UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA

Introducción a la inteligencia artificial para ciencias e ingenierías
Raul Ramos Pollan
Semestre 2022-1

Estudiantes:
Miguel Angel Castaño Cardenas cc 1152225263
Juan Sebastian Pinto Fuentes cc 1007612134

1. Introducción

Como se expuso en la entrega anterior, vamos a usar un dataset de Kaggle llamado "Respiratory Sound DataBase" (enlace: https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database). Esta base de datos contiene: 920 archivos de sonido .wav, 920 archivos de texto con anotaciones de cada medición donde se evidencia la duración de cada ciclo y si se encuentran crepitancias o sibilancias presentes, se encuentra también un archivo de texto que enumera el diagnóstico de cada paciente, un archivo de texto que explica el formato de nombre del archivo, un archivo de texto que enumera 91 nombres de los 125 pacientes y finalmente un archivo de texto que contiene información demográfica de cada paciente.

Como métricas de desempeño de Machine Learning vamos a usar las métricas accuracy, recall y el F1 score (Harmonic mean). La métrica accuracy, es el número de elementos identificados correctamente como positivo de un total de elementos identificados como positivos, es decir que esta métrica da cuenta de los falsos positivos.

La métrica de recall, es el número de elementos identificados correctamente como positivos del total de positivos verdaderos, es decir recall nos da información sobre el rendimiento de un clasificador con respecto a falsos negativos.

Por último la métrica F1, surge a partir de la necesidad de poder medir en una sola métrica varios valores que evalúan el modelo, para este caso específico con el F1 las métricas que se juntan son el Recall y Accuracy. La media armónica es una especie de promedio cuando accuracy y recall son iguales. Pero cuando accuracy y recall son diferentes, entonces está más cerca del número más pequeño en comparación con el número más grande; esto evita que se sobrecalifique un modelo que realmente puede no estar haciendo bien su trabajo.

2. Exploración descriptiva del dataset

La exploración del dataset se realizó haciendo uso de un notebook de colab, donde se encontró que los datos de los pacientes que se incluyeron en el dataset, están clasificados tanto por edades como por la posición del estetoscopio en el momento de realizar la medición. Tomando como base la clasificación por edades de la figura 1, se encontró que de los 125 pacientes hay:

- 36 bebés
- 7 infantes
- 7 adolescentes
- 1 adultos jóvenes
- 12 adultos
- 62 adultos mayores

Bebés	0-5 años
Infantes	6-12 años
Adolecentes	13- 19 años
Adultos jóvenes	20-34 años
Adultos	35 -60 años
Adultos mayores	< 60 años

Figura 1. Clasificación poblacional por edades.

Respecto a la clasificación según la posición del estetoscopio, se describe en la competencia que hay 7 posibles casos, a continuación se muestra cada una de las posiciones con el respectivo número de mediciones para cada posición. Cabe aclarar que se realizó más de una medición por paciente, con diferentes posiciones de estetoscopio, debido a esto hay más datos de mediciones que de pacientes.

- Posterior izquierda (Pl): 139 mediciones.
- Posterior derecha (Pr): 132 mediciones.
- Anterior izquierda (Al): 162 mediciones.
- Anterior derecha (Ar): 168 mediciones.
- Tráquea (Tc): 130 mediciones.
- Lateral izquierda (Ll): 77 mediciones.
- Lateral derecha (Lr): 112 mediciones.

3. Descripción del progreso alcanzado

Para el preprocesamiento del dataset decidimos clasificarlo según la posición del estetoscopio, debido a esto realizamos una tabla con la siguiente estructura:

