Tarea_1_Antonio_Bustos

April 29, 2025

0.1 Tarea 1

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn
import scipy
from scipy.stats import nbinom
import seaborn as sns
from statsmodels.iolib.summary2 import summary_col

import warnings
warnings.filterwarnings("ignore")
```

[2]:
```
def viento(valor):
    norte = ['N', 'NNE', 'NNW']
    sur = ['S', 'SSE', 'SSW']
    este = ['E', 'ENE', 'ESE', 'NE', 'SE']
    oeste = ['W', 'WNW', 'WSW', 'NW', 'SW']

    if valor in norte:
        return 'N'
    elif valor in sur:
        return 'S'
    elif valor in este:
        return 'E'
    elif valor in oeste:
        return 'W'

def log(valor):
    e = 1
    valor = valor + e
    valor = np.log(valor)
    return valor

def no_valor(valor):
```

```
    if valor==0:
        return 1
    else:
        return 0

def bin(valor):
    if valor=='Yes':
        return 1
    else:
        return 0

def trimestre(valor):
    t1 = ['January', 'Febrary', 'March']
    t2 = ['April', 'May', 'June']
    t3 = ['July', 'August', 'September']
    t4 = ['October', 'November', 'December']
    if valor in t1:
        return 't1'
    elif valor in t2:
        return 't2'
    elif valor in t3:
        return 't3'
    else:
        return 't4'
```

[3]:
```
df_original = pd.read_csv('../../data/machine_failure_data.csv', delimiter =␣
 ↪',', decimal = ',')
df_original.dtypes
```

[3]:
```
Date                  object
Location               int64
Min_Temp              object
Max_Temp              object
Leakage               object
Evaporation           object
Electricity           object
Parameter1_Dir        object
Parameter1_Speed     float64
Parameter2_9am        object
Parameter2_3pm        object
Parameter3_9am       float64
Parameter3_3pm       float64
Parameter4_9am       float64
Parameter4_3pm       float64
Parameter5_9am        object
Parameter5_3pm        object
Parameter6_9am       float64
```

```
Parameter6_3pm      float64
Parameter7_9am       object
Parameter7_3pm       object
Failure_today        object
dtype: object
```

0.2 Limpieza de Datos

[4]:
```
df = df_original

# Se elimina el Parámetro 6 que contiene muy pocos valores

df = df.drop('Parameter6_9am', axis=1)
df = df.drop('Parameter6_3pm', axis=1)

# Aplicamos el cuadrante en la dirección del viento

df['Parameter1_Dir'] = df['Parameter1_Dir'].apply(viento)
df['Parameter2_3pm'] = df['Parameter2_3pm'].apply(viento)
df['Parameter2_9am'] = df['Parameter2_9am'].apply(viento)

# Estandarizamos el Parámetro 5

df['Parameter5_9am'] = pd.to_numeric(df['Parameter5_9am'], errors='coerce')
df['Parameter5_3pm'] = pd.to_numeric(df['Parameter5_3pm'], errors='coerce')
df = df.dropna(subset=['Parameter5_9am'])
df = df.dropna(subset=['Parameter5_3pm'])

mp5_9 = np.mean(df['Parameter5_9am'])
mp5_3 = np.mean(df['Parameter5_3pm'])
sp5_9 = np.std(df['Parameter5_9am'])
sp5_3 = np.std(df['Parameter5_3pm'])

df['Parameter5_9am'] = (df['Parameter5_9am'] - mp5_9) / sp5_9
df['Parameter5_3pm'] = (df['Parameter5_3pm'] - mp5_3) / sp5_3

# Sumamos un e = 1 a todos los valores de Leakage y aplicamos Log

df['Leakage'] = pd.to_numeric(df['Leakage'], errors='coerce')
df = df.dropna(subset=['Leakage'])

df['Leakage'] = df['Leakage'].apply(log)

# Trabajamos la variable Evaporation y Electricity

df['Evaporation'] = pd.to_numeric(df['Evaporation'])
df['Electricity'] = pd.to_numeric(df['Electricity'])
```

```
df['Evaporation'] = df['Evaporation'].fillna(0)
df['Electricity'] = df['Electricity'].fillna(0)

df['No_Evaporation'] = df['Evaporation'].apply(no_valor)
df['No_Electricity'] = df['Electricity'].apply(no_valor)

# Trabajamos Failure

df['Failure_today'] = df['Failure_today'].apply(bin)

# Hacemos una columna con los Meses

df['Date'] = pd.to_datetime(df['Date'])
df['Month'] = df['Date'].dt.month_name()

# Cambio de Nombres

df.rename(columns = {'Parameter1_Dir':'Dir_Viento',
                     'Parameter1_Speed':'Vel_Viento',
                     'Parameter2_9am':'Dir_Mañana',
                     'Parameter2_3pm':'Dir_Tarde',
                     'Parameter7_9am':'Temp_Mañana',
                     'Parameter7_3pm':'Temp_Tarde',
                     'Parameter3_9am':'Vel_Mañana',
                     'Parameter3_3pm':'Vel_Tarde'
                     },
                     inplace=True)

# Se limpian las Temperaturas de Mañana y Tarde

df = df.dropna(subset=['Temp_Mañana'])
df = df.dropna(subset=['Temp_Tarde'])
df['Temp_Mañana'] = pd.to_numeric(df['Temp_Mañana'], errors='coerce')
df['Temp_Tarde'] = pd.to_numeric(df['Temp_Tarde'], errors='coerce')

# Filtramos los datos desde 2009 en adelante

df = df[df['Date'].dt.year >= 2009]

# Hacemos Dummies las variables categóricas

dummie = pd.get_dummies(df, columns=['Dir_Mañana', 'Dir_Tarde',␣
 ↪'Dir_Viento','Month'])
dummie = dummie.replace({True: 1, False:0})

# Aplicar columna con Trimestres
```

```python
df['Trimestre'] = df['Month'].apply(trimestre)

# Limpiar Temp Máximas y Mínimas

df = df.dropna(subset=['Min_Temp'])
df = df.dropna(subset=['Max_Temp'])
df['Min_Temp'] = pd.to_numeric(df['Min_Temp'], errors='coerce')
df['Max_Temp'] = pd.to_numeric(df['Max_Temp'], errors='coerce')

# Limpiamos datos que no estan

df = df.dropna(subset=['Dir_Viento'])
df = df.dropna(subset=['Vel_Viento'])
df = df.dropna(subset=['Dir_Mañana'])
df = df.dropna(subset=['Dir_Tarde'])
df = df.dropna(subset=['Vel_Mañana'])
df = df.dropna(subset=['Vel_Tarde'])
df = df.dropna(subset=['Parameter4_9am'])
df = df.dropna(subset=['Parameter4_3pm'])

df.head(10)
```

[4]:

```
          Date  Location  Min_Temp  Max_Temp  Leakage  Evaporation  \
30  2009-01-01         3      11.3      26.5      0.0          0.0
31  2009-01-02         3       9.6      23.9      0.0          0.0
32  2009-01-03         3      10.5      28.8      0.0          0.0
33  2009-01-04         3      12.3      34.6      0.0          0.0
34  2009-01-05         3      12.9      35.8      0.0          0.0
35  2009-01-06         3      13.7      37.9      0.0          0.0
36  2009-01-07         3      16.1      38.9      0.0          0.0
37  2009-01-08         3      14.0      28.3      0.0          0.0
38  2009-01-09         3      12.5      28.4      0.0          0.0
39  2009-01-10         3      17.0      30.8      0.0          0.0

    Electricity Dir_Viento  Vel_Viento Dir_Mañana  … Parameter4_3pm  \
30          0.0          W        56.0          W  …           26.0
31          0.0          W        41.0          W  …           22.0
32          0.0          S        26.0          S  …           22.0
33          0.0          W        37.0          S  …           12.0
34          0.0          W        41.0          E  …            9.0
35          0.0          W        52.0          E  …            8.0
36          0.0          W        57.0          E  …           12.0
37          0.0          W        48.0          W  …           15.0
38          0.0          E        37.0          S  …           16.0
39          0.0          E        37.0          N  …           24.0
```

