

Maestría en ciencia de datos

Aprendizaje automático

Nelson Alfonso Beas Ham - 1942687

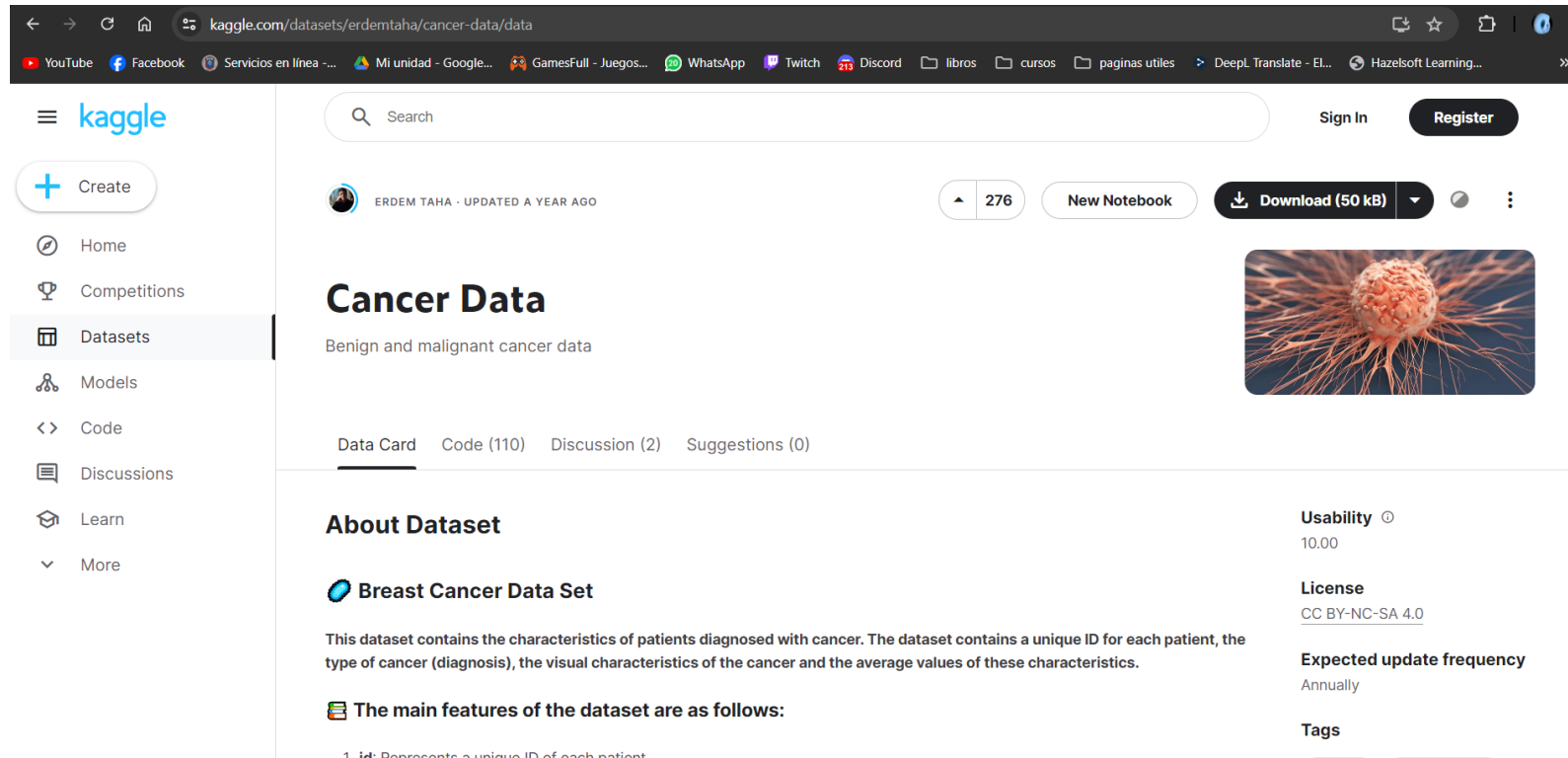


Introducción

Los datos seleccionados provienen de registros médicos que describen medidas físicas de tumores en 30 aspectos numéricos, además de una etiqueta clasificatoria que separa los registros en dos conjuntos, benignos y malignos.



