato decimal.

* **Dificultad para Caminar**

Existen dos valores posibles, "Sí" y "No", que determinan la presencia o ausencia de dificultad para caminar. Estos valores representan una variable booleana.

* **Sexo**

La característica de Sexo se compone de dos valores: "Femenino" y "Masculino", los cuales representan una variable de tipo cadena.

* **Categoría de Edad**

La característica de categoría de edad cuenta con un total de 13 valores diferentes, los cuales son representados como rangos de edades en formato de cadena. Estos rangos son: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 y 80 o más años.

* **Raza**

La característica de raza consta de seis valores diferentes, representados como cadenas de texto. Estos valores son: Indio, Americano/nativo de Alaska, Asiático, Negro, Hispano, Blanco y Otro.

* **Diabético**

La característica "diabético" se compone de cuatro valores posibles: "Sí", "No", "Sí (durante el embarazo)" y "No diabetes límite". Estos valores representan una variable booleana.

* **Actividad Física**

La característica de actividad física cuenta con dos valores posibles: "Sí" y "No". Estos valores representan una variable booleana, indicando la presencia o ausencia de actividad física.

* **Gen Salud**

La característica de GenSalud tiene un total de cinco valores distintos: "Excelente", "Aceptable", "Bueno", "Pobre" y "Muy Bueno". Estos valores se representan como cadenas de texto y describen diferentes niveles de salud general influenciado por el factor genético.

* **Tiempo de Sueño**

La característica de tiempo de sueño consta de 24 valores diferentes los cuales representan las 24 horas del día, los cuales son datos numéricos en formato decimal. Estos valores representan las horas de sueño registradas.

* **Asma**

La característica de Asma se compone de dos valores posibles: "Sí" y "No". Estos valores representan una variable booleana, indicando la presencia o ausencia de asma.

* **Enfermedad del Riñón**

La característica de Enfermedad del Riñón tiene dos valores posibles: "Sí" y "No". Estos valores representan una variable booleana que indica la presencia o ausencia de enfermedad renal.

* **Cáncer de Piel**

La característica de cáncer de piel se compone de dos valores posibles: "Sí" y "No". Estos valores representan una variable booleana que indica la presencia o ausencia de cáncer de piel.

* **Proceso básico de análisis de datos:**

**Preprocesamiento**

La codificación de variable dummy, también conocida como “one-hot enconding” o “binarización”, es una técnica de preprocesamiento utilizada para convertir variables categóricas en variables numéricas.

El proceso de codificación de variables dummy implica crear nuevas columnas binarias (0 o 1) para cada categoría en la variable original. Cada columna representa una categoría y se asigna de 1 si la instancia de datos pertenece a la categoría y 0 en caso contrario.

Visualizamos las variables y el tipo de las mismas las cuales nos determinaran las variables a aplicar la técnica dummy como se muestra en la figura 1.

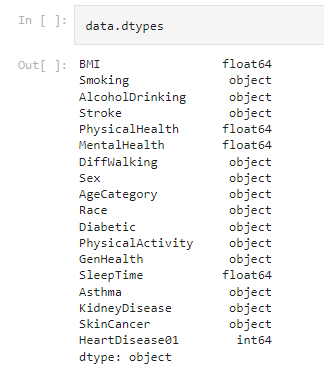


Figura 1: Tipo de variables del dataset heart disease

Aplicamos la técnica a los 9 datos de tipo booleana y 5 de tipo cadena así mismo se procede a la eliminación de las columnas las cuales son afectadas por el preprocesamiento.

Visualizamos los datos como resultado de la aplicación de la técnica que se muestra en la figura 2 a continuación.

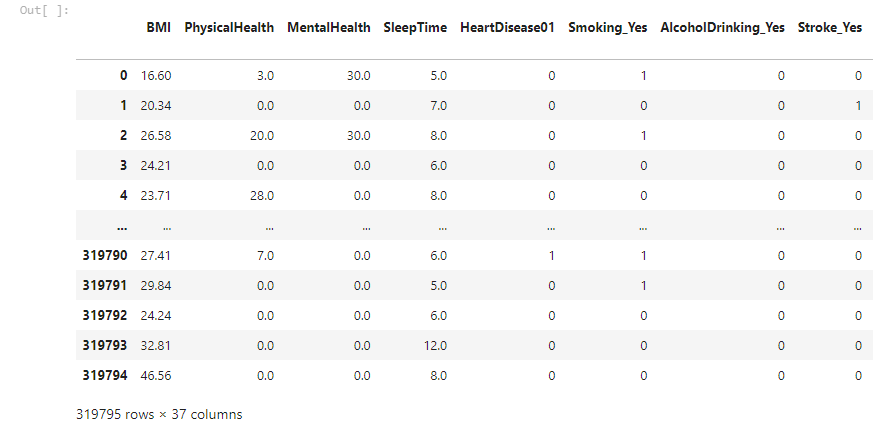


Figura 2: Resultado de la aplicación de la técnica dummy

Visualización de los datos a detalle con las variables y el tipo de dato que se muestra en la figura 3.

