

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY



Maestría en Inteligencia Artificial Aplicada (MNA)

Proyecto Integrador (TC5035)

Dra. Grettel Barceló Alonso

Dr. Luis Eduardo Falcón Morales

Asesora: Dra. María de la Paz Rico Fernández

Avance 1

Análisis exploratorio de datos

“Clasificación de ruido en laboratorio de motores eléctricos automotrices a través de métodos de inteligencia artificial”

EQUIPO 14

Andrei García Torres A01793891

Emmanuel González Calitl A01320739

Denisse María Ramírez Colmenero A01561497

Fecha: 5 de mayo de 2024

ÍNDICE

Análisis exploratorio de datos	1
Estructura de los datos	1
Análisis Univariable	2
Análisis Bi/Multivariable	3
Preprocesamiento	4
Conclusiones	5
Repositorio de GitHub	5
Bibliografía	5

Análisis exploratorio de datos

Este primer avance consiste en un análisis exploratorio de datos (EDA - *Exploratory Data Analysis*). Describe los datos utilizando técnicas estadísticas y de visualización (análisis univariante y bi/multivariante) para hacer énfasis en los aspectos más relevantes, así como aplicar y justificar operaciones de preprocesamiento, relacionadas con el manejo de valores faltantes, atípicos y alta cardinalidad. Incluye conclusiones del EDA, identificando tendencias o relaciones importantes.

Estructura de los datos

Como se explicó en el avance 0, este proyecto pretende la clasificación de ruidos de motores elevadores para la industria automotriz por medio de la utilización de métodos de inteligencia artificial. Cada muestra de ruido puede clasificarse como aceptable (OK) o rechazada (NOK). El conjunto de datos de las mediciones obtenidas en el laboratorio de ruido para las muestras de los motores elevadores (que son imágenes de espectrogramas en formato png) tiene la siguiente conformación:

349 muestras con mediciones de ruido dentro de especificación, muestras etiquetadas como OK.

265 muestras con mediciones de ruido fuera de especificación, muestras etiquetadas como NOK.

Por consiguiente, las muestras etiquetadas como OK representan el 56.84 % de la totalidad, mientras que las muestras etiquetadas como NOK representan el 43.16 %.

En este conjunto de datos no se presentan valores faltantes ni atípicos ya que todas las muestras en condiciones OK y en condiciones NOK han sido colectadas y etiquetadas por el especialista de ruido dentro del laboratorio.

