

Tarea1__Lara__Urrutia

April 30, 2025

Tarea 1 Data Analysis And Machine Learning

Gustavo Ignacio Lara Urrutia - 2021421301

- Date: data medida en frecuencia diaria
- Location: ubicacion del medidor
- Min_Temp: temperatura minima observada
- Max_Temp: temperatura maxima observada
- Leakage: Filtracion medida en el area
- Evaporation: Tasa de evaporacion
- Electricity: Consumo electrico KW
- Parameter#: Diferentes sensores de reportando direccion y velocidad de viento en distintos momentos del dia, asi como otras metricas relevantes.
- Failure today: El sensor reporta fallo (o no)

```
[2]: 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")

%matplotlib inline
```

1. Cargar la base de datos en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadísticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

```
[ ]: df = pd.read_csv('../data/machine_failure_data.csv')
df.head
```

```
[ ]: <bound method NDFrame.head of
Leakage Evaporation \
0      12/1/2008      3      13.4      22.9      0.6      NaN
1      12/2/2008      3       7.4      25.1      0.0      NaN
2      12/3/2008      3      12.9      25.7      0.0      NaN
3      12/4/2008      3       9.2      28.0      0.0      NaN
4      12/5/2008      3      17.5      32.3      1.0      NaN
...
142188 6/20/2017      42       3.5      21.8      0.0      NaN
142189 6/21/2017      42       2.8      23.4      0.0      NaN
142190 6/22/2017      42       3.6      25.3      0.0      NaN
142191 6/23/2017      42       5.4      26.9      0.0      NaN
142192 6/24/2017      42       7.8      27.0      0.0      NaN
```

```
Electricity Parameter1_Dir Parameter1_Speed Parameter2_9am ... \
0      NaN      W      44.0      W ...
1      NaN      WNW      44.0      NNW ...
2      NaN      WSW      46.0      W ...
3      NaN      NE      24.0      SE ...
4      NaN      W      41.0      ENE ...
...
142188      NaN      E      31.0      ESE ...
142189      NaN      E      31.0      SE ...
142190      NaN      NNW      22.0      SE ...
142191      NaN      N      37.0      SE ...
142192      NaN      SE      28.0      SSE ...
```

```
Parameter3_3pm Parameter4_9am Parameter4_3pm Parameter5_9am \
0      24.0      71.0      22.0      1007.7
1      22.0      44.0      25.0      1010.6
2      26.0      38.0      30.0      1007.6
3       9.0      45.0      16.0      1017.6
4      20.0      82.0      33.0      1010.8
...
142188      13.0      59.0      27.0      1024.7
142189      11.0      51.0      24.0      1024.6
142190       9.0      56.0      21.0      1023.5
142191       9.0      53.0      24.0      1021.0
142192       7.0      51.0      24.0      1019.4
```

```
Parameter5_3pm Parameter6_9am Parameter6_3pm Parameter7_9am \
0      1007.1      8.0      NaN      16.9
1      1007.8      NaN      NaN      17.2
2      1008.7      NaN      2.0      21.0
3      1012.8      NaN      NaN      18.1
4      1006.0      7.0      8.0      17.8
...
...      ...      ...      ...
```

142188	1021.2	NaN	NaN	9.4
142189	1020.3	NaN	NaN	10.1
142190	1019.1	NaN	NaN	10.9
142191	1016.8	NaN	NaN	12.5
142192	1016.5	3.0	2.0	15.1

	Parameter7_3pm	Failure_today
0	21.8	No
1	24.3	No
2	23.2	No
3	26.5	No
4	29.7	No
...
142188	20.9	No
142189	22.4	No
142190	24.5	No
142191	26.1	No
142192	26.0	No

[142193 rows x 22 columns]>

```
[6]: df.describe(include='all')
```

```
[6]:
```

	Date	Location	Min_Temp	Max_Temp	Leakage \
count	142193	142193.000000	141556.000000	141871.000000	140787.000000
unique	3436	NaN	NaN	NaN	NaN
top	12/1/2013	NaN	NaN	NaN	NaN
freq	49	NaN	NaN	NaN	NaN
mean	NaN	24.740655	12.186400	23.226784	2.349974
std	NaN	14.237503	6.403283	7.117618	8.465173
min	NaN	1.000000	-8.500000	-4.800000	0.000000
25%	NaN	12.000000	7.600000	17.900000	0.000000
50%	NaN	25.000000	12.000000	22.600000	0.000000
75%	NaN	37.000000	16.800000	28.200000	0.800000
max	NaN	49.000000	33.900000	48.100000	371.000000

	Evaporation	Electricity	Parameter1_Dir	Parameter1_Speed \
count	81350.000000	74377.000000	132863	132923.000000
unique	NaN	NaN	16	NaN
top	NaN	NaN	W	NaN
freq	NaN	NaN	9780	NaN
mean	5.469824	7.624853	NaN	39.984292
std	4.188537	3.781525	NaN	13.588801
min	0.000000	0.000000	NaN	6.000000
25%	2.600000	4.900000	NaN	31.000000
50%	4.800000	8.500000	NaN	39.000000
75%	7.400000	10.600000	NaN	48.000000

max	145.000000	14.500000	NaN	135.000000
-----	------------	-----------	-----	------------

	Parameter2_9am	...	Parameter3_3pm	Parameter4_9am	Parameter4_3pm	\
count	132180	...	139563.000000	140419.000000	138583.000000	
unique	16	...	NaN	NaN	NaN	
top	N	...	NaN	NaN	NaN	
freq	11393	...	NaN	NaN	NaN	
mean	NaN	...	18.637576	68.843810	51.482606	
std	NaN	...	8.803345	19.051293	20.797772	
min	NaN	...	0.000000	0.000000	0.000000	
25%	NaN	...	13.000000	57.000000	37.000000	
50%	NaN	...	19.000000	70.000000	52.000000	
75%	NaN	...	24.000000	83.000000	66.000000	
max	NaN	...	87.000000	100.000000	100.000000	

	Parameter5_9am	Parameter5_3pm	Parameter6_9am	Parameter6_3pm	\
count	128179.000000	128212.000000	88536.000000	85099.000000	
unique	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	
mean	1017.653758	1015.258204	4.437189	4.503167	
std	7.105476	7.036677	2.887016	2.720633	
min	980.500000	977.100000	0.000000	0.000000	
25%	1012.900000	1010.400000	1.000000	2.000000	
50%	1017.600000	1015.200000	5.000000	5.000000	
75%	1022.400000	1020.000000	7.000000	7.000000	
max	1041.000000	1039.600000	9.000000	9.000000	

	Parameter7_9am	Parameter7_3pm	Failure_today
count	141289.000000	139467.000000	140787
unique	NaN	NaN	2
top	NaN	NaN	No
freq	NaN	NaN	109332
mean	16.987509	21.687235	NaN
std	6.492838	6.937594	NaN
min	-7.200000	-5.400000	NaN
25%	12.300000	16.600000	NaN
50%	16.700000	21.100000	NaN
75%	21.600000	26.400000	NaN
max	40.200000	46.700000	NaN

[11 rows x 22 columns]

```
[7]: df.dtypes
```

```
[7]: Date          object
     Location      int64
```

```

Min_Temp          float64
Max_Temp          float64
Leakage           float64
Evaporation       float64
Electricity       float64
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    float64
Parameter5_3pm    float64
Parameter6_9am    float64
Parameter6_3pm    float64
Parameter7_9am    float64
Parameter7_3pm    float64
Failure_today     object
dtype: object

