Tecnológico de Monterrey

Proyecto Integrador (Gpo. 10)

Avance 1. Análisis Exploratorio de Datos

Equipo #6

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

Descripción de los datos: El dataset se encuentra compuesto de audios grabados en formato .wav los cuales fueron capturados por la Universidad de Saarlandes y separadas por diferentes etiquetas acerca de patologías vocales, el dataset está compuesto por grabaciones vocales de más de 2000 personas, las cuales contienen la grabación de las vocales a, i, u, adicionalmente de la frase “Guten Morgen, wie geht es ihnen”(Buenos dias, como estas.) grabadas en idioma alemán, para lo cual se realiza la transformación de los datos(audios), extrayendo espectrogramas de mel para poder analizar las características de cada audio.

Dataframe de las características extraídas:

	Media señal	Dstandar	MFCC_1	MFCC_2	MFCC_3	MFCC_4	MFCC_5	MFCC_6	MFCC_7	MFCC_8	...	MFCC_12	MFCC_13	Amáxima
0	0.000212	0.221247	-203.598648	174.003159	-0.848443	-23.611488	4.095771	12.983047	-31.629698	-15.329629	...	3.628854	4.765277	0.889053
1	-0.000394	0.246436	-111.214272	141.009583	-4.112024	-7.820776	-45.904152	4.175006	23.918943	-2.760360	...	-5.137012	30.035673	0.781513
2	0.000291	0.217159	-252.967316	166.925629	35.660683	-3.403091	-29.674986	6.641043	-17.894094	10.894828	...	14.931800	5.484111	0.530140
3	0.000053	0.199383	-216.098526	159.769394	-17.801723	-25.351851	-35.523823	2.549157	-3.534539	11.599912	...	-10.149856	2.400179	0.497984
4	0.000109	0.183465	-151.921799	140.086838	0.269669	-11.125966	-28.982672	-16.409653	12.686049	-5.591802	...	-8.897948	-2.063018	0.559642
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2036	0.000565	0.157136	-192.016083	164.215836	-37.132828	-18.451574	-8.830174	-6.031417	-3.380558	11.603949	...	15.181127	-13.768450	0.512715
2037	0.000506	0.215795	-248.503418	153.451004	15.261247	-3.907525	-3.369242	5.782655	9.785728	0.936441	...	-6.115865	-4.948894	0.548672
2038	0.000083	0.134173	-216.302200	201.367004	-38.002705	-4.974300	-18.938168	-29.329718	10.800496	11.457781	...	15.361176	-1.074535	0.850048
2039	-0.000063	0.110712	-278.585114	131.954346	10.638890	-10.060069	-56.021496	12.052641	10.749830	16.058668	...	-10.650711	-2.234592	0.368215
2040	-0.000415	0.095373	-283.017731	124.991890	-25.008957	-20.485167	-16.904882	-11.110275	10.813704	-3.239727	...	9.726804	-29.259996	0.265439

041 rows × 23 columns

Amín	AvgCS	AvgBws	Avgrolloff	AvgCrossZ	RMS	Clase
0.000000	916.593560	1364.923778	1379.593173	0.034801	0.211691	1
0.000014	1601.902180	2203.528221	3299.317383	0.059225	0.243490	1
0.000004	690.353130	1262.718469	881.665039	0.020973	0.213819	1
0.000000	952.993783	1290.571320	1266.152344	0.044678	0.196496	1
0.000000	1295.470874	1926.992803	2031.811523	0.040318	0.180750	1
...	...	...	...	...	...	...
0.000013	1047.048910	1346.701853	1496.151330	0.038270	0.154651	0
0.000002	761.188456	1467.923575	1020.520020	0.016532	0.210597	0
0.000000	933.407504	1035.757952	1254.437256	0.045602	0.129253	0
0.000000	1020.237963	1521.859301	1193.247070	0.040527	0.107454	0
0.000004	1148.032430	1622.209446	1543.379815	0.044036	0.092449	0

Las siguientes son algunas de las preguntas comunes que podrán abordar a través del EDA:

- ¿Hay valores faltantes en el conjunto de datos? ¿Se pueden identificar patrones de ausencia?

