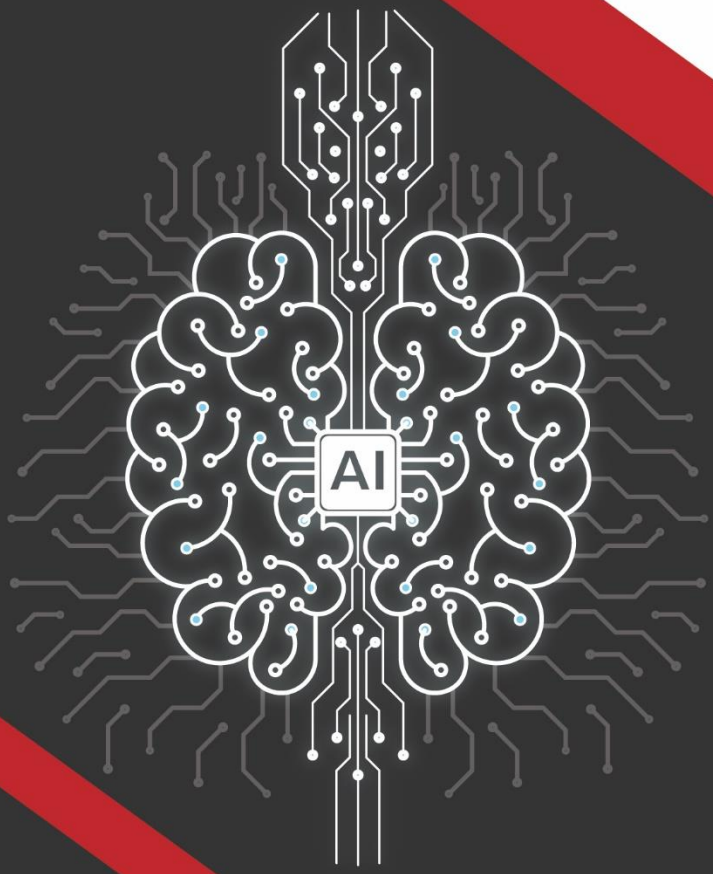
UNIVERSIDAD AUTÓNOMA DE QUERÉTARO | FACULTAD DE INGENIERÍA | DIPFI POSGRADO INGENIERÍA

CARACTERÍSTICAS DE UN DATASET

LI. Antonio García Gutiérrez
Expediente: 261724



I. OBJETIVO

Desarrollar un programa donde se puedan apreciar las características de un conjunto de datos, tal como su promedio, su desviación estándar, su distribución, su correlación, etc.

II. INTRODUCCIÓN

Recopilar datos, este es el primer paso en el desarrollo de un modelo de Machine Learning. Es un paso crítico con una influencia absoluta en cómo será de adecuado el modelo. Cuanto más y mejores datos obtengamos, mejor será el rendimiento de nuestro modelo.

Las medidas de tendencia central, más conocidas y utilizadas son: la media aritmética, la mediana, etc. . Estas medidas son utilizadas en estadística para detallar ciertos comportamientos de un grupo de datos, por ejemplo, a qué valor están cercanos, cuál es el promedio de los datos recogidos, entre otros.

Cuando se trata de datos agrupados, se hace referencia a una cantidad dada de datos que puede clasificarse, ya sea por sus cualidades cualitativas o cuantitativas, y por tal, agruparse para su análisis, esto nos posibilita cuantificar la realidad y disponer de los elementos que nos permitan, después de su tratado de limpieza, imputación, diseñar una herramienta de visualización que organizará los datos en gráficos y visualizaciones para ayudarlos a ver patrones generales en los datos, correlaciones generalizadas y posibles valores atípicos.

III. PROCESO

Para sacar toda la estadística de los datos es necesario realizar múltiples operaciones que nos darán a conocer los sesgos que tienen nuestros datos.

El **promedio** de un conjunto de datos se encuentra al sumar todos los números en el conjunto de datos y luego al dividir entre el número de valores en el conjunto.

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

La **mediana** es el valor medio cuando un conjunto de datos se ordena de menor a mayor.

La **varianza** mide el valor promedio de los cuadrados de las distancias entre los valores de los datos y la media de la muestra.

La varianza, definida como s^2, del conjunto de datos formado por los n valores numéricos $x_1, x_2, \dots, x_n$. Está definida como:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

La **covarianza** es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias.