```

A continuación, se agregan variables categóricas para las direcciones del viento y dummy para Leakage y Failure.

```

[ ]: VIENTO = list(range(1, 17))
DIRECV = ['SSW', 'S', 'NNE', 'WNW', 'N', 'SE', 'ENE', 'NE', 'E', 'SW', 'W', 'WSW', 'NNW', 'ESE', 'SSE', 'NW']
df['VIENTO'] = df['Parameter1_Dir'].replace(DIRECV, VIENTO)

df['VIENTO2_9am'] = df['Parameter2_9am'].replace(DIRECV, VIENTO)

df['VIENTO3_2pm'] = df['Parameter2_3pm'].replace(DIRECV, VIENTO)

```

```

[9]: df['VIENTO'].head

```

```

[9]: <bound method NDFrame.head of 0          11.0
1           4.0
2          12.0
3           8.0
4          11.0
...
142188      9.0
142189      9.0
142190     13.0
142191      5.0
142192      6.0
Name: VIENTO, Length: 142193, dtype: float64>

```

```
[10]: df['VIENTO2_9am'].head
```

```
[10]: <bound method NDFrame.head of 0          11.0
      1          13.0
      2          11.0
      3           6.0
      4           7.0
      ...
      142188      14.0
      142189       6.0
      142190       6.0
      142191       6.0
      142192      15.0
      Name: VIENTO2_9am, Length: 142193, dtype: float64>
```

```
[ ]: df['VIENTO3_2pm'].head
```

```
[64]: df['Failure_today'] = df['Failure_today'].str.strip().str.lower()

      df['Failure_today'] = df['Failure_today'].map({'yes': 1, 'no': 0})
```

```
[12]: df['Failure_today'].head
```

```
[12]: <bound method NDFrame.head of 0          0.0
      1          0.0
      2          0.0
      3          0.0
      4          0.0
      ...
      142188      0.0
      142189      0.0
      142190      0.0
      142191      0.0
      142192      0.0
      Name: Failure_today, Length: 142193, dtype: float64>
```

```
[13]: df['Failure_today'].value_counts()
```

```
[13]: Failure_today
      0.0    109332
      1.0     31455
      Name: count, dtype: int64
```

```
[66]: df['LK'] = df['Leakage'].apply(lambda x: 1 if x > 0 else 0)
      print(df['LK'])
```

```
0          1
1          0
```

```

2         0
3         0
4         1
..
142188    0
142189    0
142190    0
142191    0
142192    0
Name: LK, Length: 142193, dtype: int64

```

```
[15]: print(df['LK'].value_counts())
```

```

LK
0    91681
1    50512
Name: count, dtype: int64

```

Se eligen las variables con más de 50 mil NaN y se rellenan con ceros.

```
[67]: df['Evaporation'] = df['Evaporation'].fillna(0)
df['Electricity'] = df['Electricity'].fillna(0)
df['Parameter6_9am'] = df['Parameter6_9am'].fillna(0)
df['Parameter6_3pm'] = df['Parameter6_3pm'].fillna(0)

```

Se eliminan filas con NaN en el resto de variables.

```
[68]: columnas_con_nan = ['Min_Temp', 'Max_Temp', 'Leakage', 'Parameter1_Dir', '
↳ 'Parameter1_Speed',
                        'Parameter2_9am', 'Parameter2_3pm', 'Parameter3_9am', '
↳ 'Parameter3_3pm',
                        'Parameter4_9am', 'Parameter4_3pm', 'Parameter5_9am', '
↳ 'Parameter5_3pm',
                        'Parameter7_9am', 'Parameter7_3pm', 'Failure_today', '
↳ 'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm']

df = df.dropna(subset=columnas_con_nan)

```

```
[69]: parametros = ['Failure_today', 'Location', 'Min_Temp', 'Max_Temp', 'Evaporation',
                    'Electricity', 'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm', '
↳ 'Parameter4_9am',
                    'Parameter4_3pm', 'Parameter5_9am', 'Parameter5_3pm', 'Parameter6_9am',
                    'Parameter6_3pm', 'Parameter7_9am', 'Parameter7_3pm',
                    'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm', 'LK']

df_sub = df[parametros]
corr = df_sub.corr(numeric_only=True)
mask = np.triu(np.ones_like(corr, dtype=bool))

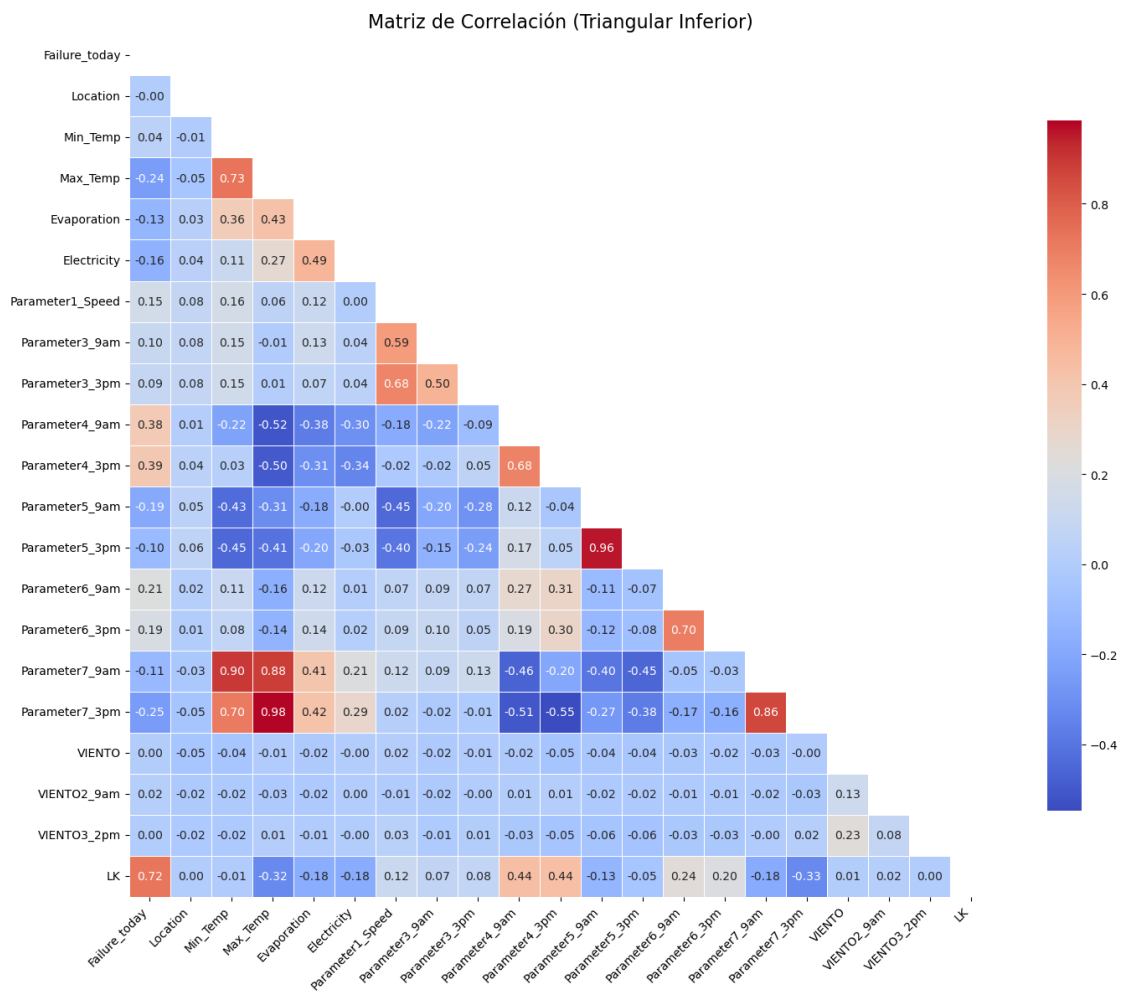
```

```
plt.figure(figsize=(15, 12))

sns.heatmap(corr, mask=mask, annot=True, fmt=".2f", cmap='coolwarm',
            square=True, linewidths=0.5, cbar_kws={"shrink": .8})

plt.title('Matriz de Correlación (Triangular Inferior)', fontsize=16)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()

plt.show()
```