No hay valores faltantes, dado que el dataset se encuentra compuesto de audios grabados en formato .wav los cuales fueron capturados por la Universidad de Saarlandes y separadas por diferentes etiquetas de los datos acerca de patologías vocales, el dataset está compuesto por grabaciones vocales de más de 2000 personas, las cuales contienen la grabación de las vocales a, i, u, adicionalmente de la frase “Guten Morgen, wie geht es ihnen”(Buenos dias, como estas.) grabadas en idioma alemán, para lo cual se realiza la transformación de los datos(audios), extrayendo espectrogramas de mel para poder analizar las características de cada audio.

```
# Verificar si hay algún valor nulo en el DataFrame
hay_nulos = df.isnull().any().any()

# Imprimir el resultado
if hay_nulos:
    print("El DataFrame tiene valores nulos.")
else:
    print("El DataFrame no tiene valores nulos.")
```

```
El DataFrame no tiene valores nulos.
```

- ¿Cuáles son las estadísticas resumidas del conjunto de datos?

Total de personas: 2225

Sanos: 869

Enfermos: 1356

```
[43] print(df.describe())
```

```
       Media señal     Dstandar       MFCC_1       MFCC_2       MFCC_3  \
count  2041.000000  2041.000000  2041.000000  2041.000000  2041.000000
mean     -0.002485     0.169235  -201.862549   142.500046   -18.298199
std       0.016945     0.053150    43.464443    30.967344    26.473400
min      -0.157546     0.042450  -368.580475    42.513683   -93.747116
25%      -0.000107     0.129600  -230.164017   120.552490   -36.241241
50%       0.000076     0.162253  -204.211319   139.752151   -18.880692
75%       0.000284     0.201403  -174.958511   164.524368    -0.691808
max       0.005719     0.374774   -21.968611   232.361237    66.196556

            MFCC_4       MFCC_5       MFCC_6       MFCC_7       MFCC_8  ...  \
count  2041.000000  2041.000000  2041.000000  2041.000000  2041.000000  ...
mean    -19.176683   -28.475712    -6.868766     8.000731     4.396277  ...
std      16.431587    16.521515    15.504835    13.790057    12.798373  ...
min     -74.443863   -80.855911   -69.161858   -41.442822   -46.571537  ...
25%     -30.439953   -39.677635   -16.314251    -1.397617    -4.302752  ...
50%     -20.043371   -28.374290    -6.323344     8.122311     4.883978  ...
75%      -7.820776   -17.014896     4.093213    17.457531    13.065988  ...
max      47.876595    22.500782    39.501770    49.409733    45.975719  ...

           MFCC_11      MFCC_12      MFCC_13      Amáxima         Amín  \
count  2041.000000  2041.000000  2041.000000  2041.000000  2.041000e+03
mean     -2.102271    -0.127836    -6.916659     0.587577  4.966712e-06
std      10.998176    12.117620    13.297187     0.183537  1.019037e-05
min     -34.729580   -40.012920   -48.706249     0.133897  0.000000e+00
25%      -9.601206    -8.745698   -16.540234     0.444264  0.000000e+00
50%      -2.317574    -0.110255    -7.469853     0.569132  1.960434e-07
75%       5.209218     8.562780     2.070260     0.719316  5.458482e-06
max      35.178688    40.838078    42.047199     0.993331  1.280038e-04

             AvgCS       AvgBws   Avgrolloff    AvgCrossZ          RMS
count  2041.000000  2041.000000  2041.000000  2041.000000  2041.000000
mean   1251.358006  1638.190229  1962.963503     0.050689     0.165430
std     384.022303   440.521880  1164.176809     0.019348     0.051887
min     586.590598   833.247854   796.939625     0.009428     0.040964
25%    1006.563428  1321.163412  1268.352475     0.039230     0.126994
50%    1180.064041  1547.986191  1540.737810     0.049186     0.158316
75%    1391.282158  1862.069264  2185.849194     0.059678     0.197186
max    4053.609489  3456.667046  8590.181996     0.231498     0.364937

[8 rows x 22 columns]
```

Se toma 23 características principales del dataset el cual lo describe estadísticamente

- ¿Hay valores atípicos en el conjunto de datos?

No, el conjunto de datos está compuesto por audios las cuales tienen su etiqueta. No hay patología que tenga valores atípicos.

- ¿Cuál es la cardinalidad de las variables categóricas?