Tabla 1. Descripción del procesamiento del dataset

Edad	Sexo	Pl	Pr	Al	Ar	Tc	Ll	Lr	Diagnóstico	Audio
74	1	0	1	0	0	0	0	0	1	

Donde la primera columna se refiere a la edad correspondiente a cada paciente, sigfuente está la columna del sexo, que se clasifica como 1(Masculino) o 0 (Femenino), las columnas referentes a cada ubicación del estetoscopio se clasifican con el valor de 1 si esa es la posición y 0 al resto de columnas, posteriormente se encuentra la columna de diagnostico que corresponde a los casos de sano y enfermo; esta clasificación se realizó tomando como base el archivo de información de cada medición donde se informaba si habían sibilancias o crepitancias presentes en los ciclos respiratorios, según esta información se clasifica al paciente como sano (sin sibilancias o crepitancias) con el número 0 y como enfermo (con sibilancias o crepitancias) con el número 1, y finalmente encontramos la columna de audio que cuyo contenido es el resultado del calculo de la densidad espectral del ciclo respiratorio al que se hace referencia en cada fila del dataframe, con estos datos se realiza el entrenamiento, prueba y validación del algoritmo de machine learning.

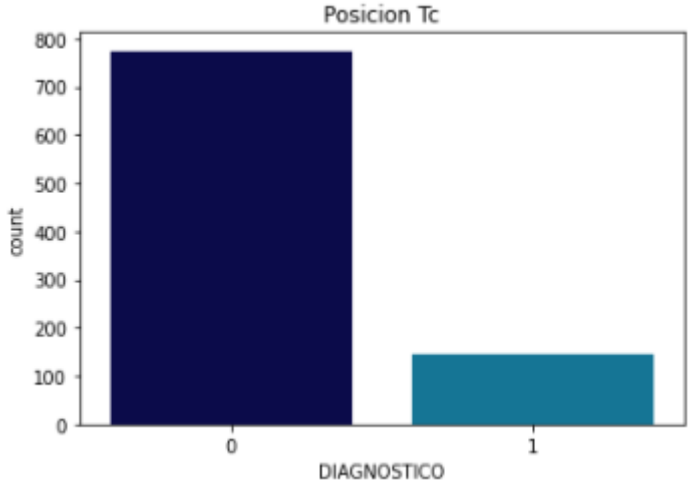

Figura 2. Diagnósticos en la posición Tc.

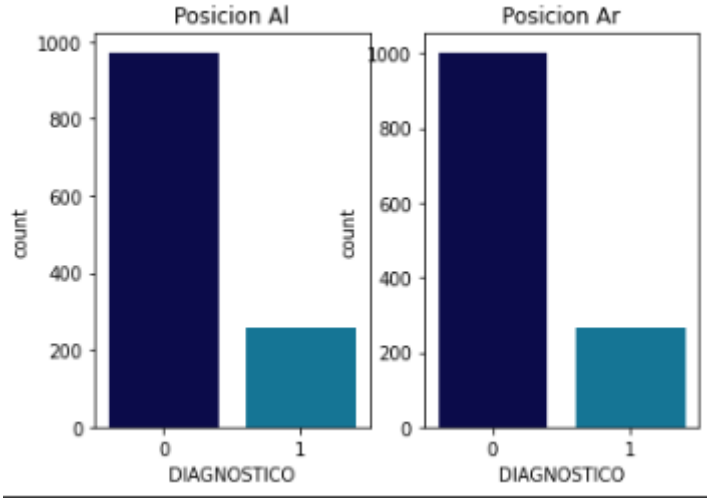

Figura 3. Diagnósticos en las posiciones Al y Ar.

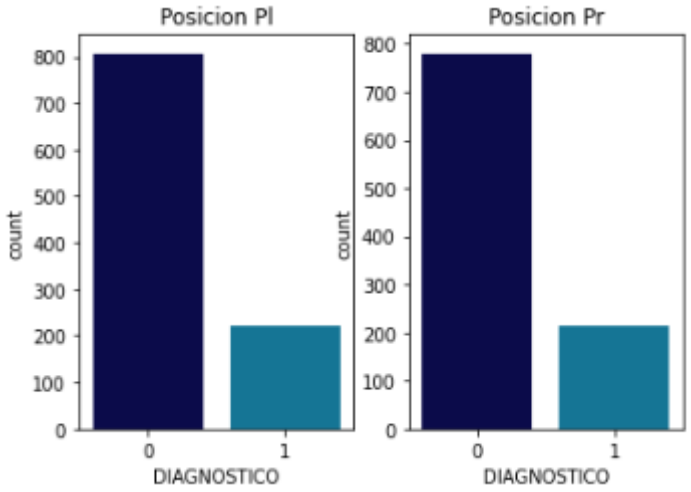

Figura 4. Diagnósticos en las posiciones Pl y Pr.