2. Ejecute un modelo de probabilidad lineal (*MCO*) que permita explicar la probabilidad de que un día 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: Se estimó un modelo de probabilidad lineal excluyendo aquellas variables con correlaciones altas, donde la variable dependiente es 'Failure_today', la cual toma valor 1 si es que existe un fallo y 0 en caso que no. De los resultados se obtuvo un R cuadrado de 0.548, es decir, explica el 54.8% de la variabilidad. Según el modelo, la locación, la temperatura máxima, el nivel de evaporación, la electricidad y varios otros se relacionan de manera negativa, es decir, reducen la probabilidad de fallo. Aquellas variables categóricas para las direcciones del viento tienen un valor-p mayor a 0.05, por lo que no son estadísticamente significativas. Finalmente, la variable 'LK' es altamente significativa y tiene un alto impacto en la probabilidad de falla.

```
[70]: #Regresion excluyendo variables de alta correlacion
y=df['Failure_today']
X=df.drop(['Date', 'Leakage', 'Parameter1_Dir',
          'Parameter2_9am', 'Parameter2_3pm', 'Failure_today',
          'Parameter7_9am', 'Parameter7_3pm', 'Parameter5_3pm'], axis=1)
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
```

OLS Regression Results

```
=====
Dep. Variable:          Failure_today    R-squared:                0.548
Model:                  OLS              Adj. R-squared:          0.548
Method:                 Least Squares    F-statistic:             5448.
Date:                  Thu, 24 Apr 2025  Prob (F-statistic):       0.00
Time:                  20:12:45          Log-Likelihood:          -16694.
No. Observations:      112925           AIC:                    3.342e+04
Df Residuals:          112907           BIC:                    3.360e+04
Df Model:              17
Covariance Type:       HC0
=====
```

```
=====
coef      std err          z      P>|z|      [0.025
0.975]
-----
----
const          4.5273      0.172     26.368      0.000      4.191
4.864
Location     -0.0002     5.74e-05     -3.278      0.001     -0.000
-7.57e-05
Min_Temp       0.0055      0.000     19.835      0.000      0.005
0.006
Max_Temp      -0.0043      0.000    -15.401      0.000     -0.005
-0.004
Evaporation   -0.0009      0.000     -4.160      0.000     -0.001
-0.000
Electricity    0.0002      0.000      1.188      0.235     -0.000
0.001
```

Parameter1_Speed	0.0022	0.000	21.293	0.000	0.002
0.002					
Parameter3_9am	0.0014	0.000	10.853	0.000	0.001
0.002					
Parameter3_3pm	-0.0023	0.000	-16.889	0.000	-0.003
-0.002					
Parameter4_9am	0.0021	6.74e-05	30.705	0.000	0.002
0.002					
Parameter4_3pm	6.755e-05	7.9e-05	0.855	0.392	-8.72e-05
0.000					
Parameter5_9am	-0.0046	0.000	-27.773	0.000	-0.005
-0.004					
Parameter6_9am	-0.0002	0.000	-0.485	0.628	-0.001
0.001					
Parameter6_3pm	0.0011	0.000	2.649	0.008	0.000
0.002					
VIENTO	2.943e-05	0.000	0.152	0.879	-0.000
0.000					
VIENTO2_9am	0.0001	0.000	0.608	0.543	-0.000
0.000					
VIENTO3_2pm	-3.908e-06	0.000	-0.020	0.984	-0.000
0.000					
LK	0.5583	0.003	204.127	0.000	0.553
0.564					

Omnibus:	11391.614	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15244.885
Skew:	-0.849	Prob(JB):	0.00
Kurtosis:	3.596	Cond. No.	1.90e+05

Notes:

- [1] Standard Errors are heteroscedasticity robust (HCO)
- [2] The condition number is large, 1.9e+05. 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: Se observa que Leakage afecta significativamente a la probabilidad de falla, es decir, un aumento de esta variable aumenta la probabilidad de falla en un 88.01 puntos porcentuales. Como el modelo no convergió correctamente debido a la influencia de esta variable, se optó por excluirla para evaluar los efectos de las demás variables. De este segundo intento, se obtuvo un pseudo R cuadrado moderado de 0.3022 y un LLR p-value de 0.000, lo que indica que el modelo sigue siendo significativo. La mayor parte de las variables aumentaron la magnitud de sus efectos marginales manteniendo el sentido. Por ejemplo, la temperatura máxima y la evaporación disminuyen la probabilidad de falla, donde un aumento de una unidad de esas variables disminuye en 12.08 y 2.83 puntos porcentuales la probabilidad de falla. Por otro lado, la variable VIENTO3_2pm no es estadísticamente significativa, mientras que el resto de variables sí. .

```
[71]: X = df[['Location', 'Min_Temp', 'Max_Temp', 'Evaporation',
            'Electricity', 'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
            ↪ 'Parameter4_9am',
            'Parameter4_3pm', 'Parameter5_9am', 'Parameter6_9am', 'Parameter6_3pm',
            'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm', 'LK']]
y = df['Failure_today']

X = sm.add_constant(X)
probit_model = sm.Probit(y, X).fit(cov_type='HCO')
print(probit_model.summary())

# Efectos marginales
mfx = probit_model.get_margeff()
print(mfx.summary())
```

Warning: Maximum number of iterations has been exceeded.
Current function value: 0.208440
Iterations: 35

Probit Regression Results

```
=====
Dep. Variable:          Failure_today    No. Observations:          112925
Model:                  Probit           Df Residuals:            112907
Method:                 MLE             Df Model:                 17
Date:                   Thu, 24 Apr 2025   Pseudo R-squ.:            0.6087
Time:                   20:12:56          Log-Likelihood:           -23538.
converged:              False            LL-Null:                  -60159.
Covariance Type:        HCO              LLR p-value:              0.000
=====
=====
```

	coef	std err	z	P> z	[0.025
0.975]					

const	17.4311	1.110	15.698	0.000	15.255
19.607					
Location	-0.0005	0.000	-1.117	0.264	-0.001
0.000					
Min_Temp	0.0858	0.003	31.901	0.000	0.081
0.091					
Max_Temp	-0.0697	0.003	-26.294	0.000	-0.075
-0.065					
Evaporation	-0.0144	0.003	-5.402	0.000	-0.020
-0.009					
Electricity	0.0105	0.002	5.256	0.000	0.007
0.014					
Parameter1_Speed	0.0143	0.001	17.142	0.000	0.013
0.016					
Parameter3_9am	0.0072	0.001	6.323	0.000	0.005

0.009					
Parameter3_3pm	-0.0125	0.001	-10.731	0.000	-0.015
-0.010					
Parameter4_9am	0.0197	0.001	31.268	0.000	0.018
0.021					
Parameter4_3pm	-0.0022	0.001	-3.683	0.000	-0.003
-0.001					
Parameter5_9am	-0.0256	0.001	-23.804	0.000	-0.028
-0.023					
Parameter6_9am	-0.0127	0.003	-4.113	0.000	-0.019
-0.007					
Parameter6_3pm	0.0124	0.003	3.981	0.000	0.006
0.019					
VIENTO	-0.0008	0.001	-0.512	0.609	-0.004
0.002					
VIENTO2_9am	-0.0002	0.001	-0.107	0.915	-0.003
0.003					
VIENTO3_2pm	-0.0003	0.001	-0.221	0.825	-0.003
0.003					
LK	7.4255	0.020	370.262	0.000	7.386
7.465					