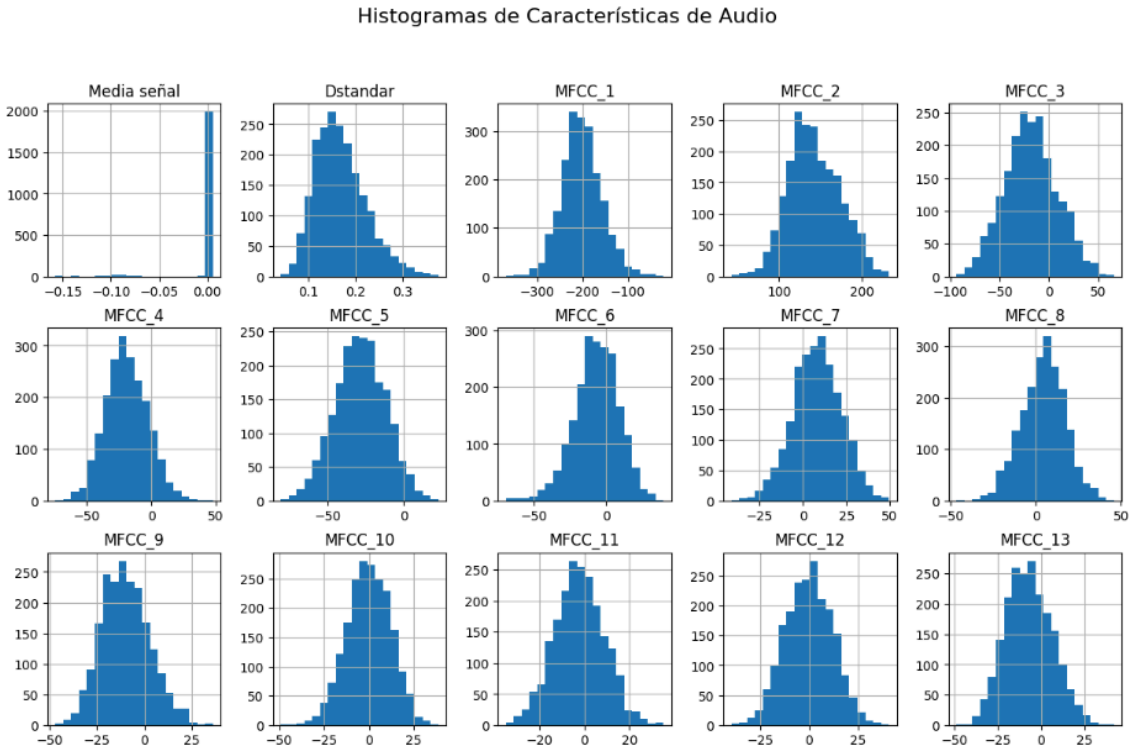
Dentro del dataset se identificó alrededor de 7 patologías oficiales.

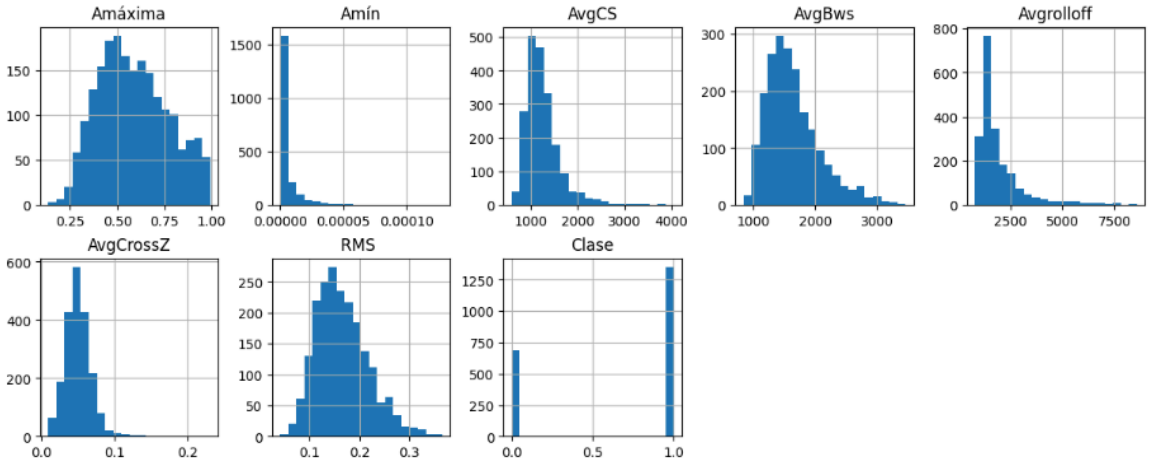
```
# Cardinalidad de las variables categóricas
cardinalidad_categoricas = df['Clase'].nunique()
print(f"Cardinalidad de la variable categórica 'Clase': {cardinalidad_categoricas}")
```