```
    Parameter5_9am  Parameter5_3pm  Temp_Mañana  Temp_Tarde  Failure_today  \
30       -1.851431       -1.713335         19.7        25.7              0
31       -0.458103       -0.306563         14.9        22.1              0
32        0.147080       -0.064995         17.1        26.5              0
33       -0.359584       -0.704438         20.7        33.9              0
34       -0.711435       -0.860746         22.4        34.4              0
35       -0.950693       -1.215991         23.1        36.8              0
36       -1.499580       -1.784385         25.2        38.4              0
37       -0.809953       -0.619179         17.9        27.6              0
38        0.020414       -0.221304         17.2        26.6              0
39       -0.598843       -1.017054         20.2        29.3              0

    No_Evaporation  No_Electricity    Month Trimestre
30               1               1  January        t1
31               1               1  January        t1
32               1               1  January        t1
33               1               1  January        t1
34               1               1  January        t1
35               1               1  January        t1
36               1               1  January        t1
37               1               1  January        t1
38               1               1  January        t1
39               1               1  January        t1

[10 rows x 24 columns]
```

Lo siguiente es para ver cuantos valores tienen las columnas de Electricity y Evaporation para ver si es bueno hacer variables indicadoras

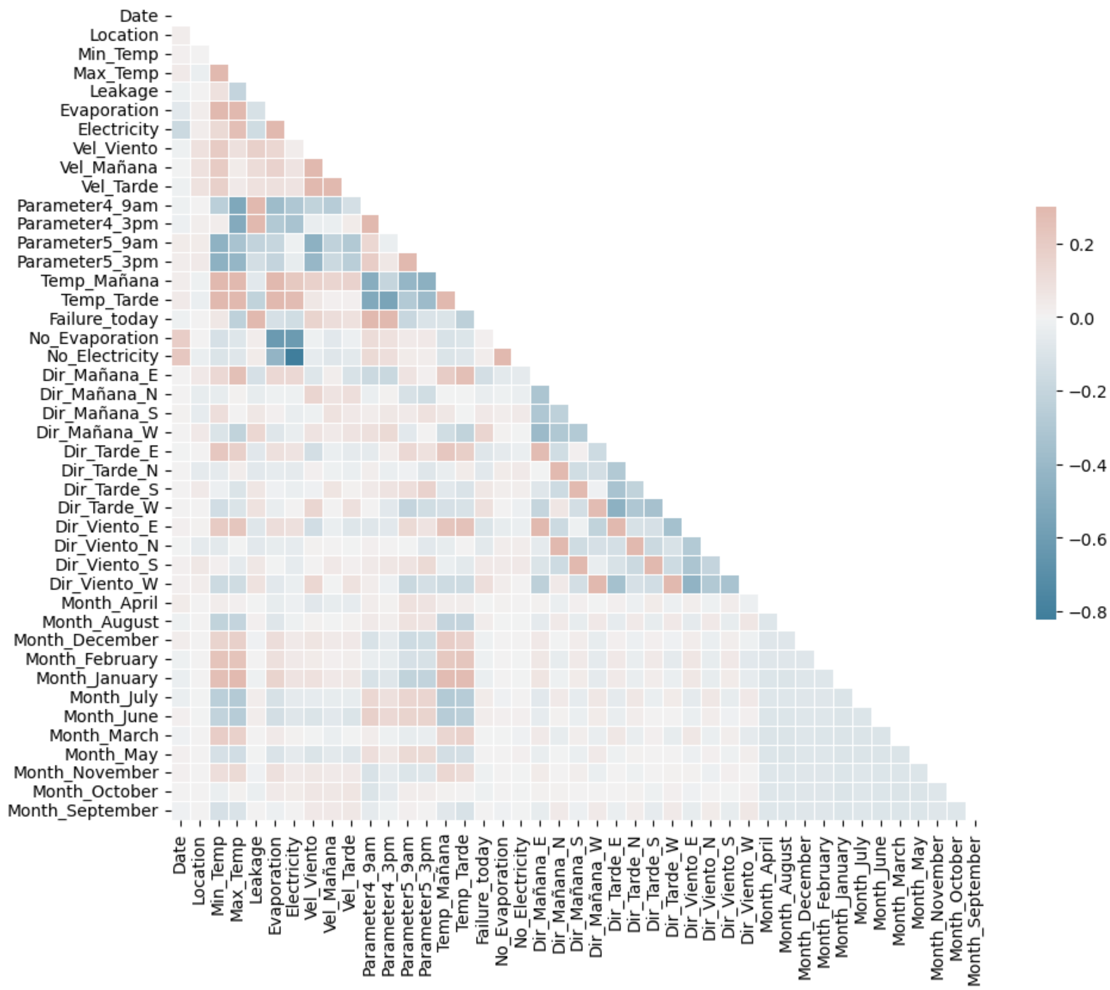
[5]:
```python
df['Elec'] = df['No_Electricity'].apply(no_valor)
df['Evap'] = df['No_Evaporation'].apply(no_valor)

group = df.groupby(['Location']).agg({
    'Elec': 'sum',
    'Evap': 'sum'
}).reset_index()

group.head(10)

no_elec = group[group['Elec'] == 0]['Location'].unique()
no_evap = group[group['Evap'] == 0]['Location'].unique()

total = len(df['Elec'])
sum_elec = np.sum(group['Elec'])
sum_evap = np.sum(group['Evap'])
p_elec = round((sum_elec/total)*100,2)
p_evap = round((sum_evap/total)*100,2)
```

```
print(f'Los sensores que no tienen valores en Electricity son: {no_elec}.\nLos␣
↪sensore que no tienen valores en Evaporation son: {no_evap}')
print(f'El {p_elec}% de los datos tienen valores en Electricity\nEl {p_evap}%␣
↪de los datos tienen valores en Evaporation')
```

```
Los sensores que no tienen valores en Electricity son: [ 3  5  6  7 15 17 18 26
27 35 41 42 44 47 48].
Los sensore que no tienen valores en Evaporation son: [ 3  5  6 15 26 27 30 41
42 44 47 48]
El 56.81% de los datos tienen valores en Electricity
El 63.02% de los datos tienen valores en Evaporation
```

[6]:
```
corr = dummie.corr()

mask = np.triu(np.ones_like(corr, dtype=bool))
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(230, 20, as_cmap=True)
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

[6]: <Axes: >



2. Ejecute un modelo de probabilidad lineal (MCO) que permita explicar la probabilidad de que un dia se reporte fallo medido por sensor, a partir de las informacion disponible. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: La variables 'Parameter4_9am', 'Parameter4_3pm' y 'Leakage' tienen mucha covariación con la variable dependiente por lo que las excluimos del modelo

Además,

- Se excluye la variable 'Vel_Tarde' porque tiene mucha correlación con 'Vel_Mañana' por lo que probablemente expliquen lo mismo
- Se excluye la variable 'Parameter5_3pm' porque tiene mucha correlación con 'Parameter5_9am' por lo que probablemente expliquen lo mismo

Luego, se puede ver que hay variables categóricas que no son significativas, sobre todo las del ID del sensor (Location), esto se puede deber a que existían muchas celdas que no tenían valores, por lo que no aporta de tan buena forma a la regresión.

Además, se hicieron interacciones entre las variables de velocidad y dirección del viento pero no se llegó a nada bueno por lo que se decidió dejar las variables de forma lineal.