=====

====

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

Probit Marginal Effects

=====

Dep. Variable: Failure_today

Method: dydx

At: overall

=====

====

	dy/dx	std err	z	P> z	[0.025
0.975]					

Location	-6.312e-05	5.65e-05	-1.117	0.264	-0.000
4.77e-05					
Min_Temp	0.0102	0.000	33.253	0.000	0.010
0.011					
Max_Temp	-0.0083	0.000	-27.050	0.000	-0.009
-0.008					
Evaporation	-0.0017	0.000	-5.409	0.000	-0.002
-0.001					
Electricity	0.0012	0.000	5.261	0.000	0.001
0.002					

Parameter1_Speed	0.0017	9.76e-05	17.340	0.000	0.002
0.002					
Parameter3_9am	0.0009	0.000	6.332	0.000	0.001
0.001					
Parameter3_3pm	-0.0015	0.000	-10.780	0.000	-0.002
-0.001					
Parameter4_9am	0.0023	7.2e-05	32.351	0.000	0.002
0.002					
Parameter4_3pm	-0.0003	7.19e-05	-3.686	0.000	-0.000
-0.000					
Parameter5_9am	-0.0030	0.000	-24.287	0.000	-0.003
-0.003					
Parameter6_9am	-0.0015	0.000	-4.115	0.000	-0.002
-0.001					
Parameter6_3pm	0.0015	0.000	3.983	0.000	0.001
0.002					
VIENTO	-9.054e-05	0.000	-0.512	0.609	-0.000
0.000					
VIENTO2_9am	-1.826e-05	0.000	-0.107	0.915	-0.000
0.000					
VIENTO3_2pm	-3.877e-05	0.000	-0.221	0.825	-0.000
0.000					
LK	0.8801	0.002	449.131	0.000	0.876
0.884					

=====

====

Se puede observar que cerca del 63% de los casos si existe filtración, entonces se produce una falla. Este es un caso de cuasi-separación perfecta, por lo que el modelo Probit no converge. Por eso, se realiza un análisis separado sin LK para conocer los efectos de las otras variables.

```
[24]: pd.crosstab(df['LK'], df['Failure_today'], normalize='index')
```

```
[24]: Failure_today      0.0      1.0
LK
0      1.000000  0.000000
1      0.367971  0.632029
```

```
[96]: X = df[['Location', 'Min_Temp', 'Max_Temp', 'Evaporation',
            'Electricity', 'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
            ↪ 'Parameter4_9am',
            'Parameter4_3pm', 'Parameter5_9am', 'Parameter6_9am', 'Parameter6_3pm',
            'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm']]
y = df['Failure_today']

X = sm.add_constant(X)
probit_model = sm.Probit(y, X).fit(cov_type='HCO')
print(probit_model.summary())
```

```
Optimization terminated successfully.  
Current function value: 0.371758  
Iterations 7
```

-0.001					
Parameter6_3pm	0.0202	0.002	8.925	0.000	0.016
0.025					
VIENTO	0.0037	0.001	3.266	0.001	0.001
0.006					
VIENTO2_9am	0.0040	0.001	3.722	0.000	0.002
0.006					
VIENTO3_2pm	0.0018	0.001	1.586	0.113	-0.000
0.004					

=====

===== Probit Marginal Effects

Dep. Variable: Failure_today
Method: dydx
At: overall

	dy/dx	std err	z	P> z	[0.025
0.975]					

Location	-0.0004	7.45e-05	-5.226	0.000	-0.001
-0.000					
Min_Temp	0.0248	0.000	64.768	0.000	0.024
0.026					
Max_Temp	-0.0252	0.000	-65.495	0.000	-0.026
-0.024					
Evaporation	-0.0059	0.001	-10.372	0.000	-0.007
-0.005					
Electricity	0.0032	0.000	9.604	0.000	0.003
0.004					
Parameter1_Speed	0.0036	0.000	30.051	0.000	0.003
0.004					
Parameter3_9am	0.0016	0.000	9.717	0.000	0.001
0.002					
Parameter3_3pm	-0.0029	0.000	-17.034	0.000	-0.003
-0.003					
Parameter4_9am	0.0075	8.79e-05	85.118	0.000	0.007
0.008					
Parameter4_3pm	-0.0008	9.17e-05	-9.130	0.000	-0.001
-0.001					
Parameter5_9am	-0.0068	0.000	-39.111	0.000	-0.007
-0.006					
Parameter6_9am	-0.0012	0.000	-2.569	0.010	-0.002
-0.000					
Parameter6_3pm	0.0042	0.000	8.934	0.000	0.003
0.005					

VIENTO	0.0008	0.000	3.267	0.001	0.000
0.001					
VIENTO2_9am	0.0008	0.000	3.723	0.000	0.000
0.001					
VIENTO3_2pm	0.0004	0.000	1.586	0.113	-8.69e-05
0.001					

=====

=====

De estos resultados se puede observar un mejor ajuste de 0.371569, lo cual es moderado.

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: Al igual que el modelo Probit, Leakage sigue teniendo un efecto significativo con un efecto marginal de aproximadamente 295 puntos porcentuales. El pseudo R^2 es de 0.6088 y el LLR p-value es de 0.000, lo que indica que el modelo es globalmente significativo. Al excluir 'LK' del modelo, se obtuvo un nuevo pseudo R^2 de 0.3107, y no se detectaron cambios significativos en las magnitudes de los efectos marginales. Se pueden interpretar algunos Odds ratios: cada grado más de temperatura mínima aumenta las probabilidades de falla en un 14.99%, mientras que cada grado más de temperatura máxima reduce las probabilidades en un 22.93% (de la fórmula $(\text{Odds} - 1) \times 100\%$).

```
[104]: X1 = df[['Location', 'Min_Temp', 'Max_Temp', 'Evaporation',
            'Electricity', 'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
            ↪ 'Parameter4_9am',
            'Parameter4_3pm', 'Parameter5_9am', 'Parameter6_9am', 'Parameter6_3pm',
            'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm', 'LK']]

y = df[['Failure_today']]
model = sm.Logit(y, X1)
logit_model = model.fit(cov_type='HCO')
print(logit_model.summary())

mfxl = logit_model.get_margeff()
print(mfxl.summary())

params = logit_model.params
conf = logit_model.conf_int()
conf['Odds Ratio'] = params
conf.columns = ['Odds Ratio', '5%', '95%']
print("Odds Ratios")
print(np.exp(conf).iloc[1:17, ])
```

Optimization terminated successfully.

Current function value: 0.208418

Iterations 14

Logit Regression Results

=====

Dep. Variable:	Failure_today	No. Observations:	112925
Model:	Logit	Df Residuals:	112908
Method:	MLE	Df Model:	16
Date:	Thu, 24 Apr 2025	Pseudo R-squ.:	0.6088
Time:	23:37:43	Log-Likelihood:	-23536.
converged:	True	LL-Null:	-60159.
Covariance Type:	HCO	LLR p-value:	0.000

=====

====

	coef	std err	z	P> z	[0.025
0.975]					

