Tecnológico de Monterrey

Proyecto Integrador (Gpo. 10)

Avance 3. Baseline

Equipo #6

Julio César Pérez Zapata	A01793880
Christian Emilio Saldaña López	A00506509
Jorge Estivent Cruz Mahecha	A0179380

Se explica por qué el algoritmo seleccionado es apropiado para el tipo de problema que se está abordando.

En base al contexto baseline, el modelo seleccionado como base para solucionar el problema en relación y entender cómo trabaja los modelos de predicción desde los simple a complejo, hemos seleccionado el modelo de **regresión logística**

Una razón fundamental por la cual la regresión logística constituye una opción sólida para construir nuestro modelo de predicción sobre la presencia de enfermedades del habla a partir de datos de audio es su capacidad para describir con precisión la relación entre las variables de entrada y la variable de salida.

La regresión logística proporciona resultados que pueden entenderse fácilmente. La regresión logística genera una probabilidad que puede **interpretarse** como la probabilidad de que un individuo esté en una clase específica (enfermo o sano). Esto ayudará a comprender los factores más importantes entre las características de audio en la clasificación.

La eficiencia computacional de la regresión logística es sobresaliente, lo que la hace aplicable para manejar conjuntos de datos de moderados a grandes. En tareas de clasificación binaria, como la que estamos tratando, la regresión logística se puede entrenar rápidamente incluso en un conjunto de datos enorme(lo cual no tenemos como base, pero si logramos obtener haciendo uso del data augmentation).

Además La regresión logística, que es un modelo de clasificación, es buena para tratar múltiples características de entrada, incluidos los coeficientes de MFCC extraídos del audio.

Inicialmente, nuestro objetivo era desarrollar un modelo que permitiera identificar patologías en la voz de las personas sometidas a prueba. Durante nuestra investigación, hemos adquirido información relevante sobre el proceso de comprensión de los datos. Al analizar el conjunto de datos completo, hemos identificado las siguientes características:

Desbalance de clases
El conjunto de datos consta de más de dos mil grabaciones de personas, tanto sanas como enfermas, con patologías detalladas. Sin embargo, observamos un desequilibrio en las clases de este conjunto de datos. Para abordar este problema, hemos llevado a cabo varios procesos de aumento de datos con el fin de normalizarlos y obtener resultados más precisos.

Formato de audio
Entre los datos iniciales del conjunto de datos, destacamos que los audios tienen una resolución de 16 bits a 50 kHz. Además, el conjunto de datos incluye información sobre el género, la edad y el diagnóstico clínico de la patología de cada individuo. Es importante señalar que hay 869 personas sanas y 1356 personas con trastornos de la voz en el conjunto de datos.

Análisis de datos

Durante el proceso de análisis e implementación del modelo inicial, así como para comprenderlo mejor, hemos utilizado los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC). Esta técnica nos ha proporcionado una forma inicial y visual de comprender los datos que estamos analizando. Al realizar transformadas matemáticas mediante convoluciones, pasamos del audio puro a obtener directamente una imagen espectral de la señal. Esta imagen muestra tanto la potencia espectral como las principales frecuencias y características de la señal de manera destacada.

Porque los MFCC?

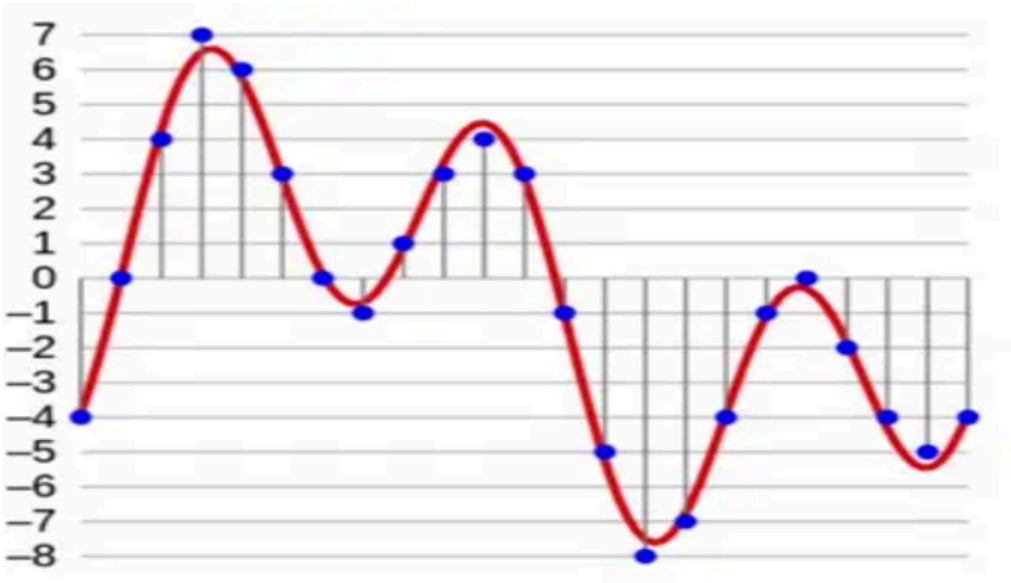
Son coeficientes para la representación del habla basados en la percepción auditiva humana. Estos surgen de la necesidad, en el área del reconocimiento de audio automático, de extraer características de las componentes de una señal de audio que sean adecuadas para la identificación de contenido relevante, así **como obviar todas aquellas que posean información poco valiosa como el ruido de fondo, emociones, volumen, tono, etc.** y que no aportan nada al proceso de reconocimiento, al contrario lo empobrecen.