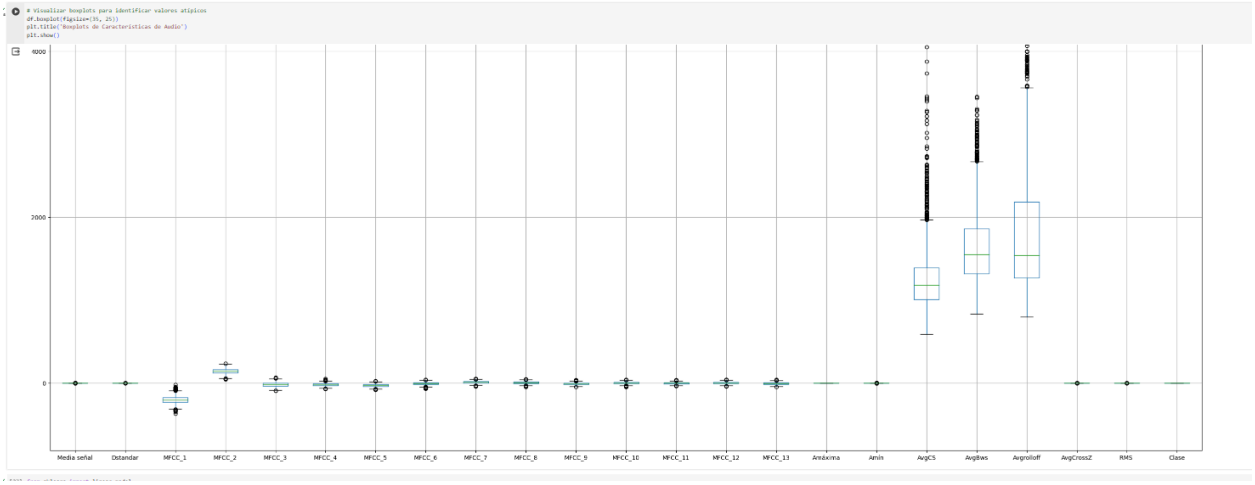
```
Cardinalidad de la variable categórica 'Clase': 2
```

- ¿Existen distribuciones sesgadas en el conjunto de datos? ¿Necesitamos aplicar alguna transformación no lineal?

```
[51] # Visualizar histogramas
     df.hist(bins=20, figsize=(15, 15))
     plt.suptitle('Histogramas de Características de Audio', y=0.95, fontsize=16)
     plt.show()
```