Location	-0.0009	0.001	-1.126	0.260	-0.002
0.001					
Min_Temp	0.1427	0.005	31.575	0.000	0.134
0.152					
Max_Temp	-0.1165	0.004	-26.234	0.000	-0.125
-0.108					
Evaporation	-0.0244	0.005	-5.281	0.000	-0.033
-0.015					
Electricity	0.0176	0.003	5.257	0.000	0.011
0.024					
Parameter1_Speed	0.0245	0.001	17.174	0.000	0.022
0.027					
Parameter3_9am	0.0122	0.002	6.376	0.000	0.008
0.016					
Parameter3_3pm	-0.0218	0.002	-11.008	0.000	-0.026
-0.018					
Parameter4_9am	0.0324	0.001	30.749	0.000	0.030
0.034					
Parameter4_3pm	-0.0038	0.001	-3.713	0.000	-0.006
-0.002					
Parameter5_9am	-0.0426	0.002	-23.693	0.000	-0.046
-0.039					
Parameter6_9am	-0.0211	0.005	-4.139	0.000	-0.031
-0.011					
Parameter6_3pm	0.0204	0.005	3.956	0.000	0.010
0.031					
VIENTO	-0.0015	0.002	-0.613	0.540	-0.006
0.003					
VIENTO2_9am	-0.0006	0.002	-0.245	0.807	-0.005
0.004					
VIENTO3_2pm	-0.0007	0.002	-0.295	0.768	-0.006
0.004					
LK	41.4266	1.862	22.243	0.000	37.776
45.077					

=====

====

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

Logit Marginal Effects

=====

Dep. Variable: Failure_today

Method: dydx

At: overall

=====

====

	dy/dx	std err	z	P> z	[0.025
0.975]					

Location	-6.351e-05	5.64e-05	-1.126	0.260	-0.000
4.7e-05					
Min_Temp	0.0102	0.000	33.261	0.000	0.010
0.011					
Max_Temp	-0.0083	0.000	-27.191	0.000	-0.009
-0.008					
Evaporation	-0.0017	0.000	-5.289	0.000	-0.002
-0.001					
Electricity	0.0013	0.000	5.264	0.000	0.001
0.002					
Parameter1_Speed	0.0017	0.000	17.414	0.000	0.002
0.002					
Parameter3_9am	0.0009	0.000	6.388	0.000	0.001
0.001					
Parameter3_3pm	-0.0016	0.000	-11.074	0.000	-0.002
-0.001					
Parameter4_9am	0.0023	7.23e-05	31.993	0.000	0.002
0.002					
Parameter4_3pm	-0.0003	7.24e-05	-3.717	0.000	-0.000
-0.000					
Parameter5_9am	-0.0030	0.000	-24.265	0.000	-0.003
-0.003					
Parameter6_9am	-0.0015	0.000	-4.142	0.000	-0.002
-0.001					
Parameter6_3pm	0.0015	0.000	3.959	0.000	0.001
0.002					
VIENTO	-0.0001	0.000	-0.613	0.540	-0.000
0.000					
VIENTO2_9am	-4.192e-05	0.000	-0.245	0.807	-0.000
0.000					
VIENTO3_2pm	-5.185e-05	0.000	-0.295	0.768	-0.000
0.000					

LK	2.9564	0.130	22.720	0.000	2.701
3.211					

====

Odds Ratios

	Odds Ratio	5%	95%
Min_Temp	1.143223e+00	1.163659e+00	1.153396e+00
Max_Temp	8.823563e-01	8.978447e-01	8.900668e-01
Evaporation	9.671625e-01	9.848031e-01	9.759429e-01
Electricity	1.011072e+00	1.024396e+00	1.017712e+00
Parameter1_Speed	1.021902e+00	1.027623e+00	1.024758e+00
Parameter3_9am	1.008504e+00	1.016113e+00	1.012301e+00
Parameter3_3pm	9.746836e-01	9.822676e-01	9.784683e-01
Parameter4_9am	1.030793e+00	1.035059e+00	1.032924e+00
Parameter4_3pm	9.942538e-01	9.982207e-01	9.962353e-01
Parameter5_9am	9.549619e-01	9.617105e-01	9.583303e-01
Parameter6_9am	9.694254e-01	9.889678e-01	9.791479e-01
Parameter6_3pm	1.010365e+00	1.031035e+00	1.020648e+00
VIENTO	9.936406e-01	1.003345e+00	9.984809e-01
VIENTO2_9am	9.947254e-01	1.004122e+00	9.994127e-01
VIENTO3_2pm	9.944696e-01	1.004101e+00	9.992737e-01
LK	2.547009e+16	3.772554e+19	9.802412e+17

```
[105]: X1 = df[['Location', 'Min_Temp', 'Max_Temp', 'Evaporation',
               'Electricity', 'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
               'Parameter4_9am',
               'Parameter4_3pm', 'Parameter5_9am', 'Parameter5_3pm', 'Parameter6_9am',
               'Parameter6_3pm', 'Parameter7_9am', 'Parameter7_3pm',
               'VIENTO', 'VIENTO2_9am', 'VIENTO3_2pm']]
y = df[['Failure_today']]
model = sm.Logit(y, X1)
logit_model = model.fit(cov_type='HCO')
print(logit_model.summary())

mfxl = logit_model.get_margeff()
print(mfxl.summary())

params = logit_model.params
conf = logit_model.conf_int()
conf['Odds Ratio'] = params
conf.columns = ['Odds Ratio', '5%', '95%']
print("Odds Ratios")
print(np.exp(conf).iloc[1:17,])
```

Optimization terminated successfully.

Current function value: 0.367194

Iterations 7

Logit Regression Results

```

=====
Dep. Variable:          Failure_today    No. Observations:          112925
Model:                  Logit            Df Residuals:              112906
Method:                 MLE             Df Model:                  18
Date:                  Thu, 24 Apr 2025   Pseudo R-squ.:            0.3107
Time:                  23:38:52          Log-Likelihood:           -41465.
converged:              True            LL-Null:                  -60159.
Covariance Type:        HCO            LLR p-value:              0.000
=====

```

```

=====
              coef      std err          z      P>|z|      [0.025
0.975]
-----
----
Location      -0.0052      0.001     -8.176      0.000     -0.006
-0.004
Min_Temp       0.1493      0.005     30.401      0.000      0.140
0.159
Max_Temp      -0.2432      0.009    -27.660      0.000     -0.260
-0.226
Evaporation   -0.0614      0.005    -12.698      0.000     -0.071
-0.052
Electricity    0.0277      0.003      9.569      0.000      0.022
0.033
Parameter1_Speed 0.0431      0.001     43.713      0.000      0.041
0.045
Parameter3_9am  0.0061      0.001      4.305      0.000      0.003
0.009
Parameter3_3pm -0.0286      0.001    -19.380      0.000     -0.032
-0.026
Parameter4_9am  0.0693      0.001     65.052      0.000      0.067
0.071
Parameter4_3pm -0.0031      0.001     -2.967      0.003     -0.005
-0.001
Parameter5_9am -0.2166      0.005    -40.988      0.000     -0.227
-0.206
Parameter5_3pm  0.2104      0.005     39.911      0.000      0.200
0.221
Parameter6_9am  0.0014      0.004      0.346      0.729     -0.006
0.009
Parameter6_3pm  0.0386      0.004      9.610      0.000      0.031
0.046
Parameter7_9am  0.0718      0.007      9.837      0.000      0.057
0.086
Parameter7_3pm  0.0631      0.010      6.508      0.000      0.044
0.082
VIENTO         0.0106      0.002      5.386      0.000      0.007
0.014

```

VIENT02_9am 0.012	0.0086	0.002	4.542	0.000	0.005
VIENT03_2pm 0.011	0.0075	0.002	3.826	0.000	0.004

=====

====

Logit Marginal Effects

Dep. Variable: Failure_today
Method: dydx
At: overall

=====

====

	dy/dx	std err	z	P> z	[0.025
--	-------	---------	---	------	--------

0.975]