MFCCs se calculan comúnmente de la siguiente forma:

- Separar la señal en pequeños tramos.
- A cada tramo aplicarle la Transformada de Fourier discreta y obtener la potencia espectral de la señal.(aquí se pasa de dominio de tiempo a dominio de frecuencia)
- Aplicar el banco de filtros correspondientes a la Escala Mel al espectro obtenido en el paso anterior y sumar las energías en cada uno de ellos.
- Tomar el logaritmo de todas las energías de cada frecuencia mel
- Aplicarle la transformada de coseno discreta a estos logaritmos.

En la gráfica de abajo una representación general de cómo se toman las características de la señal de audio en el dominio del tiempo para aplicar la transformada de fourier y pasarla a dominio de tiempo.

Imagen Gráfica de Mel

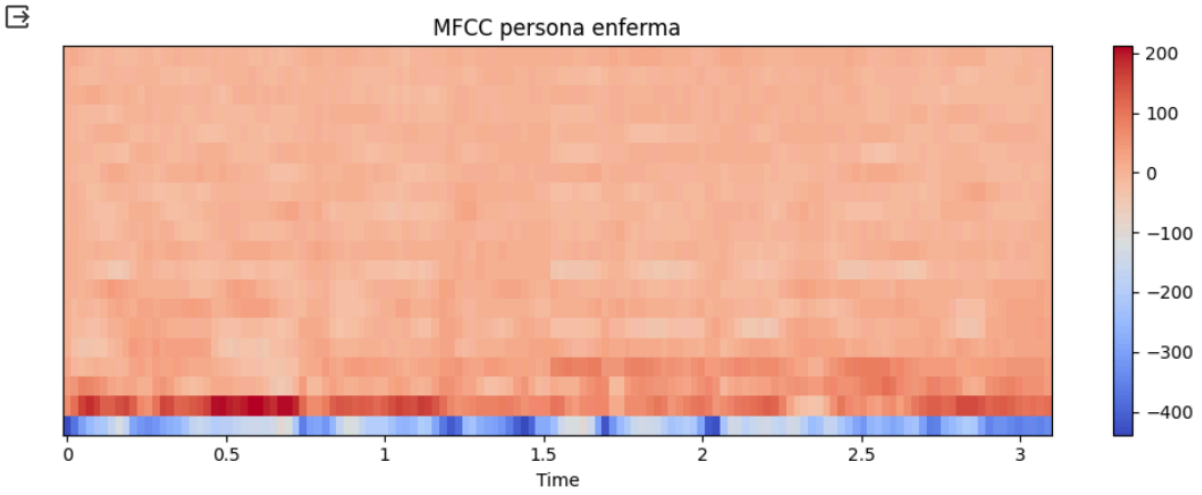


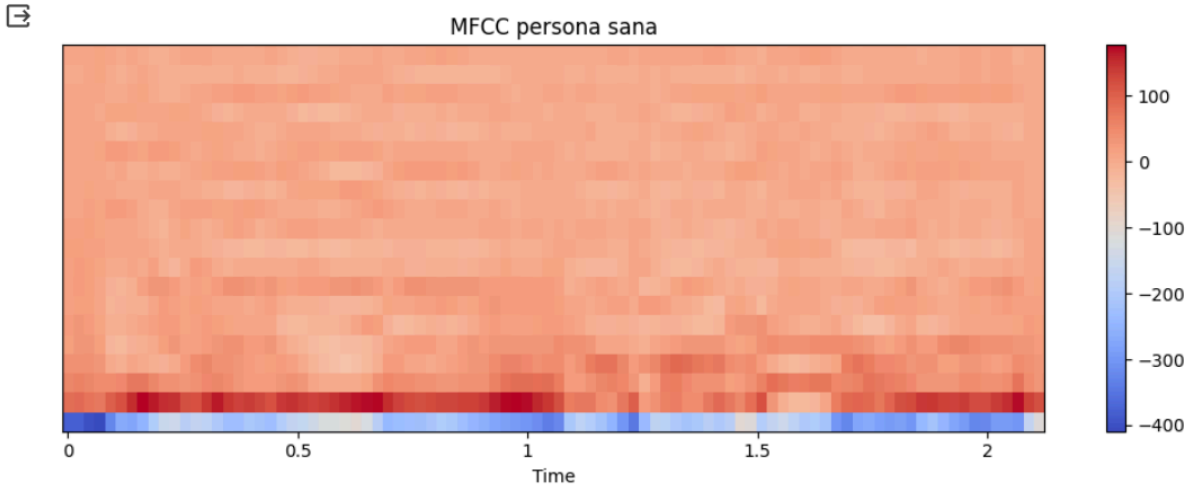
- ¿Qué algoritmo se puede utilizar como baseline para predecir las variables objetivo?