Nos permite saber cómo se comporta una variable en función de lo que hace otra variable. Es decir, cuando X sube ¿Cómo se comporta Y? Así pues, la covarianza puede tomar los siguientes valores:

- Covarianza (X,Y) es menor que cero cuando "X" sube e "Y" baja. Hay una relación negativa.
- Covarianza (X,Y) es mayor que cero cuando "X" sube e "Y" sube. Hay una relación positiva.

Los cuartiles son medidas estadísticas de posición que tienen la propiedad de dividir la serie estadística en cuatro grupos de números iguales de términos.

La **desviación** es la separación que existe entre un valor cualquiera de la serie y la media.

La desviación estándar o desviación típica es una medida que ofrece información sobre la dispersión media de una variable. La desviación estándar es siempre mayor o igual que cero. Está definida como:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

IV. RESULTADOS

El conjunto de datos consta de 10 000 puntos de datos almacenados como filas con 10 atributos en columnas

- UID: identificador único que va de 1 a 10000
- ID de producto: que consta de una letra L, M o H para baja (50 % de todos los productos), media (30 %) y alta (20 %) como variantes de calidad del producto y un número de serie específico de la variante.
- Tipo: específica la calidad del producto.
- Temperatura del aire [K].
- Temperatura del proceso [K].
- Velocidad de rotación [rpm]: calculada a partir de una potencia de 2860 W.
- Torque [Nm]: los valores de par se distribuyen normalmente alrededor de 40 Nm y sin valores negativos.
- Desgaste [min]: Las variantes de calidad H/M/L añaden 5/3/2 minutos de desgaste de herramienta a la herramienta utilizada en el proceso.
- Etiqueta de "fallo de la máquina" que indica si la máquina ha fallado en este punto de datos en particular para cualquiera de los siguientes modos de falla.
- Tipo de fallo

El data set que se utilizó cuenta con 10 atributos, de los cuales tienen el siguiente tipo:

```
Los tipos son (Atributo: valor):
UDI:<class 'numpy.int64'>
Product ID:<class 'str'>
Type:<class 'str'>
Air temperature [K]:<class 'numpy.float64'>
Process temperature [K]:<class 'numpy.float64'>
Rotational speed [rpm]:<class 'numpy.int64'>
Torque [Nm]:<class 'numpy.float64'>
Tool wear [min]:<class 'numpy.int64'>
Target:<class 'numpy.int64'>
Failure Type:<class 'str'>
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

El numero de instancias es el mismo en todos los atributos:

```
El numero de instancias por atributo es (Atributo: valor):
UDI:10000
Product ID:10000
Type:10000
Air temperature [K]:10000
Process temperature [K]:10000
Rotational speed [rpm]:10000
Torque [Nm]:10000
Tool wear [min]:10000
Target:10000
Failure Type:10000
```

El número de observaciones que tiene cada atributo es el siguiente:

```
El numero de observaciones por atributo es (Atributo: valor):
UDI:10000
Product ID:10000
Type:3
Air temperature [K]:93
Process temperature [K]:82
Rotational speed [rpm]:941
Torque [Nm]:577
Tool wear [min]:246
Target:2
Failure Type:6
```

Para esta base de datos se observa que no hay ningún dato faltante, sin embargo, existen datos atípicos.

```
El numero de datos faltantes por atributo es (Atributo: valor):
UDI:0
Product ID:0
Type:0
Air temperature [K]:0
Process temperature [K]:0
Rotational speed [rpm]:0
Torque [Nm]:0
Tool wear [min]:0
Target:0
Failure Type:0
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

El primer caso de ambigüedad está presente en los datos que no presentan la etiqueta de falla sin embargo en tipo de falla esta falla aleatoria.

	Target	Failure Type
1221	0.0	Random Failures
1302	0.0	Random Failures
1748	0.0	Random Failures
2072	0.0	Random Failures
2559	0.0	Random Failures
3065	0.0	Random Failures
3452	0.0	Random Failures
5471	0.0	Random Failures
5489	0.0	Random Failures
5495	0.0	Random Failures
5509	0.0	Random Failures
5553	0.0	Random Failures
5639	0.0	Random Failures
6091	0.0	Random Failures
6913	0.0	Random Failures
6960	0.0	Random Failures
7488	0.0	Random Failures
7868	0.0	Random Failures

La segunda situación presente y la cual tendrá que ser eliminada es el caso de los datos etiquetados con error pero que no presentan un tipo de falla, debido a que los datos que tiene presente la etiqueta de falla tienen mas valor para calcular las fallas, estos datos atípicos podrían causar que el algoritmo sea sensible a ellos, por ello se opto por eliminarlos.