Location -0.000	-0.0006	7.38e-05	-8.186	0.000	-0.001
Min_Temp 0.019	0.0174	0.001	30.816	0.000	0.016
Max_Temp -0.026	-0.0284	0.001	-28.032	0.000	-0.030
Evaporation -0.006	-0.0072	0.001	-12.781	0.000	-0.008
Electricity 0.004	0.0032	0.000	9.603	0.000	0.003
Parameter1_Speed 0.005	0.0050	0.000	45.441	0.000	0.005
Parameter3_9am 0.001	0.0007	0.000	4.306	0.000	0.000
Parameter3_3pm -0.003	-0.0033	0.000	-19.525	0.000	-0.004
Parameter4_9am 0.008	0.0081	0.000	70.263	0.000	0.008
Parameter4_3pm -0.000	-0.0004	0.000	-2.966	0.003	-0.001
Parameter5_9am -0.024	-0.0253	0.001	-42.158	0.000	-0.026
Parameter5_3pm 0.026	0.0246	0.001	40.981	0.000	0.023
Parameter6_9am 0.001	0.0002	0.000	0.346	0.729	-0.001
Parameter6_3pm 0.005	0.0045	0.000	9.622	0.000	0.004
Parameter7_9am 0.010	0.0084	0.001	9.846	0.000	0.007
Parameter7_3pm	0.0074	0.001	6.518	0.000	0.005

0.010					
VIENTO	0.0012	0.000	5.389	0.000	0.001
0.002					
VIENTO2_9am	0.0010	0.000	4.544	0.000	0.001
0.001					
VIENTO3_2pm	0.0009	0.000	3.826	0.000	0.000
0.001					
=====					
=====					
Odds Ratios					
	Odds Ratio	5%	95%		
Min_Temp	1.149880	1.172228	1.161000		
Max_Temp	0.770705	0.797732	0.784102		
Evaporation	0.931628	0.949441	0.940493		
Electricity	1.022295	1.033973	1.028117		
Parameter1_Speed	1.041979	1.046010	1.043992		
Parameter3_9am	1.003353	1.008982	1.006163		
Parameter3_3pm	0.968976	0.974602	0.971785		
Parameter4_9am	1.069548	1.074025	1.071784		
Parameter4_3pm	0.994902	0.998956	0.996927		
Parameter5_9am	0.796986	0.813665	0.805282		
Parameter5_3pm	1.221537	1.247048	1.234227		
Parameter6_9am	0.993617	1.009191	1.001373		
Parameter6_3pm	1.031171	1.047518	1.039312		
Parameter7_9am	1.059157	1.089888	1.074413		
Parameter7_3pm	1.045048	1.085504	1.065084		
VIENTO	1.006773	1.014577	1.010668		

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 investigación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: De los tres modelos se pudo observar la fuerte influencia que tienen las filtraciones en la probabilidad de fallo en los sensores. MCO no es buen modelo de especificación debido a las variables binarias, sin embargo, permite conocer las correlaciones lineales y las direcciones de influencia. En este contexto, Probit y Logit son modelos más adecuados, ya que determinan los efectos marginales sobre la probabilidad de fallo. Las diferencias identificadas son que en MCO las variables de viento no fueron estadísticamente significativas, mientras que en Probit VIENTO2_9am si lo fué y en Logit se sumó VIENTO como otra variable significativa SOLO al excluir Leakage de los modelos. En este sentido, Logit es más adecuado debido a que tiene más variables significativas que los otros modelos y la naturaleza binaria del análisis.

Variables robustas identificadas son algunas como Min_Temp, Max_Temp, Evaporation, Parameter6_3pm, entre otras, debido a su consistencia en los tres modelos. Esto refuerza su relevancia como predictores del fallo en los sensores.

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.

R: Al agregar la data a nivel mensual, excluyendo las variables categóricas, se obtuvieron un total de 4137 observaciones. El modelo Poisson mostró un pseudo R-cuadrado elevado (0.889) y convergió en solo 5 iteraciones, lo que indica buena estabilidad y ajuste. Se observaron cambios relevantes en la magnitud y dirección de algunos coeficientes respecto a modelos anteriores; por ejemplo, ahora un aumento en la temperatura máxima se asocia a un mayor riesgo de fallas, lo que contrasta con resultados previos. También cambiaron de signo variables como 'Parameter5_3pm' y 'Parameter7_3pm'. Las variables relacionadas con filtraciones, como 'LK' y 'Leakage', se mantienen como factores altamente significativos en la predicción de fallas. En cuanto a los indicadores de datos faltantes, todos resultaron significativos salvo 'I_elect', lo cual es coherente con que la variable 'Electricity' tampoco lo sea. En particular, 'I_param6_9am' mostró una fuerte asociación con menos fallas (coeficiente de -0.3168), mientras que 'I_param6_3pm' se asoció a un mayor número de fallas (coeficiente de 0.3865), lo que sugiere que la presencia o ausencia de datos en estas variables también entrega información importante sobre el comportamiento de los sensores.

```
[76]: df['Date'] = pd.to_datetime(df['Date'])
df['Month'] = df['Date'].dt.to_period('M')
df['Failure_today'] = df['Failure_today'].astype(int)

df_mes_sensor = df.groupby(['Location', 'Month']).agg({
    'Failure_today': 'sum',
    'Evaporation': 'mean',
    'Electricity': 'mean',
    'Min_Temp': 'mean',
    'Max_Temp': 'mean',
    'Parameter1_Speed': 'mean',
    'Parameter3_9am': 'mean',
    'Parameter3_3pm': 'mean',
    'Parameter4_9am': 'mean',
    'Parameter4_3pm': 'mean',
    'Parameter5_9am': 'mean',
    'Parameter5_3pm': 'mean',
    'Parameter6_9am': 'mean',
    'Parameter6_3pm': 'mean',
    'Parameter7_9am': 'mean',
    'Parameter7_3pm': 'mean',
    'Leakage' : 'mean',
    'LK' : 'mean',
}).reset_index()
```

```
[77]: df_mes_sensor['I_evap'] = (df_mes_sensor['Evaporation'] == 0).astype(int)
df_mes_sensor['I_elect'] = (df_mes_sensor['Electricity'] == 0).astype(int)
df_mes_sensor['I_param6_9am'] = (df_mes_sensor['Parameter6_9am'] == 0).
    ↪astype(int)
df_mes_sensor['I_param6_3pm'] = (df_mes_sensor['Parameter6_3pm'] == 0).
    ↪astype(int)
```

```
[79]: df_mes_sensor.head()
```

```
[79]:
```

	Location	Month	Failure_today	Evaporation	Electricity	Min_Temp	\
0	1	2008-07	10	2.110000	4.485000	7.000000	
1	1	2008-08	10	1.715789	6.147368	5.936842	
2	1	2008-09	4	4.446154	8.588462	9.461538	
3	1	2008-10	2	5.091667	9.145833	12.383333	
4	1	2008-11	5	6.178571	9.064286	14.210714	

