PROYECTO FINAL: PREDICCIÓN DE CÁNCER TIROIDEO

Proyecto enfocado en predecir la probabilidad de recurrencia en pacientes que han tenido cáncer de tiroides y actualmente se encuentran en remisión.

Samantha Citlally Castro Rojas
Machine Learning, 2025

INTRODUCCIÓN

El presente proyecto tiene como objetivo desarrollar múltiples modelos de Machine Learning y seleccionar uno capaz de predecir la probabilidad de recurrencia en pacientes con antecedentes de cáncer de tiroides por medio de un conjunto de técnicas previstas en clase. Se busca entrenar un algoritmo que detecte patrones complejos y a menudo imperceptibles mediante métodos estadísticos tradicionales. El resultado final pretende ser una herramienta de apoyo para las personas que se encuentren en dicha situación.

Ha de aclararse que dichos resultados pertenecen únicamente a la muestra de la base de datos con la que se trabajó en este proyecto, por lo que los resultados no deben tener una relación total con la teoría medica para ser considerados correctos o de provecho.

PLANTEAMIENTO DEL PROBLEMA

El cáncer de tiroides es un tumor o crecimiento malignizado localizado dentro de la glándula tiroidea y derivado de células tiroideas que pueden ser de dos tipos:

- **Foliculares:** Producen las hormonas tiroideas (T3 y T4) que segregan la proteína tiroglobulina.
- **Células C:** Producen la calcitonina.

(AECAT, s.f.)

Se estima que en 2020 se diagnosticaron 3,379 nuevos casos de cáncer de tiroides en México (2,244 en mujeres y 1,135 en hombres). Las mujeres son 2 veces más propensas a desarrollar este tipo de cáncer que los hombres. La tasa de mortalidad por cáncer de tiroides en México es de 2.9 por cada 100,000 habitantes. La edad promedio de diagnóstico es de 52 años, aunque puede presentarse a cualquier edad (SPLINT, 2024).

El 80% por ciento tiene una sobre vida de hasta 10 años. Hay recurrencia entre el 5 y 20 por ciento. Entre el 10 y el 20 por ciento llega a tener metástasis y sólo entre el 5 y 9 por ciento fallece (Aguilar, 2021).

DESCRIPCIÓN DEL DATASET

El dataset con el que se trabajó para este proyecto contiene 383 pacientes registrados y 17 variables. Los pacientes registrados representan datos de personas que sufrieron de cáncer de tiroides y que actualmente se encuentran en remisión. Por otro lado, las variables son:

Nombre de la Variable	Representación	Tipo de Dato
Age	Edad del paciente.	Numero entero.
Gender	Genero del paciente.	Binario (M/F).
Smoking	¿El paciente fuma?	Binario (si/no).
Hx Smoking	Historial de terapia por enfermedad relacionada a fumar.	Binario (si/no).
Hx Radiotherapy	Historial de radioterapia previa.	Binario (si/no).
Thyroid Function	Función tiroidea (presencia de alguna enfermedad relacionada).	**Eutiroideo:** Niveles normales.
		Hipertiroidismo: Con síntomas.
		Hipertiroidismo clínico: Sin síntomas.
		Hipotiroidismo: Con síntomas.
		Hipotiroidismo clínico: Sin síntomas.
Physical Examination	Cambios físicos por tumores.	Bocio nodular único izquierdo.
		Bocio multinodular.
		Bocio nodular único derecho.
		Normal.
		Bocio difuso.
Adenopathy	Agrandamiento de ganglios linfáticos.	Sin **adenopatía.**
		Adenopatía derecha.
		Adenopatía extensa.
		Adenopatía izquierda.
		Adenopatía bilateral.
		Adenopatía posterior.
Pathology	Tipo de cáncer anterior.	Micro papilar.
		Papilar.
		Folicular.

		Células de Hurthle.
T	Tamaño de tumores.	**T1:** Tumor pequeño. - A. - B.
		T2: Tumor más grande que T1.
		T3: El tumor supero la etapa 2. - A. - B.
		T4: Tumor que invade tejidos u órganos. - A. - B.
N	Clasificación de metástasis en ganglios linfáticos.	**N0:** No hay.
		N1A: Existe en menor medida.
		N1B: Existe en mayor medida.
M	Clasificación de metástasis distante.	**0:** No hubo.
		1: La hay.
Recurred	Probabilidad de cáncer.	Binario (si/no).