0.3 OLS

[7]:
```
ols = smf.ols("Failure_today ~ C(Dir_Mañana, Treatment(reference='N')) +␣
↪Vel_Mañana + C(Dir_Tarde, Treatment(reference='N')) + No_Electricity +␣
↪Min_Temp + Temp_Mañana + Temp_Tarde + C(Location) + C(Trimestre,␣
↪Treatment(reference='t2')) + Parameter5_9am + Evaporation + No_Evaporation +␣
↪Electricity + No_Electricity",
                data=df).fit()
print(ols.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Failure_today   R-squared:                       0.228
Model:                            OLS   Adj. R-squared:                  0.228
Method:                 Least Squares   F-statistic:                     539.2
Date:                Thu, 24 Apr 2025   Prob (F-statistic):               0.00
Time:                        21:38:42   Log-Likelihood:                -46188.
No. Observations:              111179   AIC:                         9.250e+04
Df Residuals:                  111117   BIC:                         9.310e+04
Df Model:                          61
Covariance Type:            nonrobust
================================================================================
=================================
                                                    coef    std err          t
P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
---------------------------------
Intercept                                         0.6371      0.010     60.754
0.000       0.617       0.658
C(Dir_Mañana, Treatment(reference='N'))[T.E]      0.0069      0.004      1.924
0.054      -0.000       0.014
C(Dir_Mañana, Treatment(reference='N'))[T.S]      0.0305      0.004      7.464
0.000       0.022       0.038
C(Dir_Mañana, Treatment(reference='N'))[T.W]      0.0472      0.004     13.322
0.000       0.040       0.054
C(Dir_Tarde, Treatment(reference='N'))[T.E]       0.0095      0.004      2.492
0.013       0.002       0.017
C(Dir_Tarde, Treatment(reference='N'))[T.S]       0.0456      0.004     10.973
0.000       0.037       0.054
C(Dir_Tarde, Treatment(reference='N'))[T.W]       0.0507      0.004     13.879
0.000       0.044       0.058
C(Location)[T.3]                                  0.0084      0.011      0.782
0.434      -0.013       0.030
C(Location)[T.4]                                  0.1255      0.011     11.815
0.000       0.105       0.146
C(Location)[T.5]                                  0.0242      0.011      2.219
```

```
0.026        0.003        0.046
C(Location)[T.6]                                              -0.0504      0.011     -4.741
0.000       -0.071       -0.030
C(Location)[T.7]                                              -0.0361      0.010     -3.434
0.001       -0.057       -0.015
C(Location)[T.8]                                               0.0774      0.011      7.348
0.000        0.057        0.098
C(Location)[T.9]                                               0.1077      0.011      9.984
0.000        0.087        0.129
C(Location)[T.10]                                              0.0143      0.011      1.337
0.181       -0.007        0.035
C(Location)[T.11]                                             -0.0447      0.011     -4.253
0.000       -0.065       -0.024
C(Location)[T.12]                                              0.0841      0.011      7.907
0.000        0.063        0.105
C(Location)[T.13]                                              0.0397      0.011      3.691
0.000        0.019        0.061
C(Location)[T.14]                                              0.0670      0.011      6.248
0.000        0.046        0.088
C(Location)[T.15]                                              0.0057      0.011      0.534
0.593       -0.015        0.027
C(Location)[T.16]                                             -0.1250      0.010    -11.971
0.000       -0.145       -0.105
C(Location)[T.17]                                              0.0680      0.017      4.028
0.000        0.035        0.101
C(Location)[T.18]                                             -0.0493      0.012     -4.009
0.000       -0.073       -0.025
C(Location)[T.19]                                             -0.0591      0.011     -5.256
0.000       -0.081       -0.037
C(Location)[T.20]                                             -0.0534      0.010     -5.130
0.000       -0.074       -0.033
C(Location)[T.21]                                             -0.0278      0.010     -2.714
0.007       -0.048       -0.008
C(Location)[T.22]                                              0.0621      0.010      5.975
0.000        0.042        0.082
C(Location)[T.23]                                              0.0451      0.010      4.389
0.000        0.025        0.065
C(Location)[T.26]                                             -0.0887      0.012     -7.211
0.000       -0.113       -0.065
C(Location)[T.27]                                             -0.0599      0.011     -5.708
0.000       -0.081       -0.039
C(Location)[T.28]                                             -0.0310      0.010     -2.975
0.003       -0.051       -0.011
C(Location)[T.29]                                              0.0249      0.010      2.412
0.016        0.005        0.045
C(Location)[T.30]                                              0.0931      0.011      8.260
0.000        0.071        0.115
C(Location)[T.32]                                              0.0568      0.010      5.550
```

```
0.000       0.037       0.077
C(Location)[T.33]                                 0.0752      0.010      7.334
0.000       0.055       0.095
C(Location)[T.34]                                 0.0222      0.010      2.161
0.031       0.002       0.042
C(Location)[T.35]                                 0.0314      0.011      2.798
0.005       0.009       0.053
C(Location)[T.36]                                -0.0456      0.010     -4.421
0.000      -0.066      -0.025
C(Location)[T.38]                                -0.0237      0.011     -2.159
0.031      -0.045      -0.002
C(Location)[T.39]                                -0.0123      0.010     -1.190
0.234      -0.033       0.008
C(Location)[T.40]                                 0.0232      0.011      2.163
0.031       0.002       0.044
C(Location)[T.41]                                -0.0145      0.011     -1.321
0.187      -0.036       0.007
C(Location)[T.42]                                 0.0013      0.013      0.105
0.917      -0.024       0.026
C(Location)[T.43]                                 0.0331      0.010      3.218
0.001       0.013       0.053
C(Location)[T.44]                                 0.0048      0.011      0.445
0.656      -0.016       0.026
C(Location)[T.45]                                -0.0155      0.010     -1.492
0.136      -0.036       0.005
C(Location)[T.46]                                 0.0837      0.011      7.677
0.000       0.062       0.105
C(Location)[T.47]                                 0.0526      0.011      4.818
0.000       0.031       0.074
C(Location)[T.48]                                -0.1113      0.011    -10.556
0.000      -0.132      -0.091
C(Location)[T.49]                                -0.0484      0.010     -4.695
0.000      -0.069      -0.028
C(Trimestre, Treatment(reference='t2'))[T.t1]     0.0173      0.004      4.210
0.000       0.009       0.025
C(Trimestre, Treatment(reference='t2'))[T.t3]     0.0130      0.003      4.001
0.000       0.007       0.019
C(Trimestre, Treatment(reference='t2'))[T.t4]     0.0345      0.003      9.980
0.000       0.028       0.041
Vel_Mañana                                        0.0016      0.000     10.195
0.000       0.001       0.002
No_Electricity                                   -0.0117      0.006     -2.009
0.045      -0.023      -0.000
Min_Temp                                          0.0389      0.001     77.822
0.000       0.038       0.040
Temp_Mañana                                      -0.0295      0.001    -44.914
0.000      -0.031      -0.028
Temp_Tarde                                       -0.0179      0.000    -42.400
```

```
0.000      -0.019      -0.017
Parameter5_9am                                          -0.0789      0.001     -54.669
0.000      -0.082      -0.076
Evaporation                                             -0.0107      0.000     -26.295
0.000      -0.012      -0.010
No_Evaporation                                          -0.0557      0.006     -10.014
0.000      -0.067      -0.045
Electricity                                             -0.0070      0.000     -14.467
0.000      -0.008      -0.006
==============================================================================
Omnibus:                    10624.659   Durbin-Watson:                   1.710
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13484.508
Skew:                           0.840   Prob(JB):                         0.00
Kurtosis:                       2.699   Cond. No.                     1.63e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3. Ejecute un modelo *probit* para responder a la pregunta 2. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Corriendo el mismo modelo anterior en Probit, se logró una mejor explicación de los datos.

Además, se puede ver que practicamente las mismas variables categóricas del ID del sensor no son significativa por lo que podemos suponer lo mismo que se concluyó anteriormente.

Un cambio importante es que la variable No_Electricity que es una variable indicadora (1 si no hay datos en Electricity) no es significativa, lo que puede deberse a que tiene alta correlación con la variable No_Evaporation o porque en este modelo no aporta información relevante

0.4 Probit

[8]:
```python
probit = smf.probit("Failure_today ~ C(Dir_Mañana, Treatment(reference='N')) +␣
 ↪Vel_Mañana + C(Dir_Tarde, Treatment(reference='N')) + No_Electricity +␣
 ↪Min_Temp + Temp_Mañana + Temp_Tarde + C(Location) + C(Trimestre,␣
 ↪Treatment(reference='t2')) + Parameter5_9am + Evaporation + No_Evaporation +␣
 ↪Electricity + No_Electricity",
                 data=df).fit()
print(probit.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.394917
         Iterations 7
                          Probit Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:               111179
```