Al abordar el análisis del conjunto de datos, se propone desarrollar un algoritmo que facilite una clasificación precisa de las patologías de voz (siendo estas la variable objetivo). En primer lugar, se procede a cargar los archivos de audio y a transformarlos mediante la extracción de características utilizando el espectrograma de mel.

- ¿Se puede determinar la importancia de las características para el modelo generado?

  Los coeficientes de Mel ofrecen una manera eficaz de extraer las características principales necesarias para alimentar nuestro modelo. Esta técnica nos impulsa hacia adelante en la búsqueda de resultados óptimos, ya que al comprender visualmente estos espectrogramas, podemos analizar meticulosamente cada aspecto clave de los datos. Esto, a su vez, nos capacita para preparar un modelo apropiadamente ajustado y eficaz.





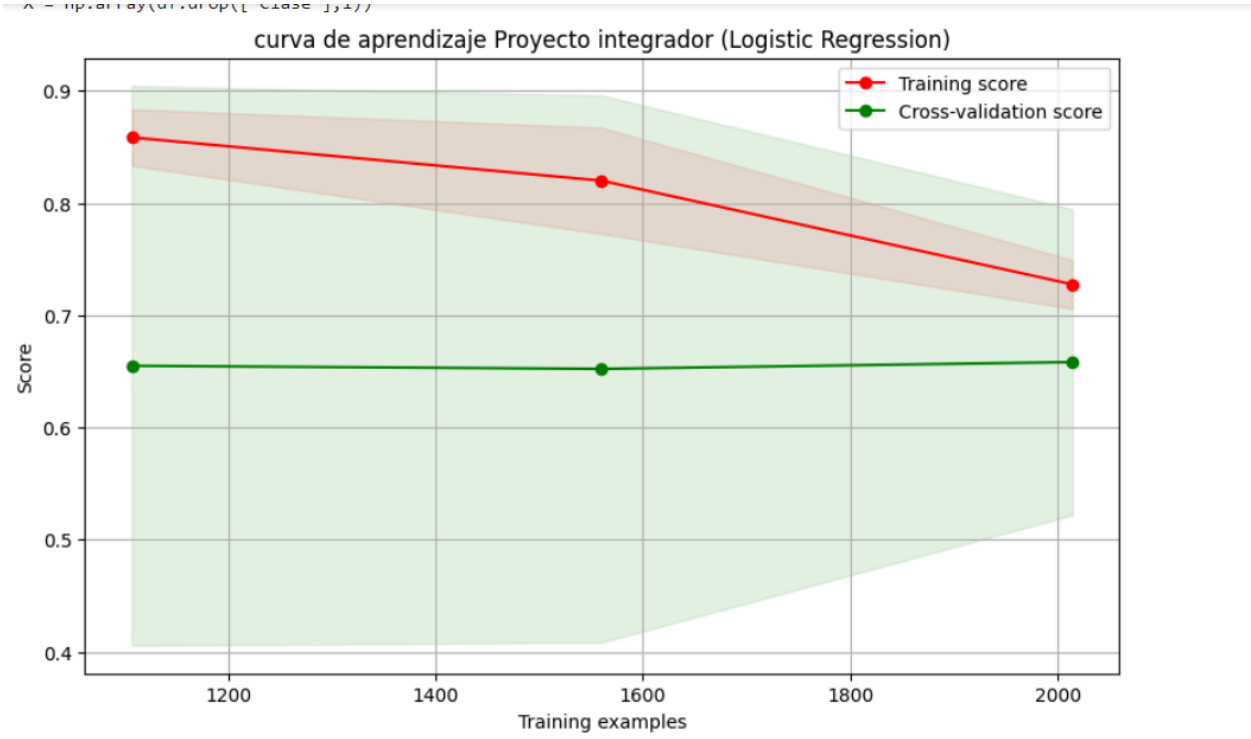
En la imagen del espectrograma de Mel se puede observar las potencias y niveles de energía de la señal procesada, por lo tanto es una herramienta muy provechosa, tanto para el entendimiento de los datos como para el análisis del algoritmo, destacando las características principales.