Limpieza de datos:

Además, del dataset no cuenta con valores nulos. El balance entre clases representa que un 71% está dedicado a personas que no tienen probabilidad de recurrir al cáncer tiroideo, mientras que el 28% del resto de personas sí, tal como se representa en la *Figura 1*.

En total se encontraron 19 valores duplicados; estos fueron eliminados ya que no eran relevantes para el análisis.

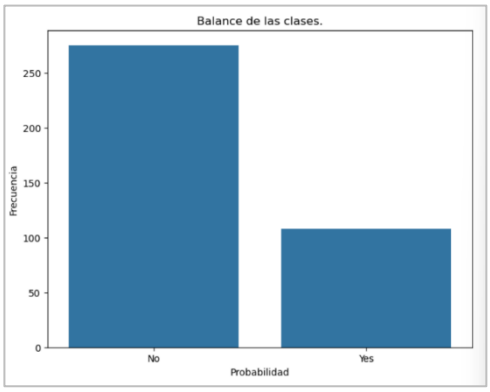

Figura 1: Balance de clases

Al momento de seleccionar las variables con las que se trabajaría, "Response", "Stage" y "Risk" fueron eliminadas; estas variables tienen relación directa con lo que sucede

después de la detección del cáncer, lo que da una respuesta previa al modelo (considerándose trampa):

- **Response:** Es la respuesta del cuerpo al tratamiento para contrarrestar el cáncer.
- **Stage:** Es el estado actual de cáncer en el paciente.
- **Riesgo:** Es un conjunto de "T", "N", "M", lo que declara la presencia de cáncer. Si se trabaja con "T", "N" y "M", la presencia de "Riesgo" es un refuerzo de lo que ya se está analizando.

Posteriormente, se codificaron las variables en los siguientes grupos:

- **Binarias:** "'Gender", "Smoking", "Hx Smoking", "Hx Radiothreapy" y "Recurred". *Variables que solo se catalogan como si o no o como dos valores únicos.*
- **Ordinales:** "T", "N", "M". *Variables que tienen un orden jerárquico importante.*
- **Nominales:** "Thyroid Function", "Physical Examination", "Adenopathy", "Pathology", "Focality". *Variables que no tienen un orden jerárquico, pero si varias opciones.*

Finalmente, realice unas graficas para entender el comportamiento de la muestra con la que estamos trabajando, concluyendo que:

Datos binarios: El dataset cuenta con mayor cantidad de mujeres, personas que no fuman y por ende, no han recibido tratamiento derivado a ello, que no han tenido un tratamiento de radioterapia previo y que no cuentan con probabilidad de volver a sufrir de cáncer tiroideo, tal como se muestra en la *Figura 2*.

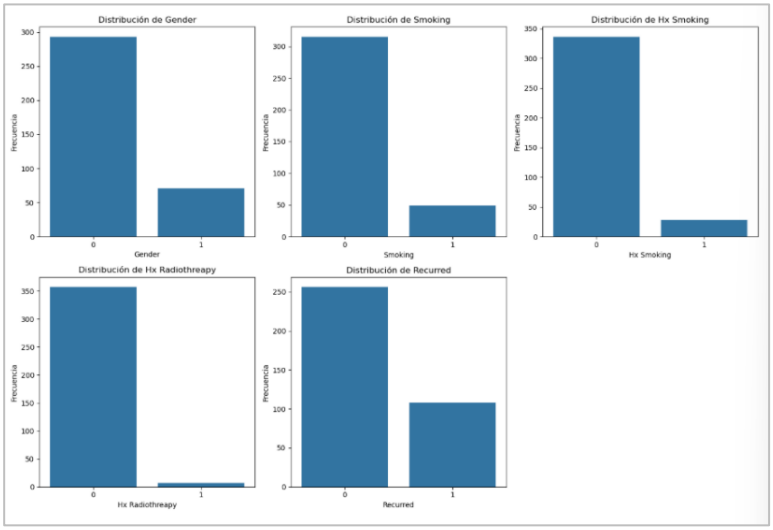

Figura 2: Variables binarias.

