IES COMERCIO

TRABAJO FINAL

Sistemas de Aprendizaje Automático

Andrei Alexandru Miu

2025

Trabajo de clase

Contenido

Acerca del Dataset 3
Algoritmos usando BigML 3
Matriz de confusión 5
Cálculo de métricas de evaluación y cuartiles 5
Gráficas con Excel 7
Procesamiento de Lenguaje Natural 9
Bibliografía 11

Acerca del Dataset

Para el trabajo final se ha escogido un conjunto de datos sobre la recurrencia del cáncer de tiroides. Lo puedes encontrar pinchando aquí. En el Excel se encuentran:

- Total de filas: 383
- Total de columnas: 13
- No hay valores faltantes/vacíos, ni atípicos.

Columnas y descripción de las mismas:

- Age: Edad del paciente (en años).
- Gender: Género del paciente (Masculino o Femenino).
- Hx Radiotherapy: Antecedentes de radioterapia (Sí o No).
- Adenopathy: Presencia de afectación ganglionar (Sí o No).
- Pathology: Tipo de cáncer de tiroides (por ejemplo, micropapilar).
- Focality: Focalidad del tumor (Unifocal o Multifocal).
- Risk: Clasificación del riesgo (Bajo, Intermedio, Alto).
- T: Clasificación del tumor primario (T1, T2, etc.).
- N: Clasificación de ganglios linfáticos (N0, N1, etc.).
- M: Clasificación de metástasis a distancia (M0, M1, etc.).
- Stage: Estadio del cáncer (Estadio I, II, III, IV).
- Response: Respuesta al tratamiento (Excelente, Indeterminada, etc.).
- Recurred: Si el cáncer reapareció (Sí o No).

La única columna numérica es la de edad (age), el resto son categóricas.

- Variables numéricas: Valores que representan cantidades o valores medibles.
- Variables categóricas: Valores que representan categorías o grupos.

Adicionalmente, en el Excel se han creado 2 macros para poder ordenar por columnas y de orden ascendente o descendente más rápido, permitiendo así visualizar antes los datos.

Algoritmos usando BigML

Dentro del propio CSV he dejado dividido por hojas con el nombre del algoritmo utilizado. A continuación, dejo la explicación de cada uno de ellos:

Supervisados:

- Model: Se emplea para predecir una variable de salida a partir de datos de entrada etiquetados (donde sabes el resultado).
    - Resultado: Se ha obtenido un 92,30% sobre la columna “Recurred”, donde indica si el cáncer reapareció, o no, algo importante de cara a

saber si un paciente volverá a recaer. Se ha usado este algoritmo porque es una variable categórica y al ser un árbol de decisión lo maneja mejor.

- Ensemble: Se utiliza para obtener un mejor rendimiento al combinar varios modelos.
    - Resultado: Se ha obtenido un 71.79% sobre la columna "Response", la cual indica la respuesta al tratamiento del paciente. Se ha utilizado este algoritmo ya que queremos que sea preciso, y es más preciso si se usan varios modelos que uno solo.
- Linear regression: Se usa cuando la variable de salida es numérica y una relación con las características de entrada.
    - Resultado: Se ha utilizado este modelo para predecir el % de éxito del futuro de la variable calculada, sin embargo no a sido posible realizar ningún experimento debido a la falta de variables numéricas (solo tenemos edad).
- Logistic regression: Se usa cuando necesitas predecir una variable categórica en 2 clases.
    - Resultado: Se ha obtenido un 74.35%. Se ha usado este algoritmo por que funciona mejor con variables que son Si/No o similares.
- Deepnet: Se usa con grandes volúmenes de datos y problemas complejos.
    - Resultado: Se ha obtenido un 91.02%. Se ha utilizado este algoritmo sobre la variable "Recurred" ya que es un patrón difícil de detectar, algo que encaja con el algoritmo.
- Time series: Se usa cuando tienes datos que dependen del tiempo, y quieres predecir el futuro basándote en el pasado.
    - Resultado: Como no tengo variables que dependen del tiempo, es inútil usar este algoritmo.
- Optiml: Es útil cuando no sabes que algoritmo usar o los parámetros para obtener un mejor resultado.
    - Resultado: Se ha obtenido un 85.89 %. Se ha utilizado este algoritmo sobre la variable "Patologhy", ya que este algoritmo encuentra la mejor combinación para la variable.