	Target	Failure Type
1437	1.0	No Failure
2749	1.0	No Failure
4044	1.0	No Failure
4684	1.0	No Failure
5536	1.0	No Failure
5941	1.0	No Failure
6478	1.0	No Failure
8506	1.0	No Failure
9015	1.0	No Failure

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

Los valores mínimos obtenidos por los atributos numéricos son los siguientes:

```
Los valores mínimos son (Atributo: valor):
UDI:1
Air temperature [K]:295.3
Process temperature [K]:305.7
Rotational speed [rpm]:1168
Torque [Nm]:3.8
Tool wear [min]:0
Target:0
```

Los valores máximos obtenidos por los atributos numéricos son los siguientes:

```
Los valores máximos son (Atributo: valor):
UDI:10000
Air temperature [K]:304.5
Process temperature [K]:313.8
Rotational speed [rpm]:2886
Torque [Nm]:76.6
Tool wear [min]:253
Target:1
```

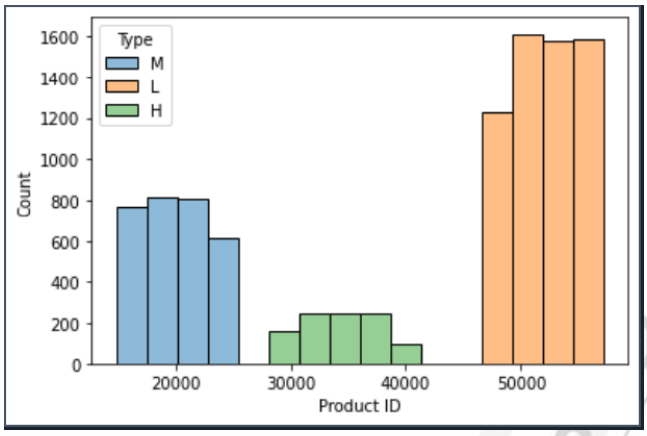
```
Los valores desviación estandar son (Atributo: valor):
UDI:2886.8956799071675
Air temperature [K]:2.000258682915751
Process temperature [K]:1.4837342191657208
Rotational speed [rpm]:179.2840959134266
Torque [Nm]:9.968933725121337
Tool wear [min]:63.654146636636355
Target:0.18098084265065364
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