Datos ordinales: En nuestro dataset hay más personas que sufrieron tumores de grado "T2" (tumor localizado en el órgano de origen) y que no resultaron en metástasis ganglionar, por ende, personas que no sufrieron de metástasis distante, tal como se muestra en la *Figura 3*.

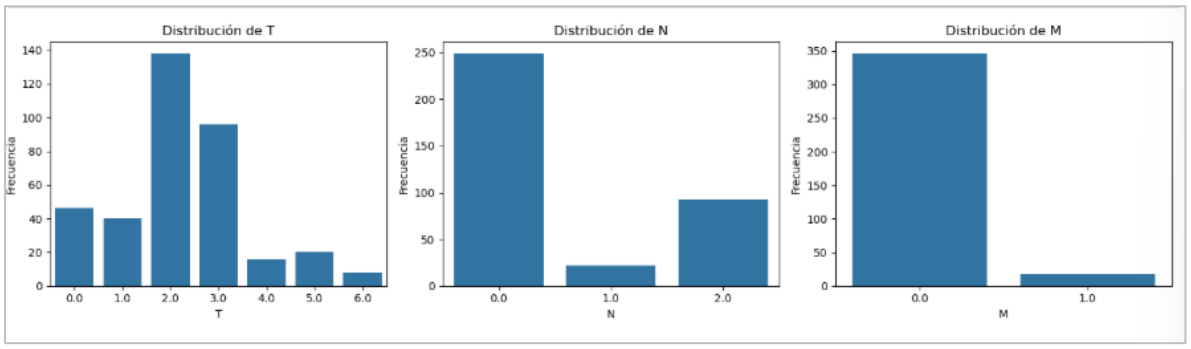

Figura 3: Datos ordinales

Datos nominales: Nuestro dataset cuenta con más personas que mantienen un estado controlado de la hormona tiroidea, sin embargo, presentan un agrandamiento de bocio, aunque no cuentan con glándulas cancerosas en los ganglios linfáticos. El cáncer previo se encontró en las células foliculares de la glándula tiroidea y el tumor se generó solo en un órgano (también, glándula tiroidea)., tal como se muestra en la *Figura 4*.

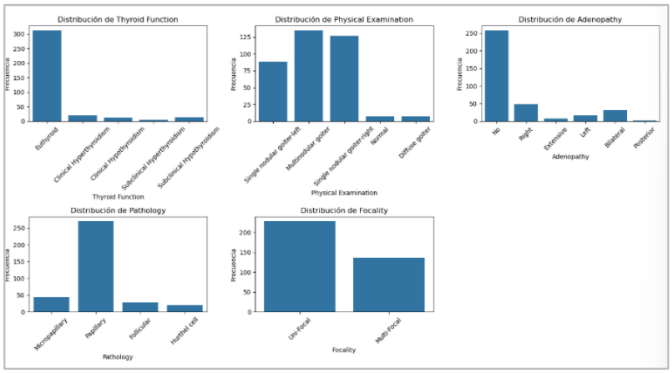

Figura 4: Datos nominales

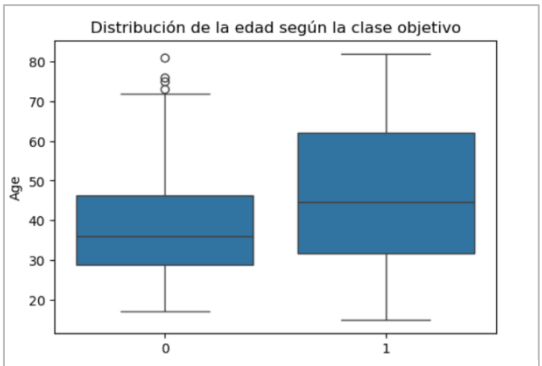

Figura 5: Comparación de Age y Recurred.

Un análisis comparativo entre la variable objetivo y la variable “age” muestra que las personas con probabilidad de sufrir cáncer nuevamente se encuentran en un rango de entre 32 y 63 años, mientras que las personas que no tienen esta probabilidad se encuentran en un rango de entre 30 y 45 años, como se muestra en la *Figura 5*.

Finalmente, con esta información y en base a la muestra con la que se realizó el análisis, se puede concluir que la probabilidad de ser positivo a caer en recurrencia es:

- Ser hombre.
- Fumar.
- Haber recibido tratamiento previo por alguna enfermedad derivada de fumar.
- Haber recibido radioterapia en el cáncer anterior.
- Haber tenido tumores de grado avanzado.
- Sufrir de hipotiroidismo clínico.
- Agrandamiento de bocio.
- Que el cáncer haya sido multifocal.