Análisis Univariable

Tabla 1

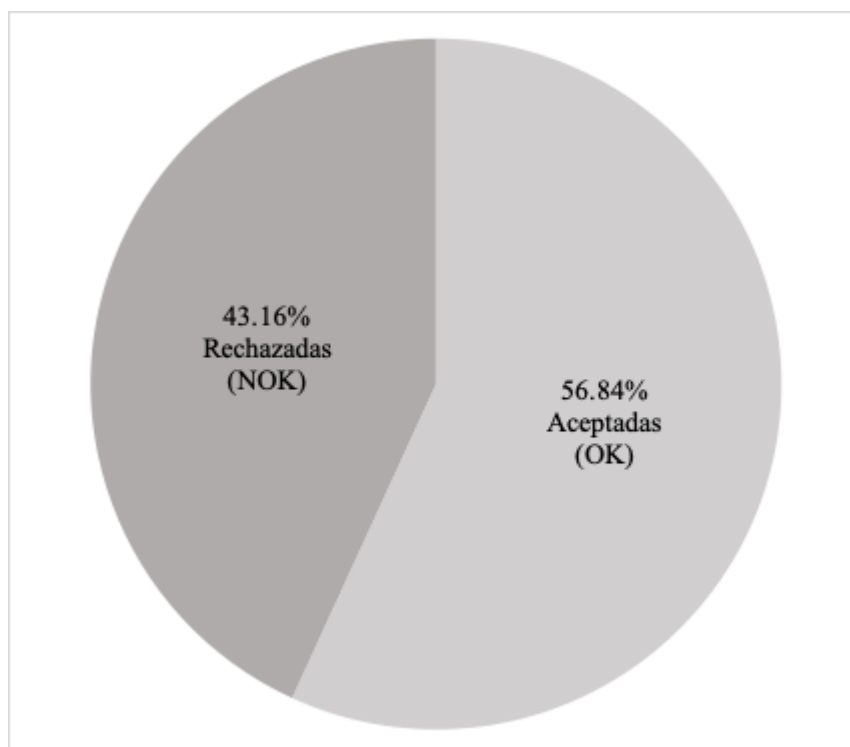
Distribución del conjunto de datos

Clase	Cantidad de muestras	Porcentaje
Aceptables (OK)	349	56,84%
Rechazadas (NOK)	265	43,16%
Total	614	100,00%

Nota. Esta tabla muestra cómo están distribuidas las muestras de ruido (imágenes de espectrogramas) en las dos clases del conjunto de datos.

Figura 1

Porcentaje de distribución de las clases



No se tiene equilibrio de las clases de la variable del objeto con los conjuntos de mediciones para la clasificación de ruido, por lo que para cada algoritmo de clasificación será necesario compensar el desequilibrio de la cantidad de imágenes y el balanceo.

Análisis Bi/Multivariable

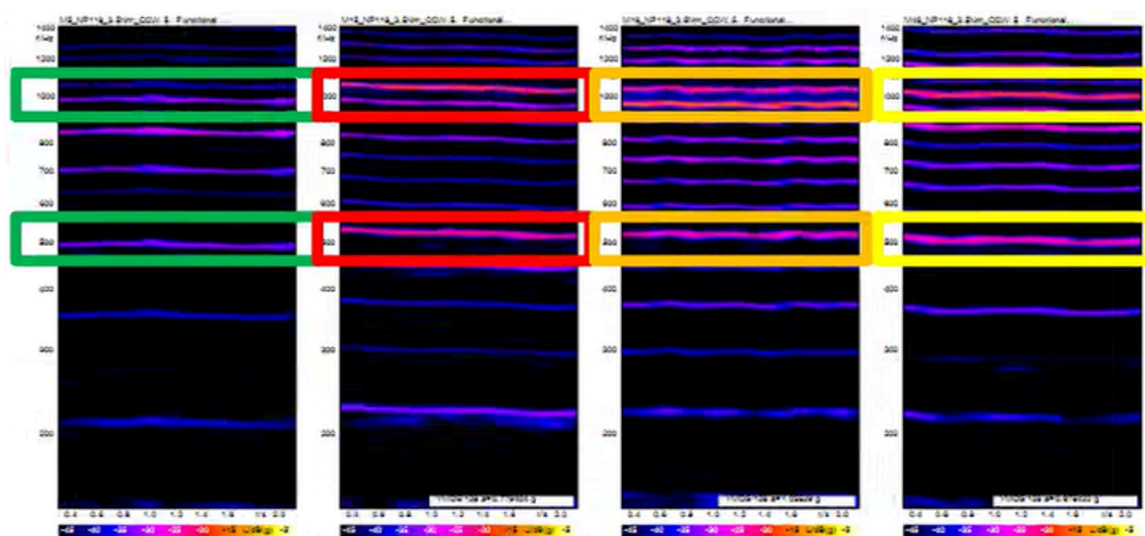
Los datos son mediciones de ruido que pueden visualizarse para cada muestra a través de imágenes de espectrogramas. Estas mediciones de ruido son gráficas que se encuentran en formato png donde el eje X representa la frecuencia en hertz y el eje Y representa la intensidad del ruido en decibeles.

No se tiene un valor específico de decibeles en el que el ruido ya es considerado no aceptable dentro de la frecuencia entre 0 y 1000 hertz, sino que el departamento de ruido de la planta ha catalogado cada muestra según su criterio propio. Las siguientes figuras representan un ejemplo de una clase.