```
Los valores de variación son (Atributo: valor):
UDI:8334166.666666667
Air temperature [K]:4.001034798579856
Process temperature [K]:2.2014672331233114
Rotational speed [rpm]:32142.787047494745
Torque [Nm]:99.37963961586156
Tool wear [min]:4051.8503840384033
Target:0.032754065406540654
```

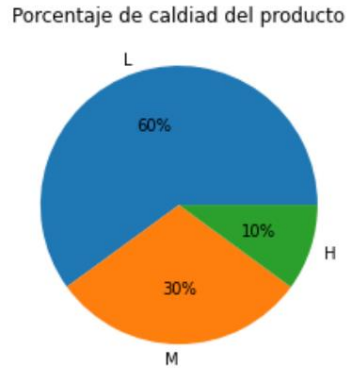
```
Los valores promedio son (Atributo: valor):
UDI:5000.5
Air temperature [K]:300.00493
Process temperature [K]:310.00556
Rotational speed [rpm]:1538.7761
Torque [Nm]:39.986909999999995
Tool wear [min]:107.951
Target:0.0339
```

Ahora dentro de las etiquetas se encuentran básicamente 3 tipos de calidades con números de series etiquetadas y enumeradas según sus características, la serie Low con los números mas altos, la Medium con los mas bajos y la High con la enumeración de en medio.

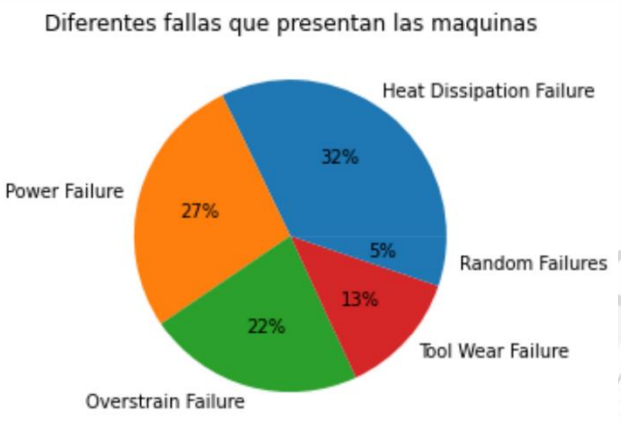


LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

Como característica este atributo se inclina demasiado por la Low, seguida por el medio y por ultima el alta.

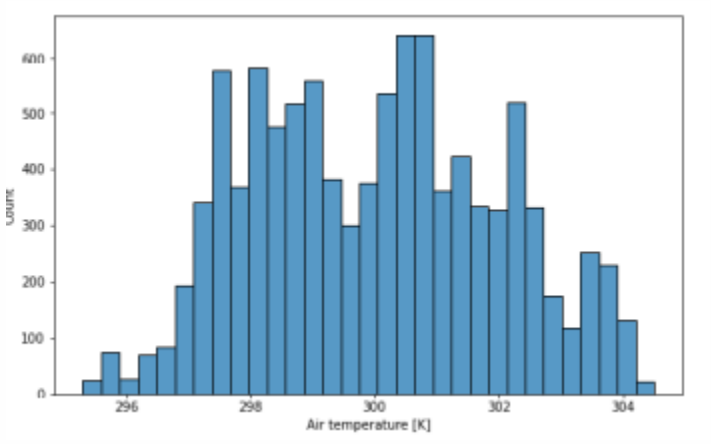


De las maquinas que presentan fallos, solo es alrededor del 3% del total de productos en los que la maquina al fabricarlos se presenta un daño, y de ese 3% los fallos tienen el siguiente balance:

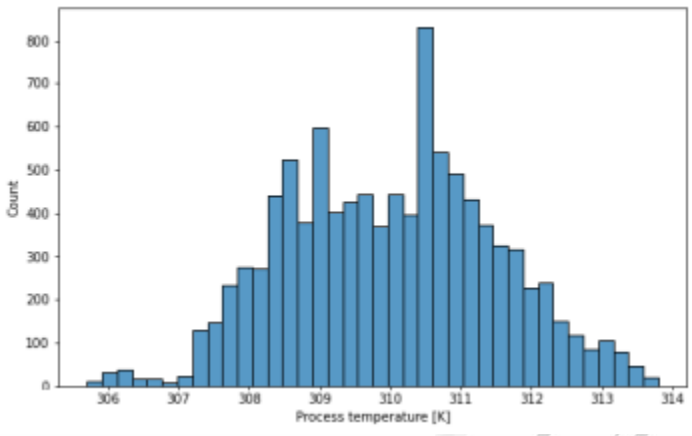


LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

La distribución de la temperatura del aire es multimodal:

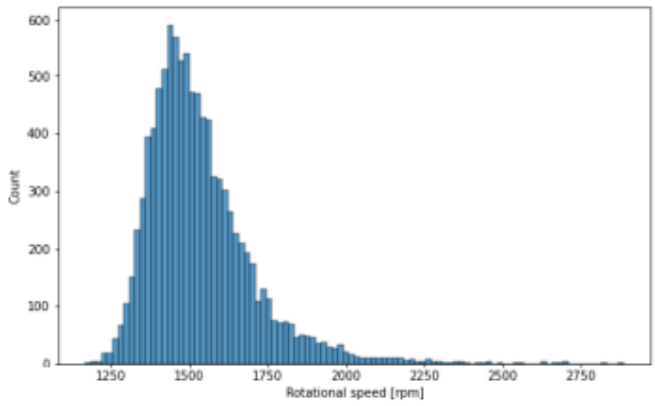


La distribución de la temperatura del proceso es bimodal:

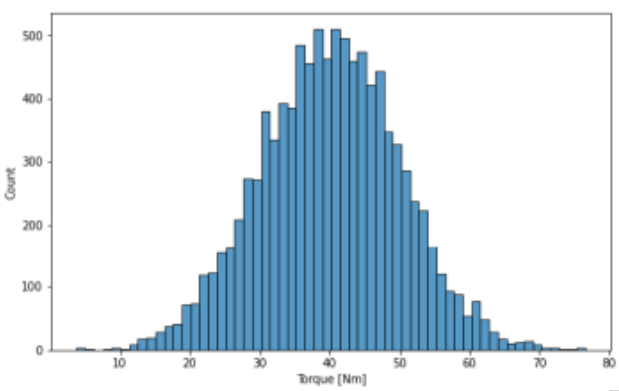


LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

La distribución de la velocidad de rotación es normal sesgada ala izquierda.

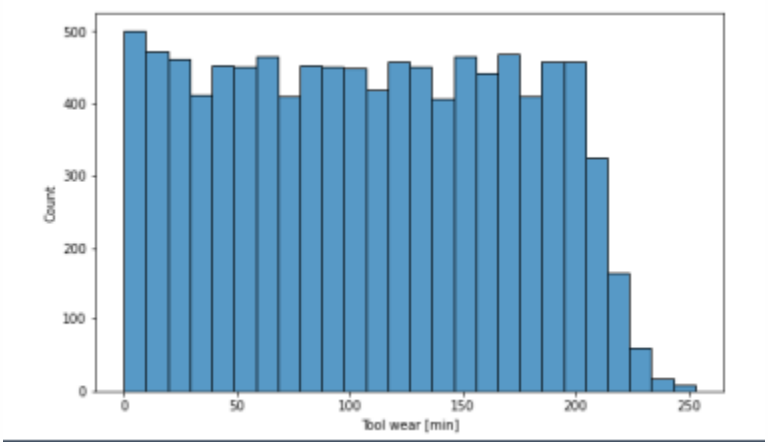


La distribución del torque es normal:

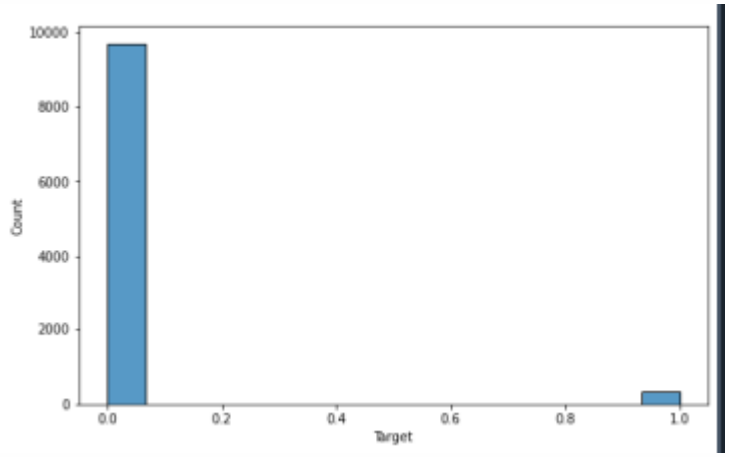


LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

La distribución dedesgaste es uniforme.



Ahora en esta grafica podemos apreciar el balance de las clases de la etiqueta de presenta fallo y no presenta fallo, es de 97% a 3 %.



LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

En general la descripción de nuestro dataset es el siguiente:

Index	UDI	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Target
count	10000	10000	10000	10000	10000	10000	10000
mean	5000.5	300.005	310.006	1538.78	39.9869	107.951	0.0339
std	2886.9	2.00026	1.48373	179.284	9.96893	63.6541	0.180981
min	1	295.3	305.7	1168	3.8	0	0
25%	2500.75	298.3	308.8	1423	33.2	53	0
50%	5000.5	300.1	310.1	1503	40.1	108	0
75%	7500.25	301.5	311.1	1612	46.8	162	0
max	10000	304.5	313.8	2886	76.6	253	1