```
Model:                         Probit   Df Residuals:                   111117
Method:                           MLE   Df Model:                           61
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.2587
Time:                        21:38:46   Log-Likelihood:                -43907.
converged:                       True   LL-Null:                       -59226.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=================================
                                                      coef    std err          z
P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
---------------------------------
Intercept                                           1.1146      0.047     23.540
0.000       1.022       1.207
C(Dir_Mañana, Treatment(reference='N'))[T.E]        0.0055      0.017      0.324
0.746      -0.028       0.039
C(Dir_Mañana, Treatment(reference='N'))[T.S]        0.1320      0.018      7.339
0.000       0.097       0.167
C(Dir_Mañana, Treatment(reference='N'))[T.W]        0.1732      0.015     11.341
0.000       0.143       0.203
C(Dir_Tarde, Treatment(reference='N'))[T.E]         0.0965      0.018      5.364
0.000       0.061       0.132
C(Dir_Tarde, Treatment(reference='N'))[T.S]         0.1987      0.019     10.479
0.000       0.162       0.236
C(Dir_Tarde, Treatment(reference='N'))[T.W]         0.2196      0.017     12.986
0.000       0.186       0.253
C(Location)[T.3]                                    0.1214      0.048      2.534
0.011       0.028       0.215
C(Location)[T.4]                                    0.4009      0.058      6.955
0.000       0.288       0.514
C(Location)[T.5]                                    0.2355      0.048      4.910
0.000       0.141       0.329
C(Location)[T.6]                                   -0.2284      0.046     -4.932
0.000      -0.319      -0.138
C(Location)[T.7]                                   -0.0990      0.048     -2.079
0.038      -0.192      -0.006
C(Location)[T.8]                                    0.5737      0.046     12.514
0.000       0.484       0.664
C(Location)[T.9]                                    0.6823      0.046     14.754
0.000       0.592       0.773
C(Location)[T.10]                                   0.1331      0.048      2.754
0.006       0.038       0.228
C(Location)[T.11]                                  -0.0682      0.050     -1.362
0.173      -0.166       0.030
C(Location)[T.12]                                   0.5123      0.045     11.325
0.000       0.424       0.601
C(Location)[T.13]                                   0.0870      0.046      1.898
0.058      -0.003       0.177
```

C(Location)[T.14]	0.6178	0.047	13.209	0.000	0.526	0.709
C(Location)[T.15]	0.2944	0.046	6.452	0.000	0.205	0.384
C(Location)[T.16]	-0.4743	0.045	-10.536	0.000	-0.563	-0.386
C(Location)[T.17]	0.6599	0.081	8.197	0.000	0.502	0.818
C(Location)[T.18]	-0.0993	0.052	-1.895	0.058	-0.202	0.003
C(Location)[T.19]	-0.1996	0.048	-4.181	0.000	-0.293	-0.106
C(Location)[T.20]	-0.1973	0.046	-4.331	0.000	-0.287	-0.108
C(Location)[T.21]	-0.2157	0.050	-4.275	0.000	-0.315	-0.117
C(Location)[T.22]	0.4571	0.049	9.235	0.000	0.360	0.554
C(Location)[T.23]	0.1493	0.044	3.368	0.001	0.062	0.236
C(Location)[T.26]	-0.3775	0.058	-6.554	0.000	-0.490	-0.265
C(Location)[T.27]	-0.1279	0.044	-2.923	0.003	-0.214	-0.042
C(Location)[T.28]	0.0016	0.043	0.037	0.970	-0.083	0.086
C(Location)[T.29]	-0.0009	0.048	-0.018	0.986	-0.094	0.093
C(Location)[T.30]	0.4732	0.051	9.242	0.000	0.373	0.574
C(Location)[T.32]	0.3155	0.047	6.718	0.000	0.223	0.408
C(Location)[T.33]	0.4335	0.047	9.197	0.000	0.341	0.526
C(Location)[T.34]	0.0128	0.043	0.296	0.767	-0.072	0.097
C(Location)[T.35]	0.2677	0.049	5.421	0.000	0.171	0.364
C(Location)[T.36]	-0.1112	0.045	-2.460	0.014	-0.200	-0.023
C(Location)[T.38]	0.0458	0.046	0.994	0.320	-0.044	0.136
C(Location)[T.39]	0.1174	0.044	2.663	0.008	0.031	0.204
C(Location)[T.40]	0.4071	0.049	8.364	0.000	0.312	0.502
C(Location)[T.41]	0.0137	0.049	0.281	0.778	-0.082	0.109

```
C(Location)[T.42]                                    0.0107      0.069      0.154
0.878      -0.125       0.147
C(Location)[T.43]                                    0.1429      0.048      2.984
0.003       0.049       0.237
C(Location)[T.44]                                    0.0605      0.045      1.354
0.176      -0.027       0.148
C(Location)[T.45]                                   -0.1032      0.045     -2.277
0.023      -0.192      -0.014
C(Location)[T.46]                                    0.5536      0.047     11.860
0.000       0.462       0.645
C(Location)[T.47]                                    0.3093      0.047      6.583
0.000       0.217       0.401
C(Location)[T.48]                                   -0.3072      0.044     -6.910
0.000      -0.394      -0.220
C(Location)[T.49]                                   -0.2940      0.054     -5.415
0.000      -0.400      -0.188
C(Trimestre, Treatment(reference='t2'))[T.t1]        0.1543      0.018      8.584
0.000       0.119       0.190
C(Trimestre, Treatment(reference='t2'))[T.t3]        0.0312      0.014      2.242
0.025       0.004       0.059
C(Trimestre, Treatment(reference='t2'))[T.t4]        0.2156      0.015     14.209
0.000       0.186       0.245
Vel_Mañana                                           0.0021      0.001      3.231
0.001       0.001       0.003
No_Electricity                                      -0.0349      0.023     -1.489
0.137      -0.081       0.011
Min_Temp                                             0.2084      0.003     79.249
0.000       0.203       0.214
Temp_Mañana                                         -0.1645      0.003    -49.992
0.000      -0.171      -0.158
Temp_Tarde                                          -0.0901      0.002    -44.404
0.000      -0.094      -0.086
Parameter5_9am                                      -0.2868      0.006    -46.804
0.000      -0.299      -0.275
Evaporation                                         -0.0660      0.002    -31.010
0.000      -0.070      -0.062
No_Evaporation                                      -0.2621      0.024    -10.925
0.000      -0.309      -0.215
Electricity                                         -0.0189      0.002     -8.978
0.000      -0.023      -0.015
================================================================================
=================================
```

4. Ejecute un modelo *logit* para responder a la pregunta 2. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R Corriendo el modelo anterior, podemos obtener resultados muy parecidos a los anteriores en temas de significancia de las variables, pero este modelo explica mejor la data.

En este caso, nuevamente las variables categóricas de Location no son significativas y la variable No_Electricity, lo que puede deberse a lo explicado en el punto anterior.

0.5 Logit