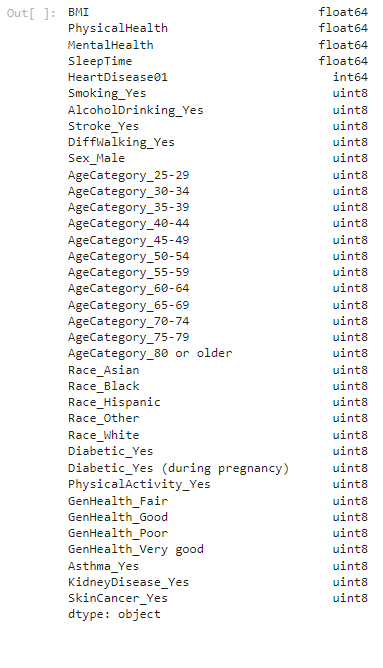


Figura 3: Listado a detalle de las variables resultantes de la técnica de preprocesamiento aplicada

**Selección del Clasificador**

**Regresión Logística**

A diferencia de su nombre, la regresión logística es un modelo de clasificación y no de regresión. Su propósito en el presente trabajo es clasificar nuevos datos utilizando la información del conjunto de datos de enfermedades cardíacas de 2020. En otras palabras, la regresión logística asigna clases a los nuevos datos provenientes de un conjunto finito. Hay dos tipos de regresión logística:

1. Binaria: los datos se clasifican en dos clases.

2. Multinomial: los datos se clasifican en más de dos clases.

En este estudio, se empleará la regresión logística binaria para clasificar los datos en dos categorías: pacientes con enfermedad cardíaca y pacientes sanos.

La regresión logística utiliza una función llamada sigmoide para modelar la probabilidad de una clase. Esta función es similar a la regresión lineal, pero restringe la función de error dentro del rango de 0 a 1.

La forma en la que se define el modelo de regresión logística es:

En nuestro caso, la variable de interés es la probabilidad de que un paciente presente una enfermedad cardíaca, representada por "p". Para facilitar la explicación, asumiremos que esta variable depende de una única variable independiente.

Si se despeja p en la ecuación nos queda de la siguiente manera

Si representáramos la función con valores específicos de β₀ y β₁, obtendríamos una gráfica similar a la mostrada en la Figura 4. En esta gráfica, la interpretación es la siguiente: supongamos que las cruces verdes representan a los pacientes sanos y los rombos rojos representan a los pacientes enfermos. Los pacientes se han ubicado en el eje x de acuerdo con los valores de la variable independiente x para cada uno de ellos. A partir de estos valores, se han calculado los coeficientes β₀ y β₁, y se ha trazado una función sigmoide. Si se nos proporcionan nuevos datos, se representan como círculos azules, simplemente colocándolos en el eje x según su nivel de masa muscular, y se observa el valor que toma la función sigmoide para cada uno de ellos. Esto representa la probabilidad de que el paciente esté sano. Por lo general, se establece un umbral, por ejemplo, t = 0.6, a partir del cual los valores se clasificarán en una clase u otra, es decir, como personas sanas o como personas que están enfermas.

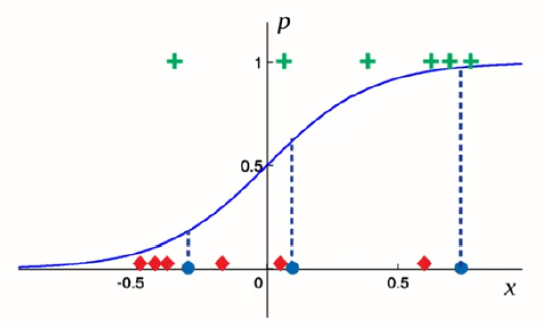


Figura 4: las cruces verdes representan datos pertenecientes a la clase 1 y los rombos rojos a la clase 2. Los círculos azules son nuevos datos.

Evidentemente la clave de la regresión logística es aproximar lo mejor posible los parámetros para que el modelo realice buenas predicciones. La función de perdida de la regresión logística es:

Donde y representa la variable dependiente, es decir las clases del modelo.

Para calcular el mínimo de esta función se utiliza el método del gradiente descendiente que en el fondo es la base de aprendizaje de numerosas técnicas de machine learning.

**División de datos: Entrenar y probar**

X\_entrena, X\_valida, Y\_entrena, Y\_valida = train\_test\_split(datos\_dummis.drop(columns = "HeartDisease01"), datos\_dummis['HeartDisease01'],train\_size = 0.80,  random\_state = 1307)

modelo\_rlog = linear\_model.LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=1000)

modelo\_rlog.fit(X\_entrena, Y\_entrena)

**Evaluación del Modelo**

La evaluación del modelo tiene como objetivo estimar precisión de generalización de un modelo en datos futuros (no vistos / fuera de la muestra).