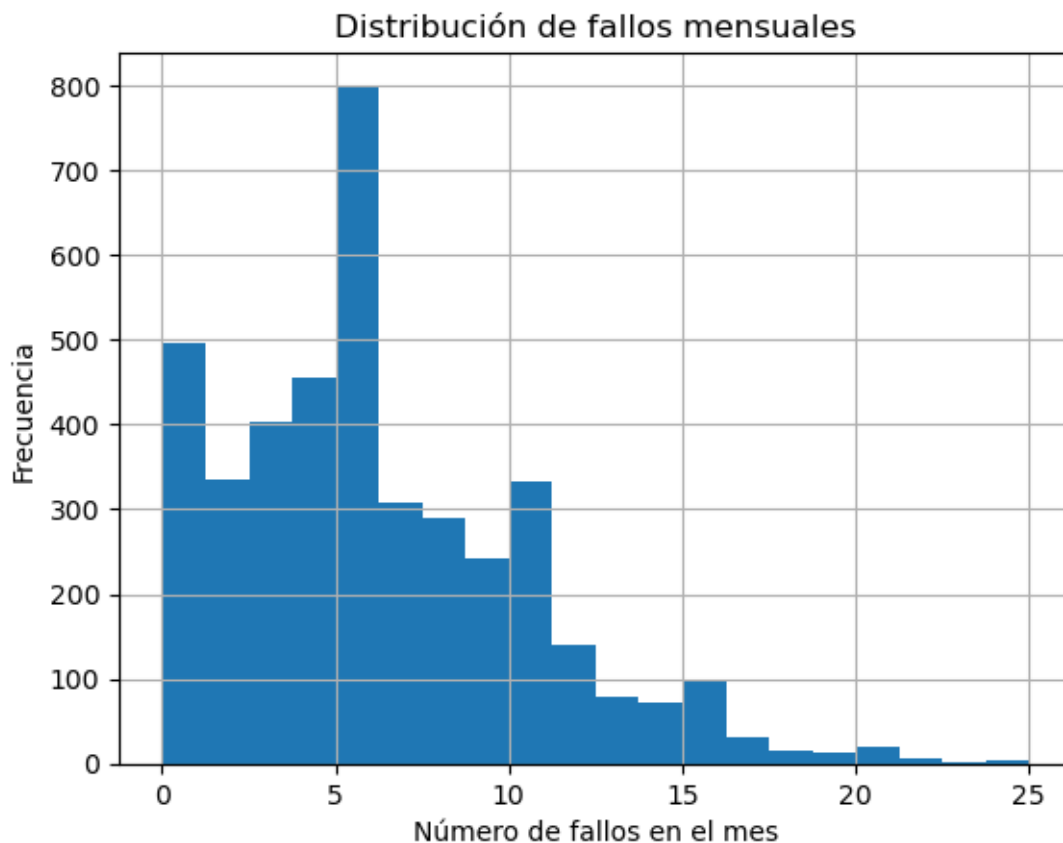
	Max_Temp	Parameter1_Speed	Parameter3_9am	Parameter3_3pm	...	\
0	14.550000	39.450000	11.950000	16.250000	...	
1	14.600000	36.105263	9.315789	15.631579	...	
2	20.234615	39.846154	14.730769	17.807692	...	
3	25.045833	37.291667	11.875000	17.458333	...	
4	24.642857	42.142857	12.607143	18.678571	...	

	Parameter6_9am	Parameter6_3pm	Parameter7_9am	Parameter7_3pm	Leakage	\
0	0.0	0.0	10.795000	13.615000	3.530000	
1	0.0	0.0	9.973684	13.484211	4.242105	
2	0.0	0.0	15.188462	19.211538	0.615385	
3	0.0	0.0	17.933333	23.941667	0.200000	
4	0.0	0.0	18.492857	23.110714	0.492857	

	LK	I_evap	I_elect	I_param6_9am	I_param6_3pm
0	0.700000	0	0	1	1
1	0.631579	0	0	1	1
2	0.346154	0	0	1	1
3	0.166667	0	0	1	1
4	0.321429	0	0	1	1

[5 rows x 24 columns]

```
[80]: df_mes_sensor['Failure_today'].hist(bins=20)
plt.title("Distribución de fallos mensuales")
plt.xlabel("Número de fallos en el mes")
plt.ylabel("Frecuencia")
plt.show()
```

```
[81]: parametros = [
    'Min_Temp', 'Max_Temp', 'Evaporation', 'Electricity',
    'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
    'Parameter4_9am', 'Parameter4_3pm', 'Parameter5_9am', 'Parameter5_3pm',
    'Parameter6_9am', 'Parameter6_3pm', 'Parameter7_9am', 'Parameter7_3pm',
    'I_evap', 'I_elect', 'I_param6_9am', 'I_param6_3pm', 'LK', 'Leakage'
]
```

```
formula = 'Failure_today ~ ' + ' + '.join(parametros)
modelo_poisson = smf.glm(formula=formula, data=df_mes_sensor, family=sm.
    families.Poisson()).fit()
print(modelo_poisson.summary())
```

Generalized Linear Model Regression Results

```
=====
Dep. Variable:          Failure_today    No. Observations:          4137
Model:                  GLM              Df Residuals:            4115
Model Family:           Poisson          Df Model:                 21
Link Function:          Log              Scale:                   1.0000
Method:                 IRLS             Log-Likelihood:          -8676.3
```

Date: Thu, 24 Apr 2025 Deviance: 3613.3
Time: 20:20:08 Pearson chi2: 3.05e+03
No. Iterations: 5 Pseudo R-squ. (CS): 0.8893
Covariance Type: nonrobust

```
=====
```

	coef	std err	z	P> z	[0.025
0.975]					

Intercept	6.2671	2.500	2.507	0.012	1.368
11.166					
Min_Temp	0.0264	0.007	3.792	0.000	0.013
0.040					
Max_Temp	0.0461	0.021	2.156	0.031	0.004
0.088					
Evaporation	-0.0334	0.005	-6.341	0.000	-0.044
-0.023					
Electricity	0.0108	0.006	1.791	0.073	-0.001
0.023					
Parameter1_Speed	0.0151	0.002	6.242	0.000	0.010
0.020					
Parameter3_9am	-0.0043	0.003	-1.478	0.139	-0.010
0.001					
Parameter3_3pm	-0.0142	0.003	-4.511	0.000	-0.020
-0.008					
Parameter4_9am	0.0109	0.002	5.228	0.000	0.007
0.015					
Parameter4_3pm	-0.0068	0.002	-2.942	0.003	-0.011
-0.002					
Parameter5_9am	-0.0051	0.013	-0.407	0.684	-0.030
0.019					
Parameter5_3pm	-0.0006	0.012	-0.048	0.962	-0.025
0.024					
Parameter6_9am	-0.0242	0.012	-2.036	0.042	-0.048
-0.001					
Parameter6_3pm	0.0636	0.011	5.614	0.000	0.041
0.086					
Parameter7_9am	0.0560	0.012	4.753	0.000	0.033
0.079					
Parameter7_3pm	-0.1199	0.024	-4.968	0.000	-0.167
-0.073					
I_evap	-0.0690	0.033	-2.092	0.036	-0.134
-0.004					
I_elect	0.0178	0.048	0.370	0.711	-0.076
0.112					
I_param6_9am	-0.3168	0.119	-2.661	0.008	-0.550
-0.083					

I_param6_3pm	0.3865	0.117	3.300	0.001	0.157
0.616					
LK	1.8186	0.055	33.105	0.000	1.711
1.926					
Leakage	0.0164	0.002	7.965	0.000	0.012
0.020					

```
=====
=====
```

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

R: El analisis muestra que existe una sobre dispersion de los datos. Esto se observa mejor cuando se compara la media con la varianza, donde la segunda es mucho mayor. Por otro lado, un posible valor óptimo de alpha se obtuvo como 0.9751.