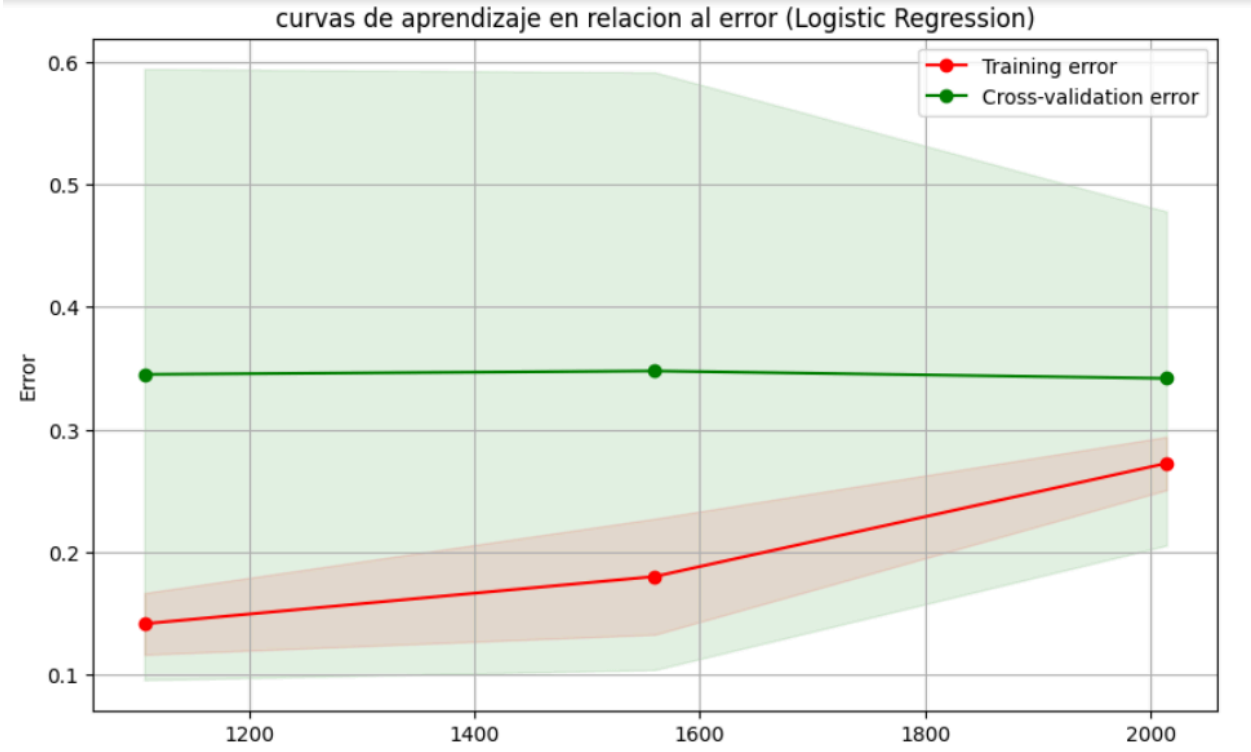
- ¿El modelo está sobre ajustado a los datos de entrenamiento?