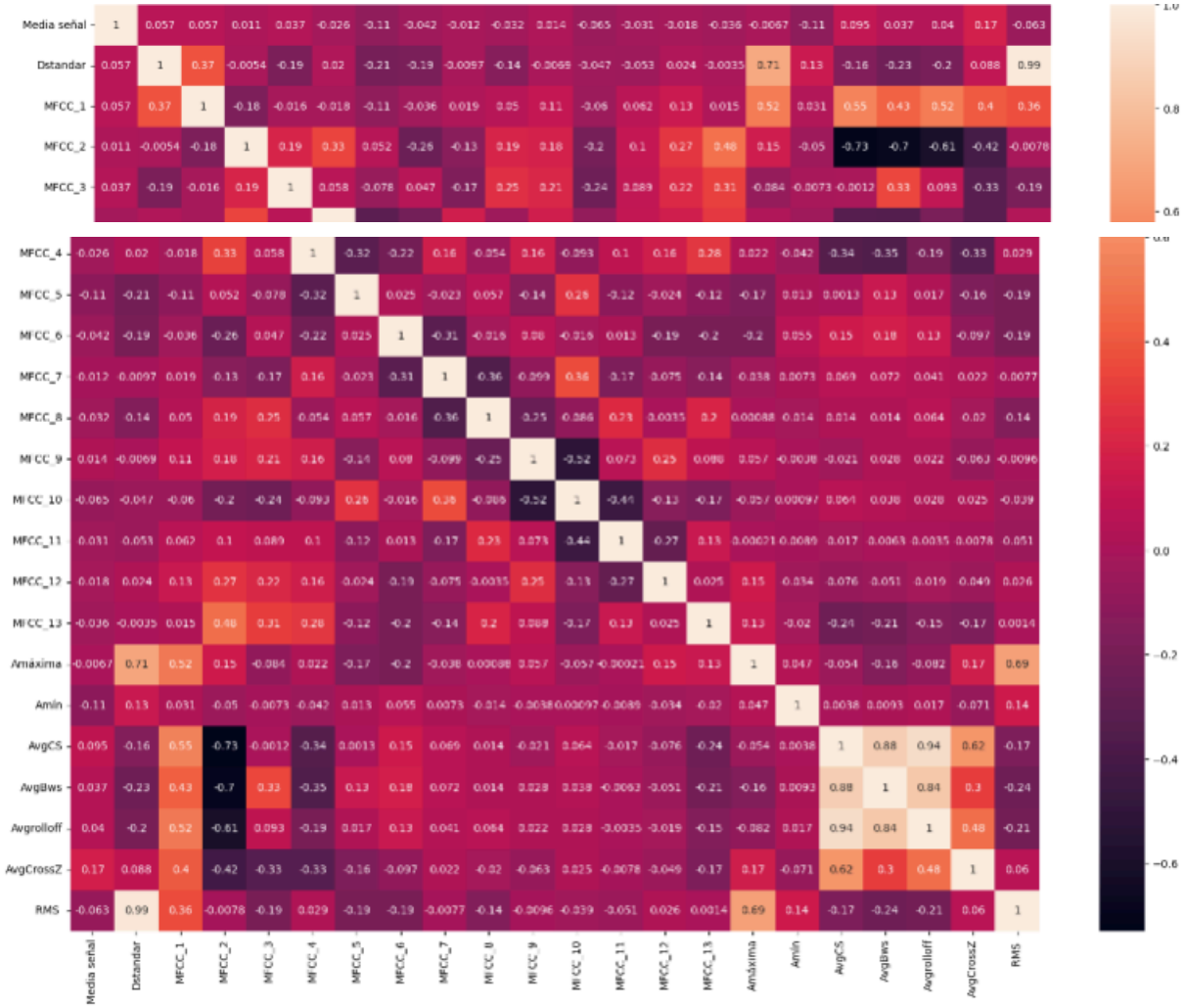


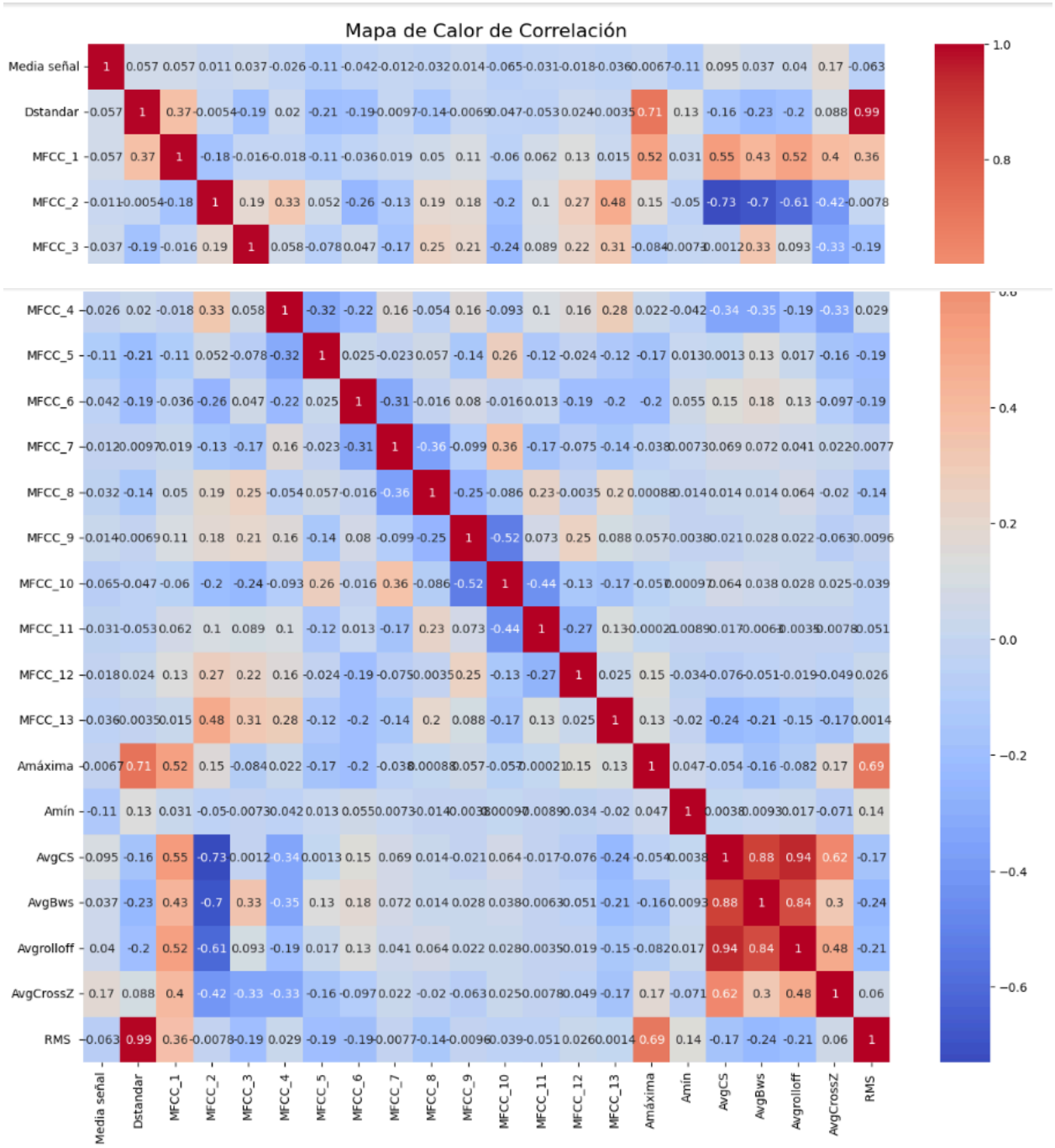
Con lo visualizado anteriormente podemos afirmar que la mayoría de características tienen un comportamiento sin sesgos aparentes aunque al tener claro la naturaleza de los datos podemos afirmar que es una muestra de una población cerrada y no define directamente la población a escalas mayores, adicionalmente los Box Plot permiten visualizar los datos atípicos de cada característica así como se encuentran agrupados los datos.