METODOLOGIA

Para realizar el análisis se hizo uso de Python y librerías para la gestión de los datos, el procesamiento y medición de los modelos y la generación de estos. Además, se separaron los datos en distintos grupos para proceder con el entrenamiento del modelo:

- **X:** Todas las variables (menos la variable objetivo).
- **Y:** Variable objetivo.
- **x_train, x_test, y_train, y_test:** Prueba y entrenamiento.
- **x_train_escalado, x_test_escalado:** Prueba y entrenamiento escalado.

Se selecciono un problema de **clasificación** ya que el propósito del proyecto es definir si existe una probabilidad de sufrir recurrencia al cáncer o no. Por otro lado, los modelos seleccionados fueron: **Regresión Logística, Random Forest, Support Vector Machine y Naive Bayes.** Estos modelos se seleccionaron gracias a que son los que mejor trabajan con datasets categóricos, tal como el que se tiene para este proyecto.

RESULTADOS

Regresión Logística:

Se seleccionaron un total de 1000 iteraciones para el entrenamiento de este modelo, para el cual se utilizaron grupos de datos ya escalados. Al hacer una comprobación de coeficientes, se puede notar que la variable con mayor impacto es “T”, algo lógico, ya que es el indicador de los tumores que se tuvieron de manera previa y que pueden ser un gran indicador de tumores futuros.

Este modelo da una precisión de 89.04% al momento de hacer la predicción, mientras que un 94.83% pertenece a la clasificación que hace el modelo. La matriz de confusión muestra que tan solo 8 datos no fueron clasificados de manera correcta, indicando un modelo muy bien entrenado. Además, el 77.27% de los casos reales fueron detectados correctamente. Finalmente, es el modelo más rápido.

Random Forest:

Se seleccionaron un total de 50 árboles, decisión que se tomo en base a una prueba previa que mostraba que, a partir de esta cantidad, la variación de la precisión es mínima. Para ese modelo, la variable que tiene más importancia es que la persona no haya sufrido de una adenopatía.

Este modelo da una precisión de 89.04% al momento de hacer la predicción, mientras que un 93.93% pertenece a la clasificación que hace el modelo. La matriz de confusión muestra que, igual que el modelo anterior, tan solo 8 datos no fueron clasificados de manera correcta, indicando un modelo que también está muy bien entrenado. Además, el 72.72% de los casos reales fueron detectados correctamente.

SVM:

Para este modelo, se seleccionó un valor de c de 1, decisión que se tomó en base a una prueba previa que mostraba que, a partir de esta cantidad, la variación de la precisión es mínima.

Este modelo da una precisión de 89.04% al momento de hacer la predicción (igual que los dos modelos anteriores), mientras que un 93.13% pertenece a la clasificación que hace el modelo. La matriz de confusión muestra que, igual que los modelos anteriores, tan solo 8 datos no fueron clasificados de manera correcta, indicando un modelo que también está muy bien entrenado. Además, el 72.72% de los casos reales fueron detectados correctamente.

Naive Bayes:

Para este modelo, se selecciono el tipo *multinomial* ya que se esta trabajando con una gran cantidad de variables categóricas, por lo que resulta perfecto para este caso.

Este modelo da una precisión de 86.30% al momento de hacer la predicción (el menor de los dos modelos anteriores), mientras que un 90.99% pertenece a la clasificación que hace el modelo. La matriz de confusión muestra que 10 datos no fueron clasificados de manera correcta, indicando un modelo que no resulta tan viable. Además, el 72.27% de los casos reales fueron detectados correctamente, siendo este uno de los modelos más rápidos al igual que regresión logística.

DISCUSIÓN

Dados nuestros resultados, podemos concluir que el mejor modelo es el de **Regresión Logística,** ya que tiene una precisión buena al momento de predecir y una precisión de clasificación igual de buena que asegura que la mayoría de los datos serán clasificados donde deben ir. Si se agregan más registros, la clasificación sin duda mejoraría. Además, es bastante rápido, lo que lo sitúa por encima del resto de los modelos.