```
[9]: logit = smf.logit("Failure_today ~ C(Dir_Mañana, Treatment(reference='N')) +␣
      ↪Vel_Mañana + C(Dir_Tarde, Treatment(reference='N')) + No_Electricity +␣
      ↪Min_Temp + Temp_Mañana + Temp_Tarde + C(Location) + C(Trimestre,␣
      ↪Treatment(reference='t2')) + Parameter5_9am + Evaporation + No_Evaporation +␣
      ↪Electricity + No_Electricity",
                      data=df).fit()
     print(logit.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.393421
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:               111179
Model:                          Logit   Df Residuals:                   111117
Method:                           MLE   Df Model:                           61
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.2615
Time:                        21:38:50   Log-Likelihood:                -43740.
converged:                       True   LL-Null:                       -59226.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=================================
                                                    coef    std err          z
P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
---------------------------------
Intercept                                         2.1309      0.083     25.602
0.000       1.968       2.294
C(Dir_Mañana, Treatment(reference='N'))[T.E]      0.0186      0.031      0.610
0.542      -0.041       0.079
C(Dir_Mañana, Treatment(reference='N'))[T.S]      0.2434      0.032      7.658
0.000       0.181       0.306
C(Dir_Mañana, Treatment(reference='N'))[T.W]      0.3136      0.027     11.665
0.000       0.261       0.366
C(Dir_Tarde, Treatment(reference='N'))[T.E]       0.1979      0.032      6.148
0.000       0.135       0.261
C(Dir_Tarde, Treatment(reference='N'))[T.S]       0.3643      0.034     10.843
0.000       0.298       0.430
C(Dir_Tarde, Treatment(reference='N'))[T.W]       0.3918      0.030     13.009
0.000       0.333       0.451
C(Location)[T.3]                                  0.2037      0.084      2.436
0.015       0.040       0.368
C(Location)[T.4]                                  0.6901      0.106      6.529
```

				0.000	0.483	0.897
C(Location)[T.5]	0.4091	0.084	4.879	0.000	0.245	0.573
C(Location)[T.6]	-0.4424	0.080	-5.510	0.000	-0.600	-0.285
C(Location)[T.7]	-0.2027	0.083	-2.432	0.015	-0.366	-0.039
C(Location)[T.8]	1.0426	0.080	12.952	0.000	0.885	1.200
C(Location)[T.9]	1.2705	0.080	15.870	0.000	1.114	1.427
C(Location)[T.10]	0.1852	0.086	2.159	0.031	0.017	0.353
C(Location)[T.11]	-0.1381	0.090	-1.536	0.125	-0.314	0.038
C(Location)[T.12]	0.9061	0.078	11.595	0.000	0.753	1.059
C(Location)[T.13]	0.1061	0.079	1.345	0.179	-0.049	0.261
C(Location)[T.14]	1.1802	0.082	14.420	0.000	1.020	1.341
C(Location)[T.15]	0.5469	0.079	6.934	0.000	0.392	0.701
C(Location)[T.16]	-0.8977	0.079	-11.383	0.000	-1.052	-0.743
C(Location)[T.17]	1.3333	0.143	9.344	0.000	1.054	1.613
C(Location)[T.18]	-0.1762	0.091	-1.932	0.053	-0.355	0.003
C(Location)[T.19]	-0.3611	0.083	-4.368	0.000	-0.523	-0.199
C(Location)[T.20]	-0.3655	0.079	-4.605	0.000	-0.521	-0.210
C(Location)[T.21]	-0.3977	0.090	-4.425	0.000	-0.574	-0.222
C(Location)[T.22]	0.7154	0.090	7.928	0.000	0.539	0.892
C(Location)[T.23]	0.2199	0.076	2.876	0.004	0.070	0.370
C(Location)[T.26]	-0.6751	0.102	-6.652	0.000	-0.874	-0.476
C(Location)[T.27]	-0.2366	0.075	-3.135	0.002	-0.384	-0.089
C(Location)[T.28]	0.0160	0.074	0.215	0.830	-0.130	0.162
C(Location)[T.29]	-0.0501	0.083	-0.602	0.547	-0.214	0.113
C(Location)[T.30]	0.7825	0.090	8.699			

```
0.000       0.606       0.959
C(Location)[T.32]                                     0.5951       0.082       7.278
0.000       0.435       0.755
C(Location)[T.33]                                     0.7872       0.082       9.548
0.000       0.626       0.949
C(Location)[T.34]                                    -0.0265       0.074      -0.358
0.721      -0.172       0.119
C(Location)[T.35]                                     0.4793       0.087       5.540
0.000       0.310       0.649
C(Location)[T.36]                                    -0.2438       0.079      -3.084
0.002      -0.399      -0.089
C(Location)[T.38]                                     0.1049       0.080       1.314
0.189      -0.052       0.261
C(Location)[T.39]                                     0.2253       0.077       2.944
0.003       0.075       0.375
C(Location)[T.40]                                     0.8584       0.086      10.025
0.000       0.691       1.026
C(Location)[T.41]                                    -0.0177       0.086      -0.207
0.836      -0.186       0.150
C(Location)[T.42]                                    -0.0435       0.128      -0.340
0.734      -0.294       0.207
C(Location)[T.43]                                     0.1653       0.085       1.939
0.053      -0.002       0.332
C(Location)[T.44]                                     0.0892       0.077       1.162
0.245      -0.061       0.240
C(Location)[T.45]                                    -0.2030       0.079      -2.581
0.010      -0.357      -0.049
C(Location)[T.46]                                     0.9515       0.082      11.631
0.000       0.791       1.112
C(Location)[T.47]                                     0.5259       0.081       6.482
0.000       0.367       0.685
C(Location)[T.48]                                    -0.5626       0.077      -7.276
0.000      -0.714      -0.411
C(Location)[T.49]                                    -0.5349       0.100      -5.345
0.000      -0.731      -0.339
C(Trimestre, Treatment(reference='t2'))[T.t1]         0.3008       0.032       9.508
0.000       0.239       0.363
C(Trimestre, Treatment(reference='t2'))[T.t3]         0.0498       0.024       2.065
0.039       0.003       0.097
C(Trimestre, Treatment(reference='t2'))[T.t4]         0.3928       0.027      14.660
0.000       0.340       0.445
Vel_Mañana                                            0.0038       0.001       3.304
0.001       0.002       0.006
No_Electricity                                       -0.0520       0.040      -1.289
0.197      -0.131       0.027
Min_Temp                                              0.3760       0.005      77.785
0.000       0.367       0.386
Temp_Mañana                                          -0.2981       0.006     -50.110
```