Como se puede observar, para que una muestra sea considerada aceptable, debería no tener mediciones mayores a -28 decibeles y la mayoría de la imagen debería contener mediciones de ruido cercanas a -50 decibeles. La figura 2 muestra la clasificación de los espectrogramas donde las curvas de calor son los la ondas sonoras que son medibles, donde las curvas que colores de rosa a amarillo presentan magnitudes de ruido fuera de la especificación del producto y las ondas sonoras de negro a azul son magnitudes audibles dentro de especificación.

Figura 2

Imágenes de espectrogramas de ruido



Nota. Espectrogramas con mediciones dentro de especificación marcados en verde y espectrogramas fuera de especificación marcados en rojo, naranja y amarillo.

Preprocesamiento

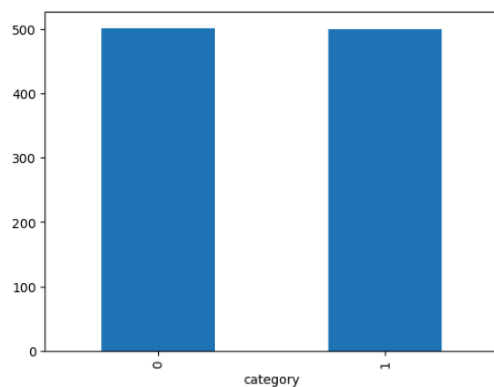
Debido a que nuestro conjunto de datos son imágenes de espectrogramas de los ruidos de los motores y ya están etiquetadas como ok y nok, no contamos con datos atípicos. Sin embargo, se podría realizar un aumento de datos y filtrado de las imágenes para adecuar el entrenamiento de las redes neuronales para la clasificación de las imágenes. En las imágenes del código se muestra la clasificación binary donde 0 se designa a imágenes nok y 1 se designa a imágenes ok.

Para ésta demostración se han aumentado las imágenes a 500 muestras para cada caso clasificación 0 y 1.

```
[5]: len(df)
[5]: 1002
[7]: df = df.drop(df.index[1])
[11]: len(df)
[11]: 1001
```

Sobre vista del conteno total de entrenamiento

```
[15]: df['category'].value_counts().plot.bar()
[15]: <Axes: xlabel='category'>
```



Se importan librerías

```
[1]: import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

from keras.preprocessing.image import ImageDataGenerator, load_img
from keras.utils import to_categorical
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import random
import os
```

Se preparan datos de entrenamiento

```
[3]: filenames = os.listdir("Drives")
categories = []
for filename in filenames:
    category = filename.split('.')[0]
    if category == 'ok':
        categories.append(1)
    else:
        categories.append(0)

df = pd.DataFrame({
    'filename': filenames,
    'category': categories
})
df.head()
```

```
[3]:
```

	filename	category
0	.ipynb_checkpoints	0
1	nok(1).png	0
2	nok(10).png	0
3	nok(100).png	0
4	nok(101).png	0

Conclusiones

Este proyecto se diferencia por tener una naturaleza de datos binaria donde no se requiere el preprocesamiento de datos para valores faltantes o atípicos y solo es necesario el etiquetado del nombre de las imágenes como aceptables (ok) y rechazadas (nok) lo cual ya fue previamente realizado. Es necesaria la técnica de aumento de datos y filtrado de imágenes para el entrenamiento de las redes neuronales y algoritmos de clasificación.

Repositorio de GitHub

Link del repositorio en Github: https://github.com/emm-gl/project_mna

Carpeta: Test_Clasification_20240505

Bibliografía

Bulentsiyah. (2019, 12 enero). *Dogs vs. Cats Classification (VGG16 Fine Tuning)*. Kaggle.

<https://www.kaggle.com/code/bulentsiyah/dogs-vs-cats-classification-vgg16-fine-tuning>

Chollet, F. (2017). *Deep learning with python*. Manning Publications.