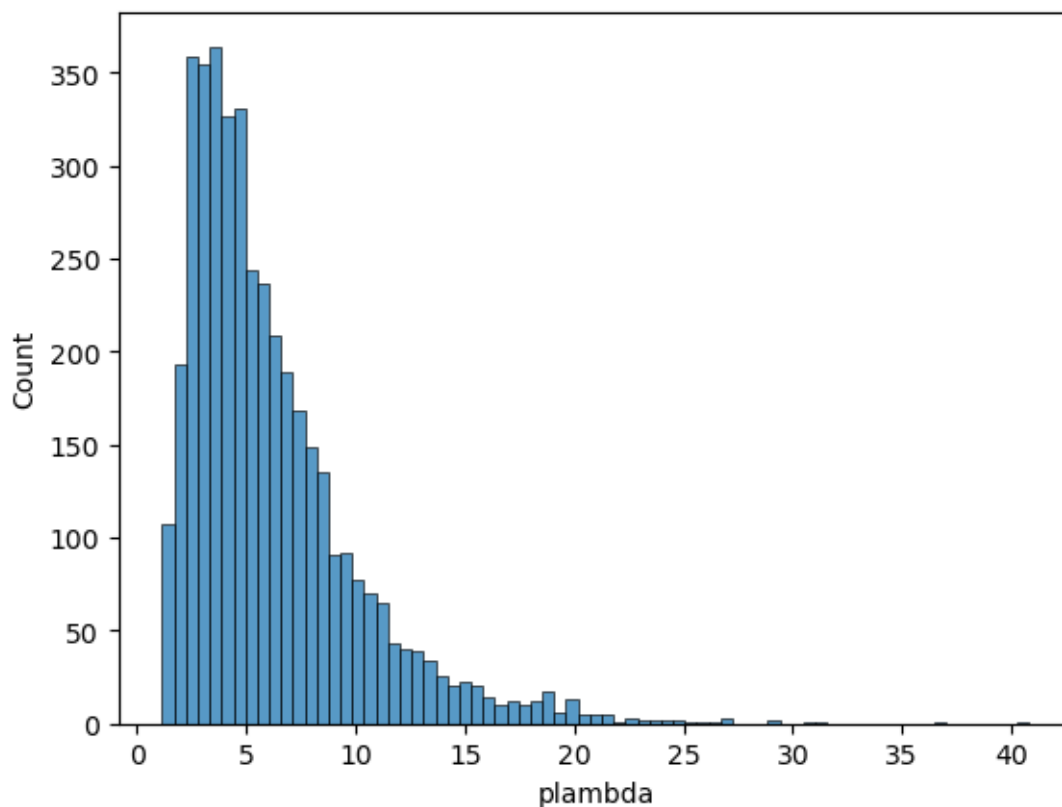
```
[100]: mean = df_mes_sensor['Failure_today'].mean()
var = df_mes_sensor['Failure_today'].var()
print("Media:", mean)
print("Varianza:", var)
```

Media: 6.132221416485375

Varianza: 18.016120620208767

```
[85]: df_mes_sensor['plambda'] = modelo_poisson.mu
sns.histplot(data=df_mes_sensor, x="plambda")
```

```
[85]: <Axes: xlabel='plambda', ylabel='Count'>
```



```
[101]: y = df_mes_sensor["Failure_today"]
mu = modelo_poisson.mu

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):
0.031
Model:                  OLS             Adj. R-squared (uncentered):
0.030
Method:                 Least Squares    F-statistic:
130.4
Date:                   Thu, 24 Apr 2025  Prob (F-statistic):
9.29e-30
Time:                   23:07:28          Log-Likelihood:
-6035.6
No. Observations:      4137             AIC:
```

```

1.207e+04
Df Residuals:          4136    BIC:
1.208e+04
Df Model:              1
Covariance Type:      nonrobust
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
x1          -0.0252      0.002     -11.419      0.000     -0.030     -0.021
=====
Omnibus:              4199.551    Durbin-Watson:              1.859
Prob(Omnibus):        0.000    Jarque-Bera (JB):        454552.246
Skew:                 4.751    Prob(JB):              0.00
Kurtosis:             53.465    Cond. No.              1.00
=====

```

Notes:

[1] R^2 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.

El posible valor óptimo de alpha se calcula como:

```
[103]: print(np.exp(-0.0252))
```

```
0.9751148695508249
```

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: Se concluye que varias de las variables en el modelo Binomial Negativo son estadísticamente significativas, cumpliendo el criterio de significancia con un valor p inferior a 0.05, como 'Evaporation', 'Parameter1_Speed', entre otros. El modelo explica aproximadamente el 31.18% de la variabilidad en la variable dependiente, que es el número de fallos por sensor observados en un mes. Las variables con mayor incidencia en el número de fallos son las de filtraciones. Por ejemplo, al aumentar LK, la probabilidad de que se presenten más fallos aumenta considerablemente, lo que sugiere una fuerte relación positiva entre esta variable y el número de fallos. Asimismo, al incrementarse la temperatura máxima, también aumenta la probabilidad de que ocurran más fallos en el sistema.

```
[93]: xx = df_mes_sensor[[
    'Min_Temp', 'Max_Temp', 'Evaporation', 'Electricity',
    'Parameter1_Speed', 'Parameter3_9am', 'Parameter3_3pm',
    'Parameter4_9am', 'Parameter4_3pm', 'Parameter5_9am', 'Parameter5_3pm',
    'Parameter6_9am', 'Parameter6_3pm', 'Parameter7_9am', 'Parameter7_3pm',
    'I_evap', 'I_elect', 'I_param6_9am', 'I_param6_3pm', 'LK', 'Leakage'
]]
```

```
negbin=sm.GLM(y,xx,family=sm.families.NegativeBinomial(alpha=0.9493)).fit()
print(negbin.summary())
```

Generalized Linear Model Regression Results

```
=====
Dep. Variable:          Failure_today    No. Observations:          4137
Model:                  GLM              Df Residuals:            4116
Model Family:          NegativeBinomial  Df Model:                  20
Link Function:          Log              Scale:                    1.0000
Method:                 IRLS             Log-Likelihood:            -11136.
Date:                   Thu, 24 Apr 2025  Deviance:                  950.26
Time:                   21:31:41          Pearson chi2:               617.
No. Iterations:         11               Pseudo R-squ. (CS):        0.3118
Covariance Type:        nonrobust
=====
```

```
=====
                                coef      std err          z      P>|z|      [0.025
0.975]
-----
-----
Min_Temp          0.0286      0.017      1.688      0.091      -0.005
0.062
Max_Temp          0.1079      0.056      1.940      0.052      -0.001
0.217
Evaporation       -0.0291      0.012     -2.367      0.018      -0.053
-0.005
Electricity       0.0026      0.015      0.175      0.861      -0.026
0.031
Parameter1_Speed  0.0159      0.006      2.636      0.008      0.004
0.028
Parameter3_9am    -0.0040      0.007     -0.548      0.584      -0.018
0.010
Parameter3_3pm    -0.0188      0.008     -2.240      0.025      -0.035
-0.002
Parameter4_9am    0.0189      0.005      3.616      0.000      0.009
0.029
Parameter4_3pm    -0.0208      0.006     -3.308      0.001      -0.033
-0.008
Parameter5_9am    -0.0437      0.033     -1.329      0.184      -0.108
0.021
Parameter5_3pm    0.0443      0.033      1.344      0.179      -0.020
0.109
Parameter6_9am    -0.0194      0.030     -0.642      0.521      -0.078
0.040
Parameter6_3pm    0.0778      0.029      2.670      0.008      0.021
0.135
Parameter7_9am    0.0962      0.030      3.192      0.001      0.037
0.155
=====
```

0.155					
Parameter7_3pm	-0.2240	0.063	-3.581	0.000	-0.347
-0.101					
I_evap	-0.0543	0.086	-0.631	0.528	-0.223
0.114					
I_elect	0.0081	0.125	0.065	0.948	-0.236
0.252					
I_param6_9am	-0.2714	0.319	-0.851	0.395	-0.896
0.353					
I_param6_3pm	0.3785	0.314	1.204	0.229	-0.238
0.995					
LK	2.2247	0.158	14.085	0.000	1.915
2.534					
Leakage	0.0400	0.007	5.335	0.000	0.025
0.055					

=====

=====

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 investigación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: Se pudo observar una clara sobredispersión (varianza > media), por lo que el modelo Poisson no es el más adecuado para estos datos. Es por eso que, en mi opinión, la Binomial Negativa es un modelo más apto, puesto que este permite modelar la sobredispersión de forma explícita. Así, 'Evaporation', 'Parameter1_Speed', 'Parameter4_9am', 'Parameter7_3pm' y 'LK' resultaron ser consistentes en los dos modelos.