Para comenzar con el análisis, lo que me llama la atención es que del primer al segundo cuartil existe un avance del aproximadamente 6 números en el torque, lo mismo al cuartil siguiente, sin embargo del 3r cuartil al final existe un salto de 20 números, que es donde creo existían los datos atípicos, de igual forma tiene el comportamiento la velocidad de rotación, en el boxplot se muestran los valores atípicos en estos dos atributos.

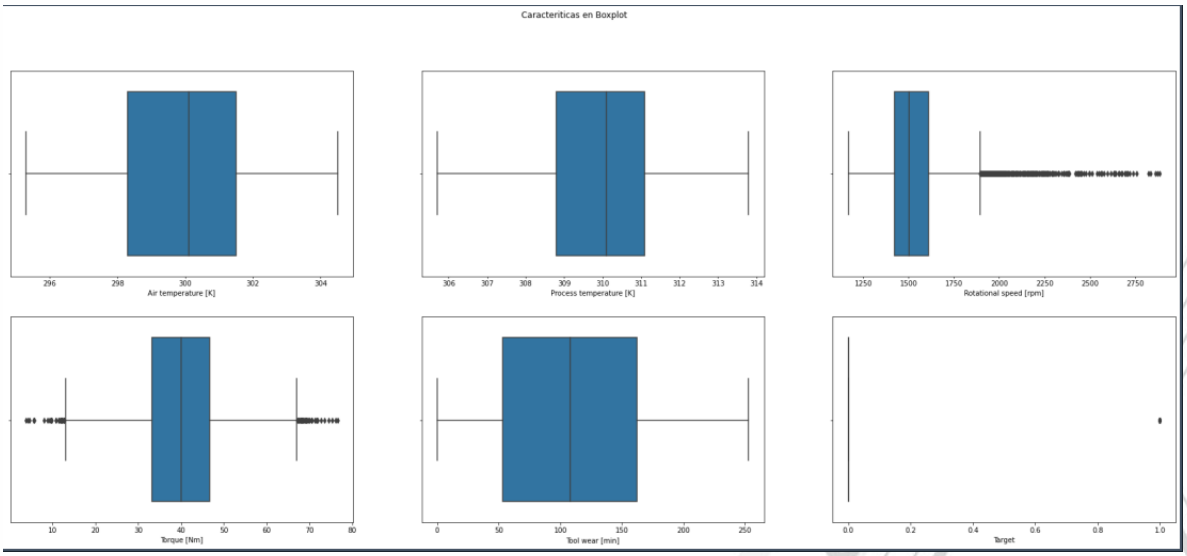
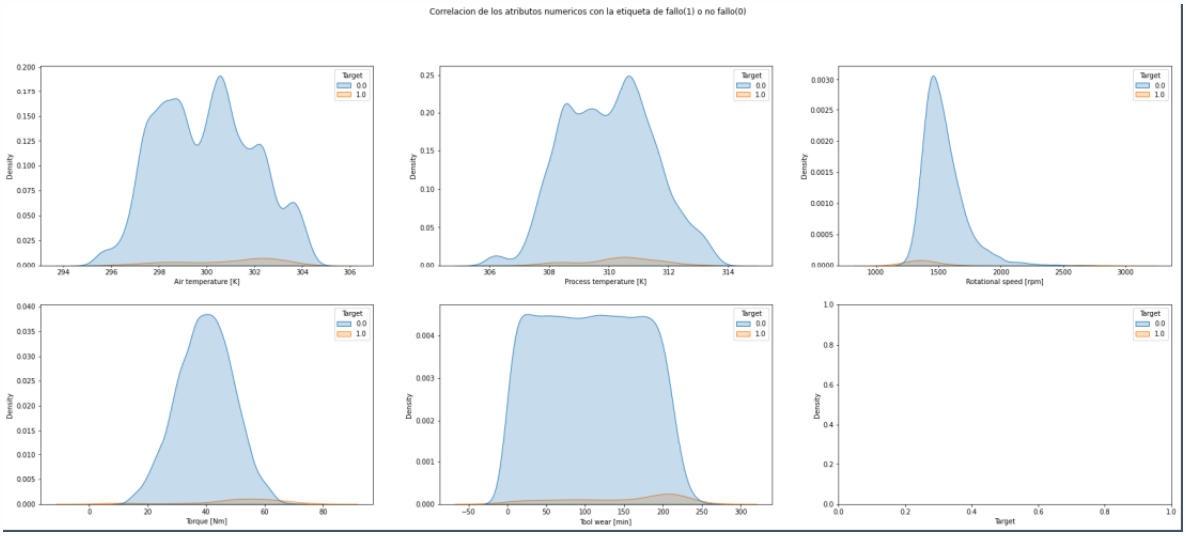


Figura 1: Boxplot de los datos

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

La correlación que existe en los atributos con la falla es negativa en la temperatura del aire y en el desgaste y positiva en todos los demás atributos.



LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

V. CONCLUSIONES

El análisis de los datos nos dan el primer esquema con el cual nos ayudara a preparar los datos para crear el modelo más conveniente, conocer y comprender los datos es fundamental ya que sin este estudio previo no se garantizara un modelo adecuado, puede que al no ser adecuado el estudio, nuestro modelo será sensible a los datos atípicos no tratados, a datos vacíos o a ambigüedades creadas incluso por la misma naturaleza de los datos, todo esto se tiene que estudiar y trabajar para crear la base de nuestro modelo.

VI. CÓDIGO DOCUMENTADO