- ¿Se identifican tendencias temporales? (En caso de que el conjunto incluya una dimensión de tiempo).

  No, se puede afirmar que el dataset no está relacionado a tendencias temporales ya que el enfoque está dado a la identificación de patologías vocales y no a sucesos capturados en líneas temporales.

- ¿Hay correlación entre las variables dependientes e independientes?


Media señal
Dstandar
MFCC_1
MFCC_2
MFCC_3
MFCC_4
MFCC_5
MFCC_6
MFCC_7
MFCC_8
MFCC_9
MFCC_10
MFCC_11
MFCC_12
MFCC_13
Amáxima
Amín
AvgCS
AvgBws
Avgrolloff
AvgCrossZ
RMS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

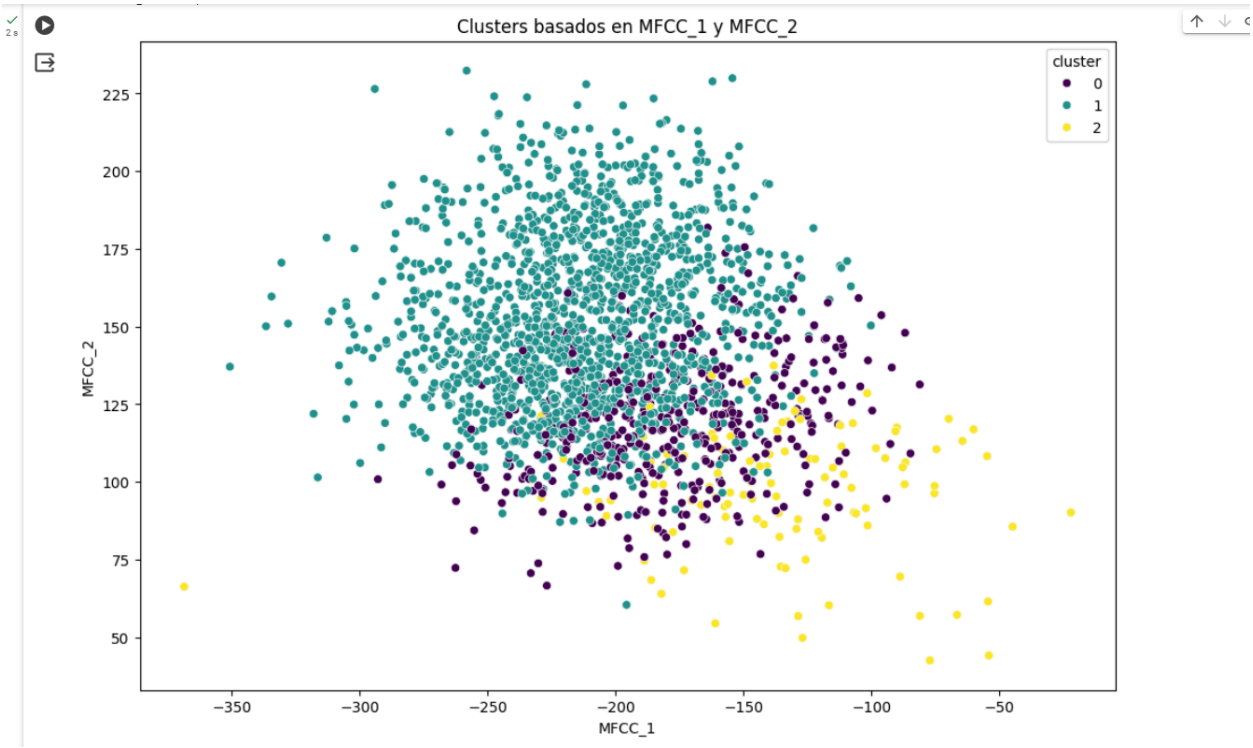


Dada la matriz de correlación se puede observar solamente correlación marcada entre los diferentes promedios de los datos directamente de las características principales no se ve correlación.

¿Cómo se distribuyen los datos en función de diferentes categorías?

- ¿Existen patrones o agrupaciones (clusters) en los datos con características similares?

Se cuenta con un conjunto de características extraídas del audio,características MFCC, espectrogramas, podemos utilizar K-Means para agrupar estos vectores de características en clusters. Esto puede ayudar a identificar patrones o segmentos específicos en tus datos de audio pero no tenemos contemplado usar esta técnica dado que para nuestro propósitos tenemos planeados usar otras técnicas y modelos.



- ¿Se deberían normalizar las imágenes para visualizarlas mejor?

Para nuestro caso no aplicar normalizar imágenes, pero si normalizar datos para que las variables sean comparables entre sí, dependiendo de los métodos y algoritmos que usemos, normalizar puede ser beneficioso. Algunos modelos de machine learning, como las máquinas de soporte vectorial (SVM) o los algoritmos basados en distancias, pueden beneficiarse de la normalización.

- ¿Hay desequilibrio en las clases de la variable objetivo?

Si, actualmente el dataset está fuertemente balanceado hacia personas con una patología. Mientras que la clase de gente “sana” es de menor proporción.

```python
# Verificar desequilibrio en las clases
desequilibrio_clases = df['Clase'].value_counts()
print("Distribución de clases:")
print(desequilibrio_clases)
# 1= enfermo
# 2= sano
```

```
Distribución de clases:
1    1354
0     687
Name: Clase, dtype: int64
```

Estamos considerando el nivelar los datos entre las etiquetas de Sano y Enfermo probando varias técnicas de balanceo que nos garanticen la precisión del modelo, como son el submuestreo y el sobremuestreo de los datos.

Al trabajar un dataset extraído de los audios basados principalmente en los coeficientes MFCC nos permite capturar las características más relevantes por medio del resumen espectral del sonido, este dataset nos brinda una confiabilidad de la información ya que reduce la dimensionalidad de los datos con representaciones más relevantes del espectro, esto a su vez ayudará el entrenamiento del modelo

Conclusiones:

- Al realizar el análisis al dataset podemos asegurar que el proyecto detalla una dificultad interesante dado que se debe ajustar y nivelar los datos para poder garantizar unas métricas finales precisas que permitan ejecutar un buen modelo.
- Se encontró el problema que algunas clases de patologías poseen muy pocos datos ocasionando un gran desbalance en los datos. En general la clase de gente sana también es minoría en los datos. Todo esto ocasiona el tener que invertir más tiempo en buscar más información o considerar la creación de datos sintéticos.
- Al trabajar un dataset extraído de los audios basados principalmente en los coeficientes MFCC nos permite capturar las características más relevantes por medio del resumen espectral del sonido, este dataset nos brinda una confiabilidad de la información ya que reduce la dimensionalidad de los datos con representaciones más significativas del espectro, esto a su vez ayudará el entrenamiento del modelo.