Primera entrega de proyecto

POR:

Sara Camila Guarín Castillo, Carlos Hernan Molina Bustos, Bryan Zuleta Vélez

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2022

1. Planteamiento del problema

Según la Organización Mundial de la Salud (OMS), más de 750 millones de personas en el mundo padecen algún tipo de patología relacionada con la tiroides y se cree que aproximadamente el 60% lo desconoce.

Se sabe que el funcionamiento inadecuado de la glándula tiroidea puede generar las siguientes patologías: hipotiroidismo, hipertiroidismo, nódulo único sólido o quístico, bocio multinodular, tiroiditis, o cáncer de tiroides. Si el paciente se hace un diagnóstico temprano y conoce su condición, un tratamiento adecuado puede determinar un buen pronóstico en la mayoría de los casos[1]; por tal motivo, se desea desarrollar un modelo que permita calcular la probailidad de que una persona pueda llegar a padecer alguna patología relacionada a la enfermedad de la tiorides (hipertiroidismo o hipotiroidismo) de acuerdo a los siguientes factores de riesgo: edad, sexo y niveles de TSH.

2. Dataset

El dataset a utilizar para el desarrollo del modelo proviene de una competencia de Kaggle, el cual se creó mediante la conciliación de un conjuntos de datos de enfermedades de la tiroides proporcionado por UCI Machine Learning Repository, los cuales fueron extraídos directamente de Garvan Institute of Medical Research. Este dataset consta de un archivo llamado **thyroidDF.csv** el cual proporciona la información personal y médica necesaria para la creación del modelo, dentro de la cual se tiene la edad, el sexo e información relevante sobre la tiroides[2].

La descripción detallada del dataset se muestra a continuación:

- **age** – edad del paciente (int)
- **sex** – sexo del paciente (str)
- **on_thyroxine** – si el paciente está tomando tiroxina (bool)
- **query on thyroxine** - *si el paciente está tomando tiroxina (Consulta) (bool)
- **on antithyroid meds** - si el paciente está tomando medicamentos antitiroideos (bool)
- **sick** - si el paciente está enfermo (bool)
- **pregnant** - si la paciente está embarazada (bool)
- **thyroid_surgery** - si el paciente se ha sometido a una cirugía de tiroides (bool)
- **I131_treatment** - si el paciente está en tratamiento con I131 (bool)
- **query_hypothyroid** - si el paciente cree que tiene hipotiroidismo (bool)
- **query_hyperthyroid** - si el paciente cree que tiene hipertiroidismo (bool)
- **lithium** - si el paciente * litio (bool)
- **goitre** - si el paciente tiene bocio (bool)
- **tumor** - si el paciente tiene tumor (bool)
- **hypopituitary** - si el paciente * glándula hiperpituitaria (float)
- **psych** - si paciente * psych (bool)
- **TSH_measured** - si se midió la TSH en la sangre (bool)
- **TSH** - nivel de TSH en la sangre del análisis de laboratorio (float)

- **T3_measured** - si se midió T3 en la sangre (bool)
- **T3** - Nivel de T3 en la sangre del trabajo de laboratorio (float)
- **TT4_measured** - si se midió TT4 en la sangre (bool)
- **TT4** - Nivel de TT4 en sangre del trabajo de laboratorio (float)
- **T4U_measured** - si se midió T4U en la sangre (bool)
- **T4U** - nivel de T4U en la sangre del análisis de laboratorio (float)
- **FTI_measured** - si FTI se midió en la sangre (bool)
- **FTI** - nivel de FTI en la sangre del análisis de laboratorio (float)
- **TBG_measured** - si se midió TBG en la sangre (bool)
- **TBG** - nivel de TBG en la sangre del análisis de laboratorio (float)
- **referral_source** - (str)
- **target** - diagnóstico médico de hipertiroidismo (str)
- **patient_id** - identificación única del paciente (str)

Como se puede apreciar, el dataset contiene 31 columnas y alrededor de 9172 observaciones, cumpliendo con los requisitos solicitados al escoger el dataset; así mismo, contiene al menos el 10% de columnas categóricas y se simularán aleatoriamente un 5% de datos faltantes en al menos 3 columnas.

3. Métricas

Para la métrica de machine learning se desea implementar *Logloss*, que se calcula mediante la siguiente expresión:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i lnp_i + (1 - y_i)\, ln(1 - p_i)]$$

Donde *N* es el número de observaciones, *i* es la observación dada, *y* es el valor actual positivo, *p* es la probabilidad de predicción y *ln* es el logaritmo natural.

Logloss indica qué tan cerca está la probabilidad de predicción del valor real/verdadero correspondiente. Cuanto más diverge la probabilidad predicha del valor real, mayor será el valor de pérdida logarítmica[3].

4. Desempeño

En cuanto a la métrica de negocio, se tiene como interés que las predicciones sean precisas en la detección de hipotiroidismo e hipertiroidismo en pacientes; para ello se espera que en la evaluación del dataset según la métrica de *Logloss* la pérdida logarítmica sea cercana a 0, de esta manera se puede garantizar que el funcionamiento del modelo es el esperado e identificar qué poblaciones son las que tienen mayor riesgo y así poder comparar estos resultados con los estudios realizados anteriormente[4][5].

5. Bibliografía

[1] de, I. (2022). 25 de mayo I Día Mundial de la Tiroides. Gob.mx. https://www.gob.mx/insabi/articulos/25-de-mayo-i-dia-mundial-de-la-tiroides

[2] Werr, E. F. (2022). Thyroid Disease Data. Kaggle.com. https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data

[3] Gaurav Dembla. (2020, November 17). Intuition behind Log-loss score - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a

[4] Las mujeres son ocho veces más propensas que los hombres de padecer enfermedades en el tiroides | Médicos y Pacientes. (2022). Medicosypacientes.com. http://www.medicosypacientes.com/articulo/las-mujeres-son-ocho-veces-mas-propensas-que-los-hombres-de-padecer-enfermedades-en-el#:~:text=Los%20problemas%20de%20tiroides%20afectan,a%20aumentar%20de%20modo%20gradual

[5] Félix, J., Corrales, B., & Soria, R. (2016). Factores de riesgo de las enfermedades tiroideas. Hospital del Seguro Social Ambato. Revista de Ciencias Médicas de Pinar Del Río, 20(5), 113–128. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-31942016000500014