No supervisados:

- Cluster: Se usa cuando no hay etiquetas y quieres agrupar datos similares.
    - Resultado: Se han obtenido 2 clústeres (0,1), donde vemos que el clúster 1 es más grande que el 0. Se ha utilizado este algoritmo sobre la variable "Focality", ya que queremos agrupar pacientes similares con un mismo problema.

- Anomaly: Se usa para detectar anomalías (fraudes o fallos en el sistema).
    - Resultado: Se ha utilizado este algoritmo sobre la variable "Hx Radiothreapy", ya queremos encontrar casos donde se haya podido diagnosticar mal a un paciente.
- Association: Se usa cuando quieres encontrar patrones.
    - Resultado: Se ha utilizado este algoritmo sobre la variable "T", ya queremos encontrar relaciones entre posibles columnas como N, M o el estado, y el algoritmo es para eso precisamente.
- PCA: Se usa cuando tienes muchas variables y quieres reducir la complejidad.
    - Resultado: Se ha utilizado este algoritmo sobre la variable "T", ya queremos simplificar la información, visualizando así mejor los datos.

Matriz de confusión

La matriz de confusión sirve para mostrar de forma explícita cuándo una clase es confundida con otra, lo cual nos permite trabajar de forma separada con distintos tipos de error. En este caso, haremos una matriz de confusión de Genero y Edad, donde dividiremos la edad en 2 grupos:

- Edad <= 40 -> Grupo Joven
- Edad > 40 -> Grupo Mayor

Pudiendo así comparar la columna de genero con edad de forma binaria.

		Edad	
		Grupo Joven (Positivo)	Grupo Mayor (Negativo)
Género	Hombre	VP = 30	FN = 41
	Mujer	FP = 190	VN = 122

Interpretación de la tabla:

- **VP (Verdaderos Positivos):** Hombres jóvenes (<= 40 años).
- **FN (Falsos Negativos):** Hombres mayores (> 40 años).
- **FP (Falsos Positivos):** Mujeres jóvenes (<= 40 años).
- **VN (Verdaderos Negativos):** Mujeres mayores (> 40 años).

Cálculo de métricas de evaluación y cuartiles

Cálculo de métricas

Exactitud (Accuracy):

(VP + VN) / (VP + VN + FN + FP) = (30 + 122) / (30 + 122 + 41 + 190) = 39.63%

Es el porcentaje de predicciones correctas. En este caso un 00.00% no es que sea un modelo muy fiable.

Precisión:

VP / (VP + FP) = 30 / (30 + 190) = 13.64%

De todas las veces que el modelo predijo la clase positiva, el 13.64% fueron correctas.

Sensibilidad (Recall):

VP / (VP + FN) =30 / (30 + 41) = 42.25%

De todos los casos realmente positivos, el modelo detectó correctamente el 42.25%

Especificidad:

VN / (VN + FP) = 122 / (122 + 190) = 39.10%

Indica qué tan bien el modelo detecta los negativos. En este caso un 39.10% es que hay muchos falsos positivos

F1-Score:

2 x ((Precisión y Recall) / (Precisión+Recall)) = 20.61%

Esta métrica es útil cuando hay clases desbalanceadas o cuando queremos equilibrar precisión y recall.

Tasa de Error:

(FN + FP) / (VP + VN + FN + FP) = (41 + 190) / (30 + 122 + 41 + 190) = 60.37%

Porcentaje de predicciones incorrectas.

Prevalencia:

(VP + FN) / (VP + VN + FN + FP) = (30 + 41) / (30 + 122 + 41 + 190) = 18.54%

Proporción de casos positivos reales en el dataset.