**Precisión**

La precisión describe la precisión con la que un modelo puede clasificar correctamente. Es el numero de predicciones correctas hechas como una proporción de toda predicción hecha.

print(modelo\_rlog.score(X\_entrena, Y\_entrena))



A partir de ese resultado podemos interpretar si la precisión del modelo que obtenemos es de alrededor del 91, 6 % en el conjunto de validación.

**Matriz de confusión**

La matriz de confusión es una herramienta útil para evaluar un modelo de clasificación. Proporciona información sobre las predicciones realizadas por el modelo para cada clase y las instancias reales clasificadas en cada clase. Una ventaja importante de las matrices de confusión es que permiten identificar fácilmente si el modelo está confundiendo las diferentes clases o generando resultados incorrectos.

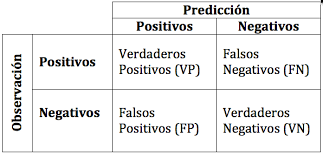


Figura 5: Matriz de Confusión

De la matriz de confusión podemos derivar la precisión que viene dada por la suma de las predicciones corregidas divida por el numero total de predicciones:

Precision = VP + VN / VP +FP + FN + VN

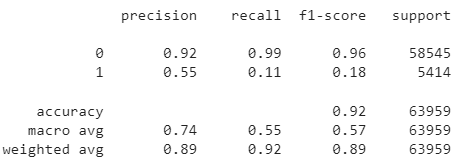
print(confusion\_matrix(comparaciones['HeartDisease\_Real'], comparaciones['HeartDisease\_Pred']))



La matriz de confusión anterior muestra 58074 + 584 = 58658 predicciones correctas y 4830 + 471 = 5301 incorrectas

* Verdaderos positivos (VP): 58074
* Verdaderos negativos (VN): 584
* Falsos positivos (FP): 4830 (error tipo I)
* Falsos negativos (FN): 471 (error tipo II)

print(classification\_report(comparaciones['HeartDisease\_Real'], comparaciones['HeartDisease\_Pred']))



La precisión del modelo = 92%

Precisión o valor pronostico positivo = 92%

Valor predicho negativo = 55%

La sensibilidad o tasa de verdaderos positivos (TPR) = 99%

La especificidad o tasa de verdaderos negativos (TNR) = 11%

Podemos mencionar que los valores positivos se predicen con mayor precisión que los negativos.

**Área bajo la curva (AUC)**

El área bajo la curva ROC cuantifica la precisión de la clasificación del modelo; cuanto mayor sea el área, mayor será la disparidad entre los positivos verdaderos y falsos, y mas fuerte será el modelo en la clasificación de los miembros del conjunto de datos de entrenamiento. Un área de 0,5 corresponde a un modelo que no funciona mejor que la clasificación aleatoria y un buen clasificador se mantiene lo mas alejado posible de eso. Un área de 1 es ideal. Cuanto mas cerca este el AUC de 1 es mejor.

accuracy = accuracy\_score(

    y\_true = comparaciones['HeartDisease\_Real'],

    y\_pred = comparaciones['HeartDisease\_Pred'],

    normalize = True

    )

print(f"El estadístico accuracy es: {accuracy}")



**Referencias**

Berry J, Dyer A, Cai X, Garside D, Ning H, Thomas A, et al. Lifetime Risks of Cardiovascular Disease. N Engl J Med 2012; 366:321-329.

Gupta S, Gudapati R, Gaurav K, Bhise M. Emerging risk factors for cardiovascular diseases: Indian context. Indian Journal of Endocrinology and Metabolism 2013; 17:806-814.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

J. H. . Campo Mendoza, K. D. Parra García, F. Mendoza Palechor, y A. De La Hoz Manotas, «ANÁLISIS DE TÉCNICAS DE MD EN DIAGNÓSTICO DE ENFERMEDADES CARDIOVASCULARES», EIEI ACOFI, jul. 2015.

Rojas, Susan, Querales, Marvin, & Villarino Vivas, Angel. (2016). Evaluación de los factores de riesgo que predisponen a la hipertensión arterial a través de un modelo de regresión logística. Salus, 20(2), 18-23. Recuperado en 31 de mayo de 2023, de <http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-71382016000200005&lng=es&tlng=es>.

Sagaró del Campo, Nelsa María, & Zamora Matamoros, Larisa. (2019). Análisis estadístico implicativo versus Regresión logística binaria para el estudio de la causalidad en salud. Multimed, 23(6), 1416-1440. Recuperado en 29 de mayo de 2023, de <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1028-48182019000601416&lng=es&tlng=es>.