Origen de los datos



The screenshot shows the Kaggle website interface. The browser address bar displays the URL: [kaggle.com/datasets/erdemtaha/cancer-data/data](https://www.kaggle.com/datasets/erdemtaha/cancer-data/data). The left sidebar contains navigation links: Create, Home, Competitions, Datasets (selected), Models, Code, Discussions, Learn, and More. The main content area features a search bar, 'Sign In' and 'Register' buttons, and a dataset card for 'Cancer Data' by ERDEM TAHA, updated a year ago. The card shows 276 views and a 'Download (50 kB)' button. The dataset title 'Cancer Data' is prominently displayed, followed by the description 'Benign and malignant cancer data'. Below this, tabs for 'Data Card', 'Code (110)', 'Discussion (2)', and 'Suggestions (0)' are visible. The 'About Dataset' section includes the title 'Breast Cancer Data Set' and a description: 'This dataset contains the characteristics of patients diagnosed with cancer. The dataset contains a unique ID for each patient, the type of cancer (diagnosis), the visual characteristics of the cancer and the average values of these characteristics.' It also lists the main features: 'The main features of the dataset are as follows:'. On the right, a 'Usability' score of 10.00, a 'License' of CC BY-NC-SA 4.0, and an 'Expected update frequency' of Annually are shown. A 'Tags' section is also present.

- <https://www.kaggle.com/datasets/erdemtaha/cancer-data/data>



Preprocesamiento

Tendremos tres diferentes preprocesamientos para cada algoritmo:

1° Algoritmo supervisado

2° Algoritmo no supervisado

3° grafo (árbol)



Preprocesamiento

Supervisado:

Se tomarán las 31 características independientes y se realizará un PCA para realizar una imagen de 4x4 de las 16 características más representativas de la etiqueta.

No supervisado:

Se retira la etiqueta y se acota a 8 variables con PCA.

Árbol:

Se poda el árbol a 16 variables.



Metodología

La etiqueta generada para cada registro será el resultado de calcular la moda entre el conjunto de etiquetas generadas individualmente por algoritmos descritos a continuación.



Algoritmo supervisado:

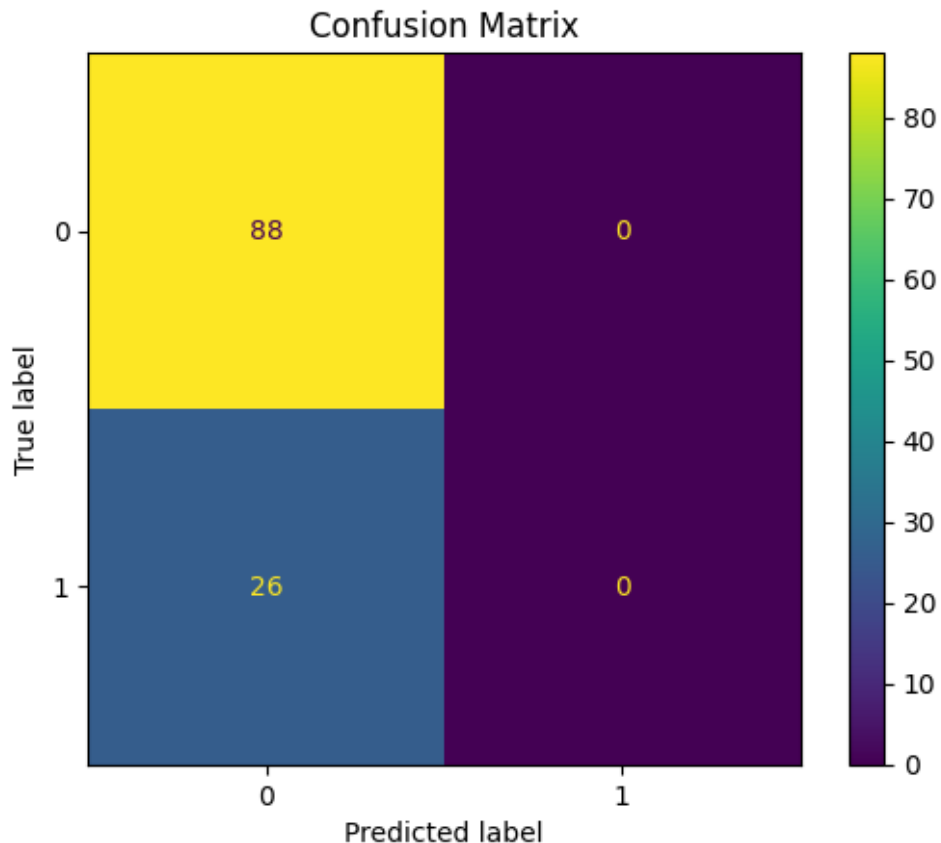
El modelo se compondrá de una red neuronal convolucional con las siguientes características:

- De forma secuencial, la red está formada por una capa convolucional de entrada $4 \times 4 \times 1$, la cual cuenta con 4 núcleos de 2×2 , activada mediante la función relu.
- Posteriormente, se encuentra una capa de maxpool con dimensión 2×2 .
- Una capa de reducción de dimensiones 2D - 1D
- Una capa densa de 8 neuronas con activación relu.
- Una capa densa de 1 neurona con activación softmax.



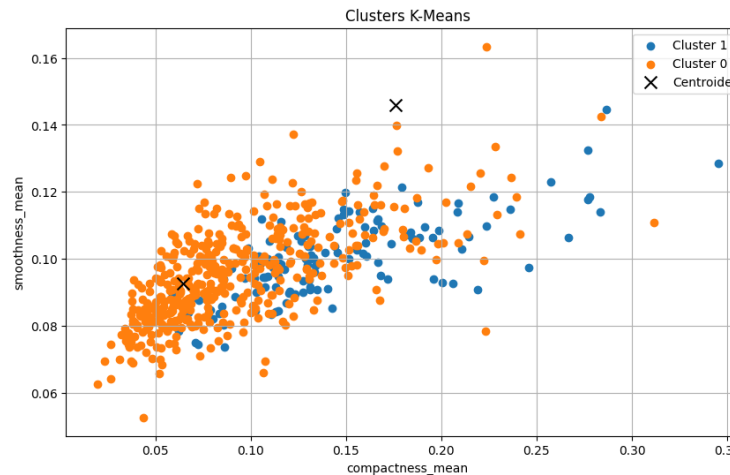
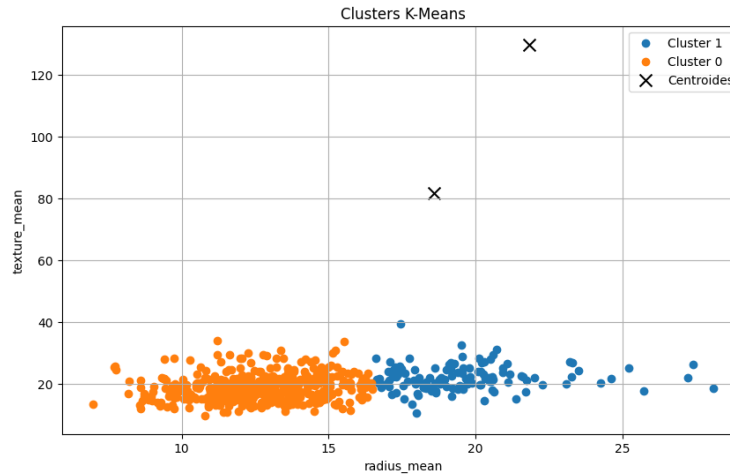
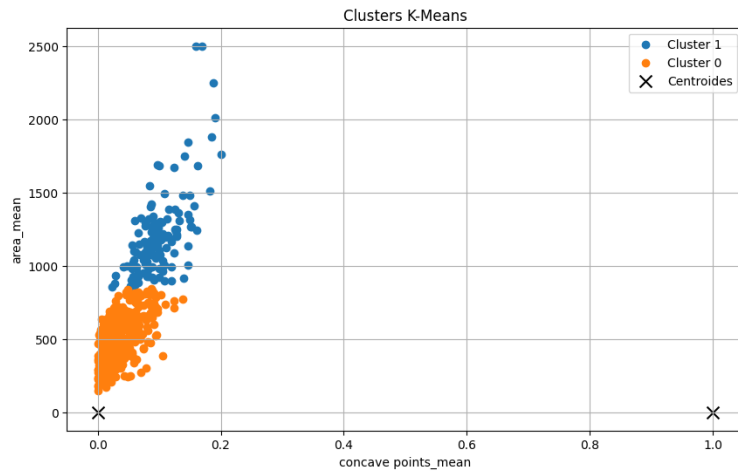
Resultados