Índice de Jaccard:

(VP + VN) / ((VP + VN + FN + FP) + (FN + FP)) = (30 + 122) / ((30 + 122 + 41 + 190) + (41 + 190)) = 24.76%

Es similar al F1-Score pero penaliza más los errores.

Tasa de Falsos Positivos (FPR):

FP / (VN + FP) = 190 / (190 + 122) = 60.90%

Indica el porcentaje de negativos mal clasificados como positivos. En este caso es un 60.90%, un porcentaje bastante alto.

Tasa de Falsos Negativos (FNR):

FN / (VP + FN) = 41 / (41 + 30) = 57.75%

Indica el porcentaje de positivos mal clasificados como negativos, el cual también es alto

Cálculo de cuartiles

Calculo hecho sobre la columna edad (age)

Q1: Para encontrar Q1, se toma el punto de datos en el 25%.

- Q1 = 15

Q2: También conocido como mediana, ya que es el 50%.

- Q2 = 37

Q3: Es el punto del 75% de los datos.

- Q3 = 51

Q4: Es el 100% de los datos.

- Q4 = 82

IQR: El rango intercuartil es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Mide la difusión del 50% medio de los datos. Para calcularlo:

- IQR = Q3 – Q1 = 51 – 15 = 36

Límite inferior: Valor usado para detectar valores atípicos por debajo de lo esperado. Para calcularlo:

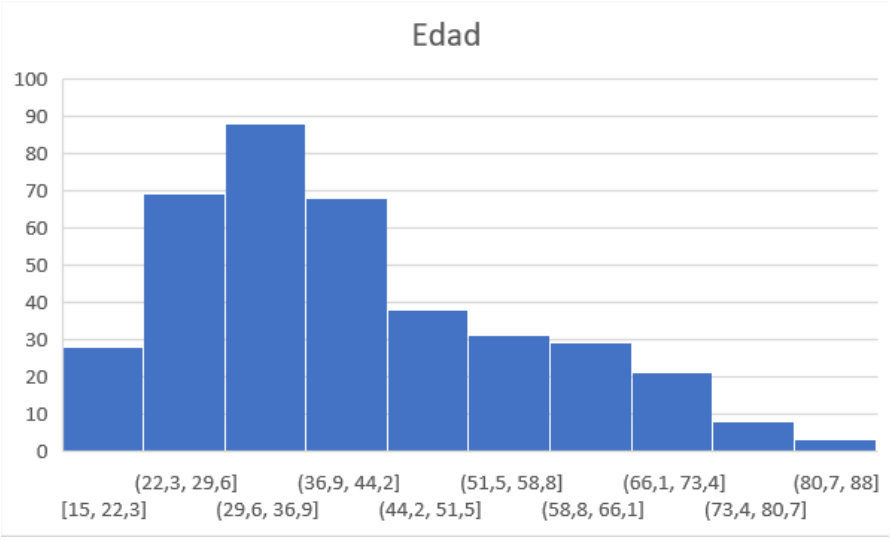
- Limite inferior = Q1 – 1.5 * IQR = 15 – 1.5 * 36 = -39

Límite superior: Valor usado para detectar valores atípicos por encima de lo esperado. Para calcularlo:

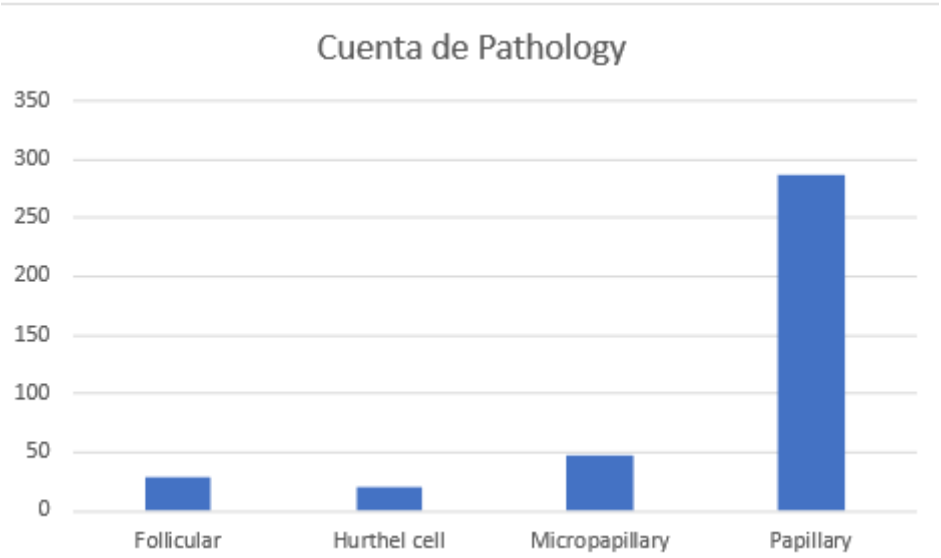
- Limite superior = Q3 + 1.5 * IQR = 51 + 1.5 * 36 = 105

Gráficas con Excel

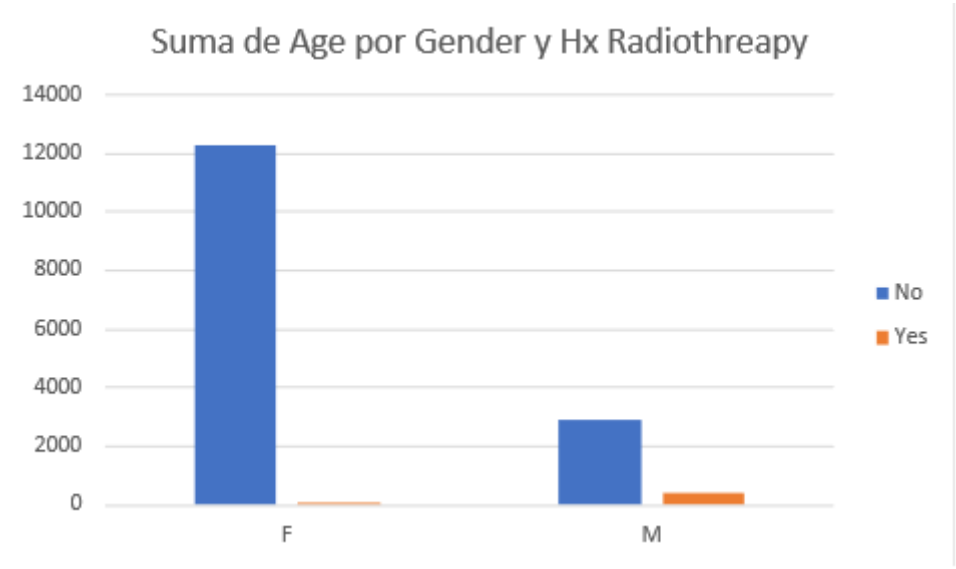
Se han realizado un par de gráficos en Excel, tanto para la búsqueda de posibles valores atípicos, como para visualizar los datos. A continuación dejo un par de imágenes con las explicaciones de cada una.