```
0.000       -0.310      -0.286
Temp_Tarde                                                  -0.1608      0.004     -43.823
0.000       -0.168      -0.154
Parameter5_9am                                              -0.4978      0.011     -46.252
0.000       -0.519      -0.477
Evaporation                                                 -0.1509      0.005     -33.415
0.000       -0.160      -0.142
No_Evaporation                                              -0.5559      0.043     -13.058
0.000       -0.639      -0.472
Electricity                                                 -0.0265      0.004      -7.112
0.000       -0.034      -0.019
================================================================================
=================================
```

5. Comente los resultados obtenidos en 2, 3 y 4. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: Según los resultados, OLS explica poco de la regresión y no lo explica como uno esperaria ya que solo toma valores continuos y puede entregar erroneamente valores binarios. Por otro lado, logit y probit son mejores en este caso, ya que explican una probabilidad, lo que favorece en este modelo porque la variable que queremos explicar es binaria.

Como Logit explica mejor el modelo, podemos decir que es la mejor opción para explicar las Fallas.

Además, muchas de las variables son robustas exceptuando:

- C(Dir_Mañana)[E] No significativa en Logit ni Probit ($p > 0.5$).
- C(Location)[T.11] No significativa en Logit ni Probit.
- C(Location)[T.13] No significativa en Logit, débil en Probit.
- C(Location)[T.18] Marginal en Logit ($p = 0.053$), no robusta.
- C(Location)[T.28] No significativa ($p = 0.830$).
- C(Location)[T.29] No significativa ($p = 0.547$).
- C(Location)[T.34] No significativa ($p = 0.721$).
- C(Location)[T.38] No significativa ($p = 0.189$).
- C(Location)[T.41] No significativa ($p = 0.836$).
- C(Location)[T.42] No significativa ($p = 0.734$).
- C(Location)[T.43] Marginal ($p = 0.053$), no robusta.
- C(Location)[T.44] No significativa ($p = 0.245$).
- No_Electricity No significativa en Logit ($p = 0.197$), débil en otros.
- C(Trimestre)[T.t3] Muy débil en Logit ($p = 0.039$), no siempre aparece significativa en OLS.

Y hay algunas con comportamientos dudosos como:

- Electricity
- No_Evaporation
- Vel_Mañana

0.6 Poisson

6. Agregue la data a nivel mensual, usando la data promedio de las variables (ignorando aquellas categoricas, como la direccion del viento). En particular, genere una variable que cuente la cantidad de fallos observados en un mes, utilice un valor de 0 si en ese mes no se reporto fallos en ningun dia. Use un modelo Poisson para explicar el numero de fallas por mes. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

Para hacer el modelo de Poisson, hacemos un nuevo df con los valores por mes en promedio y sin considerar las variables categóricas. En este caso, trabajaremos con,

1. Velocidad (Mañana y Tarde)
2. Parameter4 (Mañana y Tarde)
3. Parameter5 (Mañana y Tarde)
4. Temperatura (Mañana y Tarde)
5. Temperatura (Máxima y Mínima)
6. Location

Se decidió excluir las variables Evaporation y Electricity y sus respectivos indicadores porque tenían muy pocos datos, lo que no favorece a este tipo de análisis.

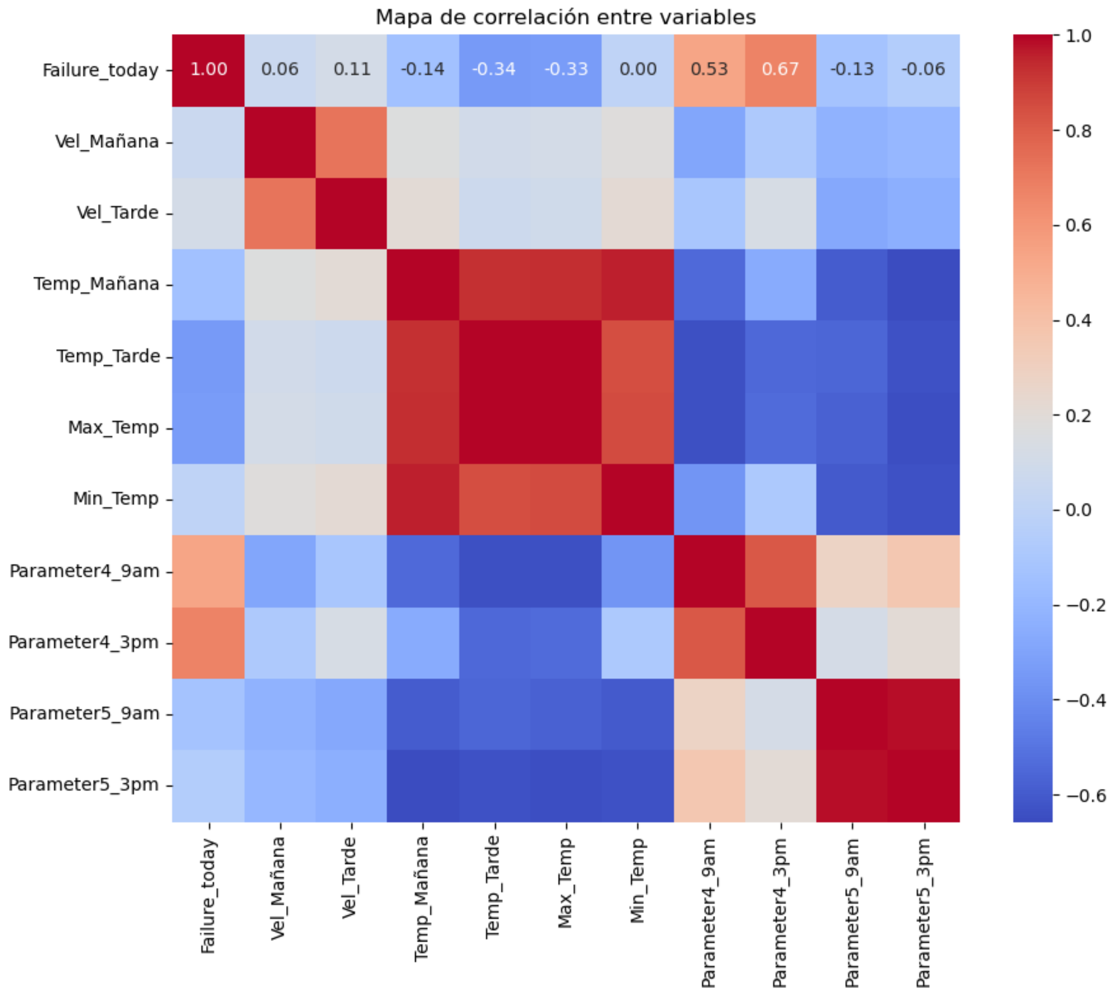
[10]:
```python
df['Periodo'] = df['Date'].dt.to_period('M')
df_mensual = df.groupby(['Periodo','Location']).agg({
    'Failure_today' : 'sum',
    'Vel_Mañana' : 'mean',
    'Vel_Tarde' : 'mean',
    'Temp_Mañana' : 'mean',
    'Temp_Tarde' : 'mean',
    'Max_Temp' : 'mean',
    'Min_Temp' : 'mean',
    'Parameter4_9am' : 'mean',
    'Parameter4_3pm' : 'mean',
    'Parameter5_9am' : 'mean',
    'Parameter5_3pm' : 'mean'
}).reset_index()

df_mensual.head(10)
```

[10]:

	Periodo	Location	Failure_today	Vel_Mañana	Vel_Tarde	Temp_Mañana	\
0	2009-01	1	0	10.448276	17.931034	23.510345	
1	2009-01	3	1	11.935484	18.548387	22.993548	
2	2009-01	4	3	18.516129	25.032258	29.241935	
3	2009-01	5	3	7.551724	17.758621	22.524138	
4	2009-01	6	0	20.172414	22.241379	18.637931	
5	2009-01	7	0	14.645161	20.645161	21.087097	
6	2009-01	8	6	10.857143	16.964286	26.721429	
7	2009-01	9	17	11.933333	14.733333	27.653333	
8	2009-01	10	4	10.322581	19.161290	20.083871	
9	2009-01	11	3	17.233333	14.600000	27.073333	

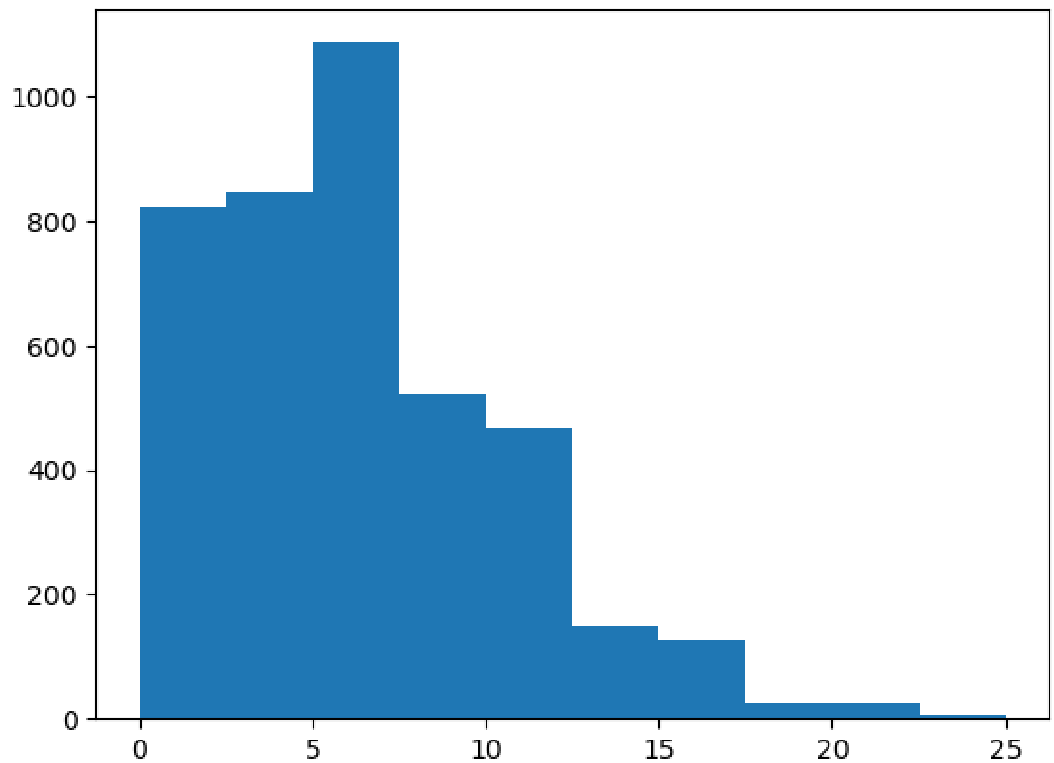
```
   Temp_Tarde   Max_Temp   Min_Temp  Parameter4_9am  Parameter4_3pm  \
0   30.579310  31.868966  17.975862       38.689655       23.827586
1   32.964516  34.658065  16.312903       41.903226       17.870968
2   34.487097  36.058065  22.422581       37.096774       24.516129
3   31.279310  32.872414  16.455172       65.724138       36.206897
4   26.589655  28.548276  10.617241       50.586207       24.379310
5   31.338710  32.864516  13.125806       42.096774       14.064516
6   28.114286  30.207143  21.285714       63.035714       56.000000
7   29.336667  30.906667  24.030000       77.500000       71.566667
8   29.322581  31.022581  14.003226       57.903226       26.516129
9   34.420000  36.623333  22.203333       37.866667       19.900000