Para el algoritmo supervisado, tuvimos un sesgo muy marcado hacia los resultados con la etiqueta B, la cual es mapeada como un 0 para el algoritmo, esto puede deberse al tamaño reducido de la imagen formada o a el tamaño de la red neuronal.



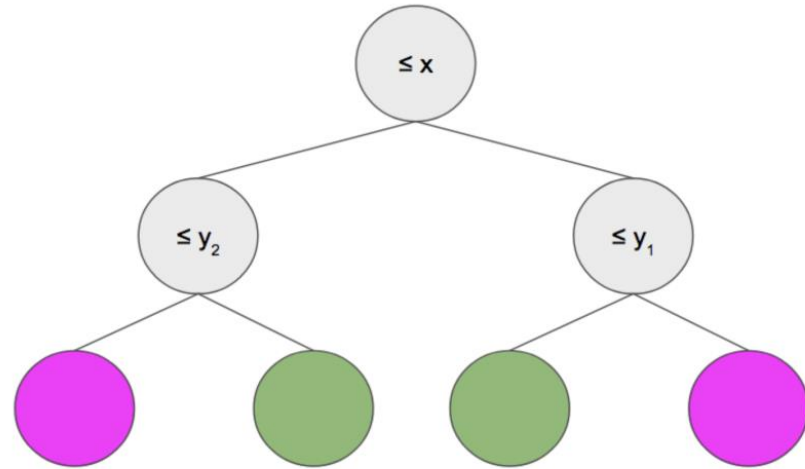
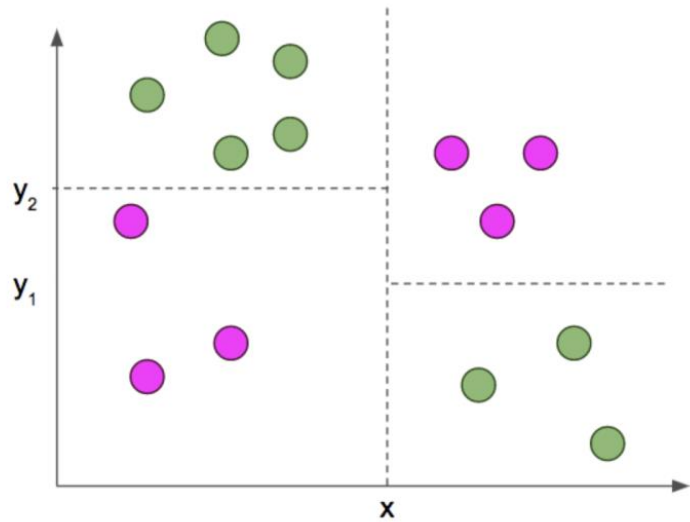
Algoritmo no supervisado y resultado:

Tendremos una clasificación k-means con $K = 2$, logrando un 83% de acierto con respecto a la etiqueta real.



Arbol

Pte



Gracias