Formulación del modelo:

Este modelo realmente es de clasificación probabilística, no de regresión. El núcleo de la regresión logística es transformar una regresión lineal estándar para que su salida se encuentre entre 0 y 1, lo cual al final es interpretado como una probabilidad. En lugar de declarar una hipótesis nula, se declara una hipótesis con sigmoide (que funciona de manera logística): en regresión logística, el sigmoide es la función matemática que transforma la salida en un rango de 0 a 1, tal como lo que se busca en un inicio.

Función de costo:

En este modelo en específico no podemos hacer uso del Error Cuadrático Medio (MSE) porque, al introducir la sigmoide no lineal, la función de costo se volvería "no convexa", lo que significa que tendría varios mínimos locales, lo que dificultaría la optimización del modelo. Es por eso que en su lugar, se recomienda la Entropía Cruzada o Log Loss; Esta función es convexa, lo que garantiza que el algoritmo de optimización encuentre el mínimo global. Esta medida lo que nos da es un resultado de que tan bien o mal esta funcionando el modelo al momento de predecir la clase correcta.

Algoritmo de aprendizaje:

Cuando se trata de minimizar, el método fundamental es el Descenso de Gradiente; este es un algoritmo de optimización que permite encontrar los parámetros como pesos y/o sesgos óptimos que minimizan la función de perdida o error en el modelo. De esta manera el modelo puede aprender de sus errores. Para esto, el gradiente será un vector que apunta en la dirección de máximo ascenso en la función de perdida.

Propiedades teóricas relevantes:

Algunas de sus propiedades son:

- Es un modelo generalizado, lo que significa que es un modelo que encuentra la relación entre las variables predictoras y su respuesta categórica.
- Función Logística: El modelo utiliza la función logística para transformar el valor resultante en un valor entre 0 y 1.
- Variables categóricas: El modelo esta pensado para variables categóricas.
- No tiene supuestos de normalidad.
- Tiene baja multicolinealidad.

Ventajas y limitaciones matemáticas:

Sus ventajas son:

1. El resultado es directamente una probabilidad de entre 0 y 1, lo que hace que su interpretación sea más sencilla.
2. Es un modelo bastante eficiente aun con una gran cantidad de datos, teniendo rapidez al momento de calcular.
3. El modelo puede ser regularizado fácilmente.

Sus limitaciones son:

1. Existe un supuesto de linealidad logit, lo que asume una relación lineal entre las variables predictoras y el logaritmo del resultado.
2. Necesita un tamaño de muestra grande para que sus predicciones sean estables.
3. Es más útil en datos categóricos o discretos, no en valores continuos.

CONCLUSIONES

Todos los modelos pueden generar una buena predicción, pero esto no significa que ser un buen modelo: lo que es realmente importante es reconocer si esta predicción viene de una buena clasificación. El modelo de regresión logística es un modelo completamente útil y rápido que permite una predicción alta con variables categóricas, excelente para el problema que se busco resolver en este proyecto. La cantidad de valores predichos y clasificados es buena, asegurando un buen tratamiento de la información.

Sin duda un proyecto bastante interesante y que aún puede ser mejorado de muchas maneras.

TRABAJO FUTURO

En un futuro, lo ideal seria agregar mas registros al dataset para que de esta manera, el modelo tenga aún más información de donde aprender. Además, este modelo una vez perfeccionado puede ser de gran utilidad en distintos hospitales oncológicos de México para ayudar a una detección temprana.

REFERENCIAS

AECAT. (s.f.). El Cáncer de Tiroides. Obtenido de AECAT: https://www.aecat.net/el-cancer-de-tiroides/sobre-el-cancer-de-tiroides/

Aguilar, F. G. (9 de Diciembre de 2021). *Gaceta UNAM*. Obtenido de Tiroides: las señales de un cáncer que padecen más las mujeres: https://www.gaceta.unam.mx/tiroides-las-senales-de-un-cancer-que-padecen-mas-las-mujeres/

SPLINT. (16 de Abril de 2024). *SPLINT*. Obtenido de Cáncer de tiroides en México: Un enemigo silencioso en aumento: https://splint.mx/publicaciones/cancer-de-tiroides-en-mexico-un-enemigo-silencioso-en-aumento/#:~:text=Se%20estima%20que%20en%202020,puede%20presentarse%20a%20cualquier%20edad.