```
# -*- coding: utf-8 -*-
"""
Created on Sun Aug 14 00:51:36 2022

@author: garci
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

class Estadistica:

    def __init__(self, df):
        self.df = df
        self.min = self.min()
        self.max = self.max()
        self.mean = self.mean()
        self.var = self.var()
        self.std = self.std()
        self.numberOfInstanceOfAtrr = self.numberOfInstanceOfAtrr()
        self.numberOfEmptyInstanceOfAtrr = self.numberOfEmptyInstanceOfAtrr()
        self.observations = self.observations()

    def min(self):
        minimos = {}
        for i in self.df:
            minimos[i] = self.df[i].min()
        return minimos

    def max(self):
        maximos = {}
        for i in self.df:
            maximos[i] = self.df[i].max()
        return maximos

    def mean(self):
        promedios = {}
        for i in self.df:
            if(type(self.df[i].min()) != str):
                promedios[i] = self.df[i].mean()
        return promedios

    def std(self):
        desviacion = {}
        for i in self.df:
            if(type(self.df[i].min()) != str):
                desviacion[i] = self.df[i].std()
        return desviacion

    def var(self):
        v = {}
        for i in self.df:
            if(type(self.df[i].min()) != str):
                v[i] = self.df[i].var()
        return v
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

```python
def desc(self):
    return self.df.describe()

def numberOfInstanceOfAtrr(self):
    numInstancias = {}
    for i in self.df:
        numInstancias[i] = self.df[i].count()
    return numInstancias

def observations(self):
    numInstancias = {}
    for i in self.df:
        numInstancias[i] = len(self.df[i].unique().tolist())
    return numInstancias

def numberOfEmptyInstanceOfAtrr(self):
    numInstancias = {}
    for i in self.df:
        numInstancias[i] = self.df[i].isnull().sum()
    return numInstancias

def numeroAtributos(self):
    numAtributos = 0
    for i in self.df:
        numAtributos += 1
    return numAtributos

def plotAll(self):
    self.df['Tool wear [min]'] = self.df['Tool wear [min]'].astype('float64')
    self.df['Rotational speed [rpm]'] = self.df['Rotational speed [rpm]'].astype('float64')
    self.df['Target'] = self.df['Target'].astype('float64')

    # Primerla letra del id
    self.df['Product ID'] = self.df['Product ID'].apply(lambda x: x[1:])
    self.df['Product ID'] = pd.to_numeric(self.df['Product ID'])
    sns.histplot(data=self.df, x='Product ID', hue='Type')
    plt.show()


    # Porcentaje
    value = self.df['Type'].value_counts()
    porcentaje = 100*value/self.df.Type.shape[0]
    labels = porcentaje.index.array
    x = porcentaje.array
    plt.pie(x, labels = labels, colors=sns.color_palette('tab10')[0:3], autopct='%.0f%%')
    plt.title('Porcentaje de caldiad del producto')
    plt.show()
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com

```
# Tipo de Distribucion

encabezados = [col for col in self.df.columns
          if self.df[col].dtype=='float64' or col =='Type']
nuevosEncabezados = [header for header in encabezados if  self.df[header].dtype=='float64']
fig, axs = plt.subplots(nrows=2, ncols=3, figsize=(30,12))
fig.suptitle('Caracteristicas en Histograma')
for j, feature in enumerate(nuevosEncabezados):
    sns.histplot(ax=axs[j//3, j-3*(j//3)], data= self.df, x=feature)
plt.show()