   Parameter5_9am  Parameter5_3pm
0       -0.468294       -0.416811
1       -0.646059       -0.779612
2       -1.293918       -1.495604
3       -0.294553       -0.444251
4       -0.682316       -0.540780
5       -0.642881       -0.648515
6       -0.424928       -0.361372
7       -1.289877       -1.370405
8       -0.477171       -0.599468
9       -0.774299       -0.802012
```

[11]:
```python
df_corr = df_mensual.drop(columns=['Location','Periodo'])

corr_matrix = df_corr.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap="coolwarm", square=True)
plt.title("Mapa de correlación entre variables")
plt.tight_layout()
plt.show()
```



[12]:
```
plt.hist(df_mensual.Failure_today)
```

[12]:
```
(array([ 822.,  846., 1086.,  522.,  466.,  148.,  127.,   27.,   26.,
          6.]),
 array([ 0. ,  2.5,  5. ,  7.5, 10. , 12.5, 15. , 17.5, 20. , 22.5, 25. ]),
 <BarContainer object of 10 artists>)
```



R: De la correlación, se decidió excluir el Parameter4_3pm y Parameter5_3pm porque estaban muy correlacionados con el mismo parámetro pero en tomado en la tarde. Lo mismo pasa con las variables de Temp_Mañana y Min_Temp, y con Temp_Tarde y Max_Temp, por lo que solo dejamos las variables Temp_Mañana y Temp_Tarde.

Así, se formó la regresión viendo que la mayoría de las variables son significativas, exceptuando algunas ubicaciones del sensor.

[13]:
```
poisson = smf.poisson('Failure_today ~ Vel_Mañana + Temp_Mañana + Temp_Tarde +␣
 ↪C(Location) + Parameter4_9am + Parameter5_9am',
                      data=df_mensual).fit()
print(poisson.summary())
```

```
Optimization terminated successfully.
         Current function value: 2.288684
         Iterations 6
                          Poisson Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:                 4076
Model:                        Poisson   Df Residuals:                     4027
Method:                           MLE   Df Model:                           48
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.2856
Time:                        21:38:51   Log-Likelihood:                -9328.7
```

```
converged:                       True   LL-Null:                       -13059.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
Intercept             0.0423      0.123      0.344      0.731      -0.199
0.283
C(Location)[T.3]     -0.4002      0.063     -6.329      0.000      -0.524
-0.276
C(Location)[T.4]     -0.1579      0.081     -1.957      0.050      -0.316
0.000
C(Location)[T.5]     -0.5093      0.064     -8.018      0.000      -0.634
-0.385
C(Location)[T.6]     -0.8705      0.068    -12.833      0.000      -1.004
-0.738
C(Location)[T.7]     -0.4649      0.064     -7.312      0.000      -0.590
-0.340
C(Location)[T.8]     -0.2288      0.061     -3.726      0.000      -0.349
-0.108
C(Location)[T.9]     -0.5602      0.062     -9.006      0.000      -0.682
-0.438
C(Location)[T.10]    -0.4490      0.066     -6.802      0.000      -0.578
-0.320
C(Location)[T.11]    -0.1715      0.069     -2.489      0.013      -0.307
-0.036
C(Location)[T.12]    -0.4057      0.060     -6.745      0.000      -0.524
-0.288
C(Location)[T.13]    -0.6410      0.061    -10.567      0.000      -0.760
-0.522
C(Location)[T.14]    -0.6335      0.065     -9.815      0.000      -0.760
-0.507
C(Location)[T.15]    -0.7182      0.068    -10.546      0.000      -0.852
-0.585
C(Location)[T.16]    -0.4434      0.059     -7.507      0.000      -0.559
-0.328
C(Location)[T.17]    -0.6234      0.110     -5.653      0.000      -0.840
-0.407
C(Location)[T.18]    -0.6768      0.069     -9.755      0.000      -0.813
-0.541
C(Location)[T.19]    -0.3730      0.066     -5.692      0.000      -0.501
-0.245
C(Location)[T.20]    -0.4832      0.065     -7.401      0.000      -0.611
-0.355
C(Location)[T.21]    -0.4017      0.075     -5.380      0.000      -0.548
-0.255
```

C(Location)[T.22]	-0.3502	0.073	-4.768	0.000	-0.494	-0.206
C(Location)[T.23]	-0.4234	0.059	-7.156	0.000	-0.539	-0.307
C(Location)[T.26]	-0.5719	0.084	-6.787	0.000	-0.737	-0.407
C(Location)[T.27]	-0.6891	0.060	-11.542	0.000	-0.806	-0.572
C(Location)[T.28]	-0.7804	0.065	-12.047	0.000	-0.907	-0.653
C(Location)[T.29]	-0.3641	0.062	-5.913	0.000	-0.485	-0.243
C(Location)[T.30]	-0.3202	0.066	-4.874	0.000	-0.449	-0.191
C(Location)[T.32]	-0.1214	0.060	-2.037	0.042	-0.238	-0.005
C(Location)[T.33]	-0.0898	0.063	-1.433	0.152	-0.213	0.033
C(Location)[T.34]	-0.4702	0.057	-8.296	0.000	-0.581	-0.359
C(Location)[T.35]	-0.5924	0.065	-9.113	0.000	-0.720	-0.465
C(Location)[T.36]	-0.6297	0.063	-10.017	0.000	-0.753	-0.507
C(Location)[T.38]	-0.2543	0.061	-4.199	0.000	-0.373	-0.136
C(Location)[T.39]	-0.2909	0.061	-4.783	0.000	-0.410	-0.172
C(Location)[T.40]	-0.9440	0.069	-13.717	0.000	-1.079	-0.809
C(Location)[T.41]	-0.3853	0.064	-6.035	0.000	-0.510	-0.260
C(Location)[T.42]	-0.1673	0.108	-1.546	0.122	-0.379	0.045
C(Location)[T.43]	-0.2276	0.064	-3.554	0.000	-0.353	-0.102
C(Location)[T.44]	-0.6331	0.057	-11.120	0.000	-0.745	-0.522
C(Location)[T.45]	-0.4045	0.059	-6.837	0.000	-0.521	-0.289
C(Location)[T.46]	-0.5299	0.063	-8.351	0.000	-0.654	-0.406
C(Location)[T.47]	-0.4808	0.058	-8.242	0.000	-0.595	-0.366
C(Location)[T.48]	-0.7583	0.062	-12.158	0.000	-0.881	-0.636
C(Location)[T.49]	-0.6442	0.087	-7.371	0.000	-0.815	-0.473