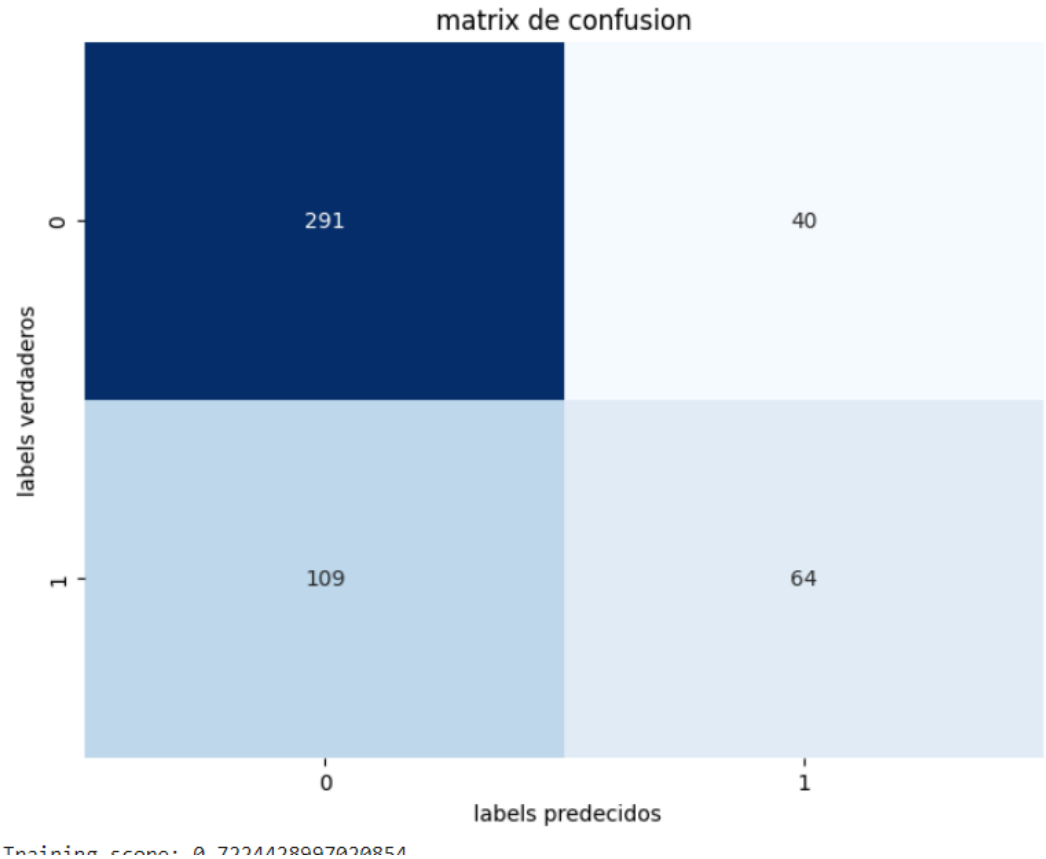
El modelo no está sobreajustado esta sub ajustado.El modelo presenta ajustes deficientes en los datos de aprendizaje.

Ahora el rendimiento deficiente podría deberse a que el modelo es demasiado sencillo.

En las entregas venideras demostraremos otros modelos más complejos que pueden dar mejores resultados.


curva de aprendizaje Proyecto integrador (Logistic Regression)
Training score
Cross-validation score
Score
Training examples
0.9
0.8
0.7
0.6
0.5
0.4
1200
1400
1600
1800
2000


curvas de aprendizaje en relacion al error (Logistic Regression)
Training error
Cross-validation error
Error
0.6
0.5
0.4
0.3
0.2
0.1
1200
1400
1600
1800
2000



- ¿Cuál es la métrica adecuada para este problema de negocio?

La métrica adecuada se presenta mediante la precisión y sensibilidad del modelo esto debido a que debemos garantizar una precisión aceptable y bajar la fluctuación de los resultados del modelo para así disminuir los falsos positivos y falsos negativos, como podemos observar en la matriz de confusión se tienen resultados con una gran cantidad de falsos negativos por lo tanto debemos trabajar en mejorar estos resultados.

```
Training score: 0.7224428997020854
Test score: 0.7043650793650794
```

- ¿Cuál debería ser el desempeño mínimo a obtener?

Se está buscando inicialmente una clasificación binaria de mínimo el 80% de Precisión y de la clasificación multiclase de una mínimo de 60% por clase,esto dado a las investigaciones relacionadas y modelos actuales que alcanzan este mínimo requerido sin embargo aun con modelo inicial no hemos alcanzado el

umbral deseado para garantizar lo anterior por lo tanto es necesario la prueba de más modelos y diferentes tratamientos a los datos de data augmentation para poder alcanzar los resultados esperados.

Referencias

Equipo GitHub
https://github.com/julioperezzapata/Proyecto_integrador_grupo_6