# boxplot
fig, axs = plt.subplots(nrows=2, ncols=3, figsize=(30,12))
fig.suptitle('Caracteriticas en Boxplot')
for j, encab in enumerate(nuevosEncabezados):
    sns.boxplot(ax=axs[j//3, j-3*(j//3)], data=self.df, x=feature)
plt.show()

idx_fail = self.df.loc[self.df['Failure Type'] != 'No Failure'].index
df_fail = self.df.loc[idx_fail]
df_fail_percentage = 100*df_fail['Failure Type'].value_counts()/df_fail['Failure Type'].shape[0]
# Pie plot
plt.title('Diferentes fallas que presentan las maquinas')
plt.pie(x=df_fail_percentage.array, labels=df_fail_percentage.index.array,
        colors=sns.color_palette('tab10')[0:4], autopct='%.0f%%')
plt.show()

fig, axs = plt.subplots(nrows=2, ncols=3, figsize=(30,12))
fig.suptitle('Correlacion de los atributos numericos con las fallas')
custom_palette = {'L', 'M', 'H'}
for j, feature in enumerate(nuevosEncabezados):
    sns.kdeplot(ax=axs[j//3, j-3*(j//3)], data=self.df, x=feature,
                hue='Type', fill=True)
plt.show()

fig, axs = plt.subplots(nrows=2, ncols=3, figsize=(30,12))
fig.suptitle('Correlacion de los atributos numericos con la etiqueta de fallo(1) o no fallo(0)')
enumerate_features = enumerate(nuevosEncabezados)
for j, feature in enumerate_features:
    sns.kdeplot(ax=axs[j//3, j-3*(j//3)], data=self.df, x=feature,
                 hue='Target', fill=True)
plt.show()
return
```

```
    def __str__(self):
        self.plotAll()

        returnTypes = ""
        for i in self.min:
            returnTypes += (i +":"+ str(type(self.min[i])) + "\n")
        returnObservations = ""
        for i in self.observations:
            returnObservations += (i +":"+ str(self.observations[i]) + "\n")
        returnInstanceOf = ""
        for i in self.numberOfInstanceOfAtrr:
            returnInstanceOf += (i +":"+ str(self.numberOfInstanceOfAtrr[i]) + "\n")
        returnEmpyInstance = ""
        for i in self.numberOfEmptyInstanceOfAtrr:
            returnEmpyInstance += (i +":"+ str(self.numberOfEmptyInstanceOfAtrr[i]) + "\n")
        returnMin = ""
        for i in self.min:
            returnMin += (i +":"+ str(self.min[i]) + "\n" if (type(self.min[i]) != str) else "")
        returnMax = ""
        for i in self.max:
            returnMax += (i +":"+ str(self.max[i]) + "\n" if (type(self.max[i]) != str) else "")
        returnMean = ""
        for i in self.mean:
            returnMean += (i +":"+ str(self.mean[i]) + "\n" if (type(self.mean[i]) != str) else "")
        returnVar = ""
        for i in self.var:
            returnVar += (i +":"+ str(self.var[i]) + "\n" if (type(self.var[i]) != str) else "")
        returnStd = ""
        for i in self.std:
            returnStd += (i +":"+ str(self.std[i]) + "\n" if (type(self.std[i]) != str) else "")
        return ("El número de atributos es: " + str(self.numeroAtributos()) + "\n\n" +
                "Los tipos son (Atributo: valor): \n" + returnTypes + "\n\n" +
                "El numero de instancias por atributo es (Atributo: valor): \n" + returnInstanceOf + "\n\n" +
                "El numero de observaciones por atributo es (Atributo: valor): \n" + returnObservations + "\n\n" +
                "El numero de datos faltantes por atributo es (Atributo: valor): \n" + returnEmpyInstance + "\n\n" +
                "Los valores mínimos son (Atributo: valor): \n" + returnMin + "\n\n" +
                "Los valores máximos son (Atributo: valor): \n" + returnMax + "\n\n" +
                "Los valores desviación estandar son (Atributo: valor): \n" + returnStd + "\n\n" +
                "Los valores de variación son (Atributo: valor): \n" + returnVar + "\n\n" +
                "Los valores promedio son (Atributo: valor): \n" + returnMean + "\n\n")



url = 'https://drive.google.com/file/d/1_dElPy5hHvvgqvO9qYgx3JfaGJmXP8cz/view?usp=sharing'
path = 'https://drive.google.com/uc?export=download&id='+url.split('/')[-2]

dfCaracteristicas = Estadistica(pd.read_csv(path))
a = dfCaracteristicas.desc()
print(dfCaracteristicas)
```

LI. Antonio García Gutiérrez
Expediente: 261724
antonigarga34@gmail.com