```
Vel_Mañana             0.0223      0.003      7.709      0.000       0.017
0.028
Temp_Mañana            0.1507      0.007     20.886      0.000       0.137
0.165
Temp_Tarde            -0.1558      0.006    -23.966      0.000      -0.169
-0.143
Parameter4_9am         0.0374      0.001     37.552      0.000       0.035
0.039
Parameter5_9am        -0.3221      0.018    -18.275      0.000      -0.357
-0.288
================================================================================
=====
```

7. Determine sobre dispersion en la data y posible valor optimo de alpha para un modelo Binomial Negativa.

R: De aquí podemos suponer que hay sobredispersión en la data y que el valor de Alpha será de aproximadamente 1.03

[14]:
```python
y = df_mensual['Failure_today']
mu = poisson.predict()
aux=((y-mu)**2-mu)/mu
auxr=sm.OLS(aux,mu).fit()
print(auxr.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:          Failure_today   R-squared (uncentered):
0.005
Model:                            OLS   Adj. R-squared (uncentered):
0.005
Method:                 Least Squares   F-statistic:
22.37
Date:                Thu, 24 Apr 2025   Prob (F-statistic):
2.33e-06
Time:                        21:38:51   Log-Likelihood:
-7644.9
No. Observations:                4076   AIC:
1.529e+04
Df Residuals:                    4075   BIC:
1.530e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0167      0.004      4.729      0.000       0.010       0.024
==============================================================================
```

```
Omnibus:                     2989.703   Durbin-Watson:                   1.834
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            65023.910
Skew:                           3.295   Prob(JB):                         0.00
Kurtosis:                      21.424   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

[15]:
```python
alpha = np.exp(0.0273)
print(f'El valor esperado de Alpha será {alpha} y se puede ver que hay␣
 ↪sobredispersión')
```

```
El valor esperado de Alpha será 1.027676059340493 y se puede ver que hay
sobredispersión
```

8. Usando la informacion anterior, ejecute un modelo Binomial Negativa para responder a la pregunta 6. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Aplicando la regresión definida anterior con el modelo de Binomial Negativa, se puede ver que se obtuvo un Alpha practicamente igual de 1.03.

Además, las variables tienen la misma significancia que en el modelo Poisson.

Del modelo también se puede ver que la ubicación del sensor aporta mucho a la probabilidad de una falla, al igual que la Temperatura y el Parameter5.

0.7 Binomial Negativa

[16]:
```python
nbin = smf.negativebinomial('Failure_today ~ Vel_Mañana  + Temp_Mañana +␣
 ↪Temp_Tarde + C(Location) + Parameter4_9am + Parameter5_9am'
                             , data=df_mensual).fit()
print(nbin.summary())
```

```
         Current function value: 2.285015
         Iterations: 35
         Function evaluations: 43
         Gradient evaluations: 43
                     NegativeBinomial Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:                 4076
Model:               NegativeBinomial   Df Residuals:                     4027
Method:                           MLE   Df Model:                           48
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.1748
Time:                        21:38:52   Log-Likelihood:                -9313.7
converged:                      False   LL-Null:                        -11287.
```

```
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----
Intercept            -0.0129      0.132     -0.097      0.923      -0.272
0.247
C(Location)[T.3]     -0.3824      0.068     -5.630      0.000      -0.516
-0.249
C(Location)[T.4]     -0.1593      0.084     -1.892      0.058      -0.324
0.006
C(Location)[T.5]     -0.5111      0.068     -7.552      0.000      -0.644
-0.378
C(Location)[T.6]     -0.8659      0.073    -11.800      0.000      -1.010
-0.722
C(Location)[T.7]     -0.4477      0.068     -6.558      0.000      -0.581
-0.314
C(Location)[T.8]     -0.2390      0.066     -3.636      0.000      -0.368
-0.110
C(Location)[T.9]     -0.5921      0.068     -8.770      0.000      -0.724
-0.460
C(Location)[T.10]    -0.4328      0.070     -6.151      0.000      -0.571
-0.295
C(Location)[T.11]    -0.1583      0.073     -2.169      0.030      -0.301
-0.015
C(Location)[T.12]    -0.4152      0.065     -6.396      0.000      -0.542
-0.288
C(Location)[T.13]    -0.6421      0.066     -9.778      0.000      -0.771
-0.513
C(Location)[T.14]    -0.6813      0.070     -9.679      0.000      -0.819
-0.543
C(Location)[T.15]    -0.7329      0.073    -10.028      0.000      -0.876
-0.590
C(Location)[T.16]    -0.4393      0.063     -6.919      0.000      -0.564
-0.315
C(Location)[T.17]    -0.6591      0.117     -5.626      0.000      -0.889
-0.429
C(Location)[T.18]    -0.6681      0.075     -8.966      0.000      -0.814
-0.522
C(Location)[T.19]    -0.3699      0.070     -5.268      0.000      -0.507
-0.232
C(Location)[T.20]    -0.4790      0.070     -6.829      0.000      -0.616
-0.342
C(Location)[T.21]    -0.3860      0.079     -4.894      0.000      -0.541
-0.231
C(Location)[T.22]    -0.3498      0.078     -4.499      0.000      -0.502
```

						-0.197
C(Location)[T.23]	-0.4263	0.064	-6.637	0.000	-0.552	-0.300
C(Location)[T.26]	-0.5631	0.090	-6.263	0.000	-0.739	-0.387
C(Location)[T.27]	-0.6905	0.064	-10.768	0.000	-0.816	-0.565
C(Location)[T.28]	-0.7897	0.070	-11.309	0.000	-0.927	-0.653
C(Location)[T.29]	-0.3544	0.066	-5.361	0.000	-0.484	-0.225
C(Location)[T.30]	-0.3272	0.070	-4.674	0.000	-0.464	-0.190
C(Location)[T.32]	-0.1340	0.064	-2.105	0.035	-0.259	-0.009
C(Location)[T.33]	-0.1037	0.067	-1.543	0.123	-0.235	0.028
C(Location)[T.34]	-0.4769	0.062	-7.743	0.000	-0.598	-0.356
C(Location)[T.35]	-0.5909	0.069	-8.560	0.000	-0.726	-0.456
C(Location)[T.36]	-0.6293	0.067	-9.346	0.000	-0.761	-0.497
C(Location)[T.38]	-0.2566	0.065	-3.931	0.000	-0.385	-0.129
C(Location)[T.39]	-0.2926	0.065	-4.471	0.000	-0.421	-0.164
C(Location)[T.40]	-0.9840	0.074	-13.329	0.000	-1.129	-0.839
C(Location)[T.41]	-0.3671	0.068	-5.392	0.000	-0.501	-0.234
C(Location)[T.42]	-0.1581	0.112	-1.405	0.160	-0.378	0.062
C(Location)[T.43]	-0.2107	0.069	-3.070	0.002	-0.345	-0.076
C(Location)[T.44]	-0.6414	0.062	-10.409	0.000	-0.762	-0.521
C(Location)[T.45]	-0.3968	0.064	-6.236	0.000	-0.522	-0.272
C(Location)[T.46]	-0.5345	0.068	-7.865	0.000	-0.668	-0.401
C(Location)[T.47]	-0.4972	0.063	-7.880	0.000	-0.621	-0.374
C(Location)[T.48]	-0.7660	0.067	-11.440	0.000	-0.897	-0.635
C(Location)[T.49]	-0.6362	0.091	-6.978	0.000	-0.815	-0.458
Vel_Mañana	0.0228	0.003	7.284	0.000	0.017	

```
0.029
Temp_Mañana            0.1540      0.008     19.952      0.000       0.139
0.169
Temp_Tarde            -0.1576      0.007    -22.715      0.000      -0.171
-0.144
Parameter4_9am         0.0379      0.001     35.271      0.000       0.036
0.040
Parameter5_9am        -0.3239      0.019    -17.061      0.000      -0.361
-0.287
alpha                  0.0200      0.004      4.914      0.000       0.012
0.028
================================================================================
=====
```

[17]:
```python
alpha = np.exp(0.0328)

print(f'El valor que toma alpha es de {alpha}')
```

```
El valor que toma alpha es de 1.03334384980309
```

9. Comente los resultados obtenidos en 6, 7 y 8. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: Ambos modelos son para data de conteo. El de poisson es bueno bajo el supuesto que la varianza ~ media (no hay sobredispersion), mientras que el modelo binomial negativa es bueno cuando si existe sobredispersion (varianza » media).

Bajo lo anterior, se puede decir que el modelo de binomial negativa es mejor, ya que existe sobredispersion. Además, tiene una mejor estimación de la Log-Likelihood (muy poca deferencia).

Podemos concluir que todas las variables son robustas (mismo signo y significativas) exceptuando las categóricas:

- T.4
- T.33
- T.42