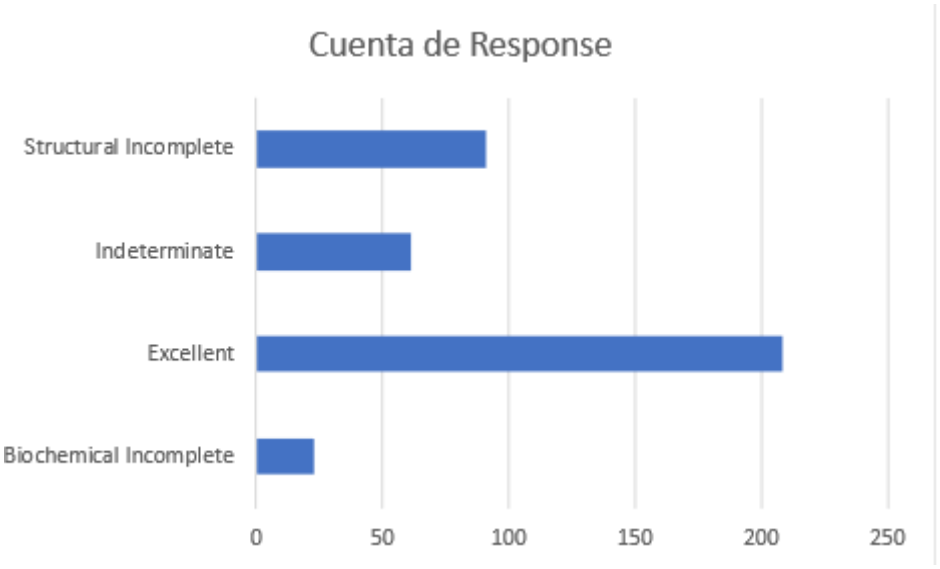
En este histograma podemos ver la distribución de edad del dataset. Vemos que se concentra mucho en un rango de 22.3 a 44.2 años de edad.



En este gráfico de columna, podemos ver los tipos de patologías más comunes, siendo así que el más común es el papilar.



En este gráfico de columna podemos ver el género dividido en 2 y la suma total de si han tenido o no antecedentes de radioterapia.



En este gráfico de barras podemos observar que la respuesta al tratamiento por lo general suele ser excelente.

Al no haber muchas variables numéricas no puedo sacar más tipos de gráficos como podría ser uno de dispersión o de caja y bigotes.

Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural (NLP) es una tecnología de machine learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano.

- Por ejemplo: Utilizan software de NLP para procesar de forma automática estos datos, analizan la intención y responden en tiempo real a la comunicación humana.

Casos de uso en ámbito empresarial:

- Eliminación de información confidencial
- Interacción con clientes
- Análisis empresarial

¿Cómo funciona NLP?

Combina modelos de lingüística computacional, machine learning y aprendizaje profundo para procesar el lenguaje humano.

- Lingüística computacional: Ciencia de entender y crear modelos de lenguaje humano con computadoras y herramientas de software.
- Machine learning: Tecnología que entrena a una computadora con datos de muestra para mejorar su eficiencia.
- Aprendizaje profundo: Campo específico del machine learning que enseña a las computadoras a aprender y pensar como humanos.
- Pasos de la implementación del NLP: Recopilación y preparación de datos
- Preprocesamiento: Uso de distintas técnicas (como la creación de tokens, derivación, lematización y eliminación) para preparar los datos.
- Capacitación: Se utilizan los datos preprocesados y el machine learning para entrenar modelos NLP.
- Despliegue e inferencia: El modelo de NLP recibe entradas y predice un resultado para el caso de uso específico para el que está diseñado el modelo.

Que son las tareas de NLP: Dividen el texto en partes más pequeñas, por ejemplo:

- Etiquetado de parte de la voz
- Desambiguación del sentido de las palabras
- Reconocimiento de voz
- Traducción automática
- Reconocimiento de entidades con nombre
- Análisis de opiniones

Enfoques procesamiento de lenguaje natural:

- NLP supervisado
- NLP no supervisado
- Comprensión de lenguaje natural
- Generación de lenguaje natural

Bibliografía

https://profesordata.com/2020/08/07/evaluando-los-modelos-de-clasificacion-en-aprendizaje-automatico-la-matriz-de-confusion-claramente-explicada/

https://telefonicatech.com/blog/como-interpretar-la-matriz-de-confusion-ejemplo-practico

https://escueladedatos.online/tutorial/estadistica-avanzada-detectando-valores-atipicos-y-datos-inconsistentes/

https://fastercapital.com/es/contenido/Metodo-de-cuartil--analisis-de-datos-mediante-actualizacion-de-cuatro-partes-iguales.html

https://aws.amazon.com/es/what-is/nlp/

https://www.ibm.com/think/topics/machine-learning

https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

https://es.slideshare.net/slideshow/introduccin-al-machine-learning-con-bigml/76397091

https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/