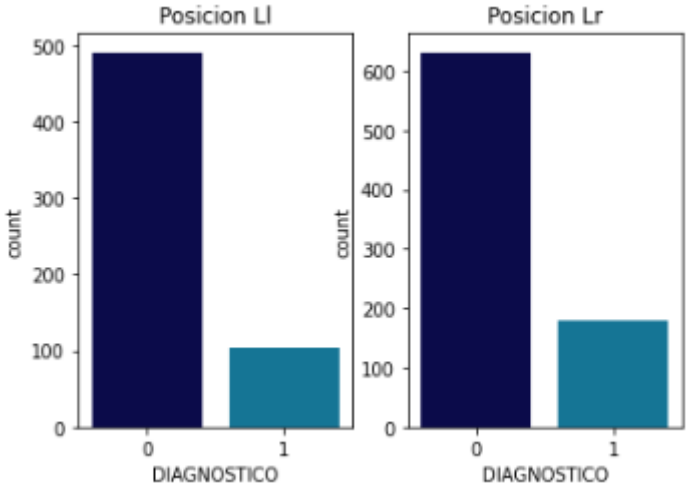

Figura 5. Diagnósticos en las posiciones Ll y Lr.

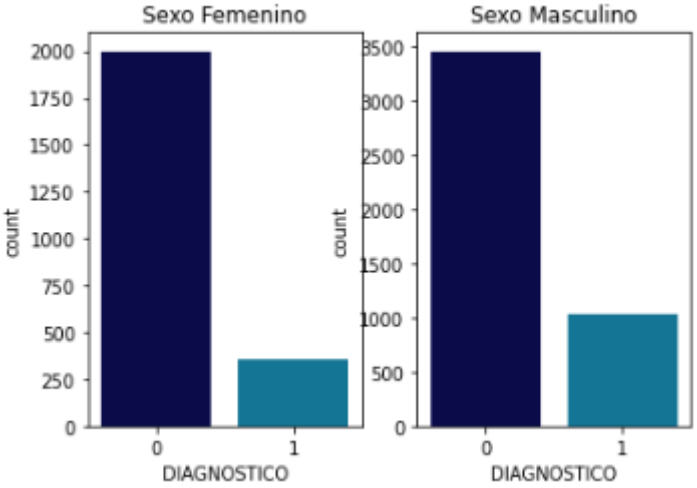

Figura 6. Diagnósticos segun el sexo.

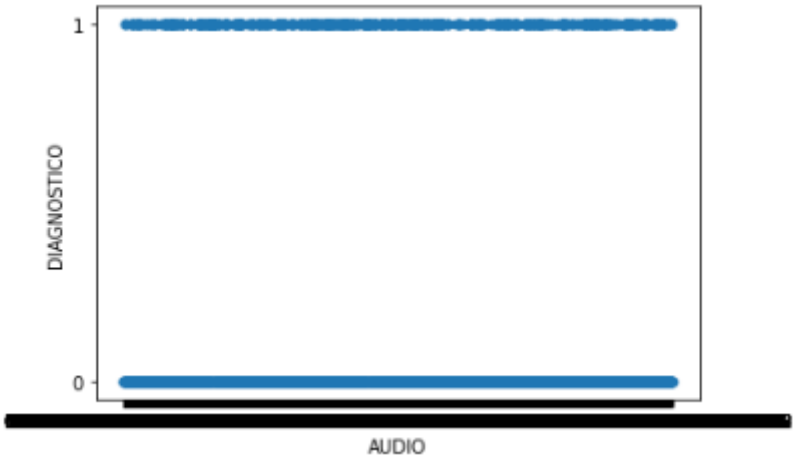

Figura 7. Diagnósticos según la densidad espectral.