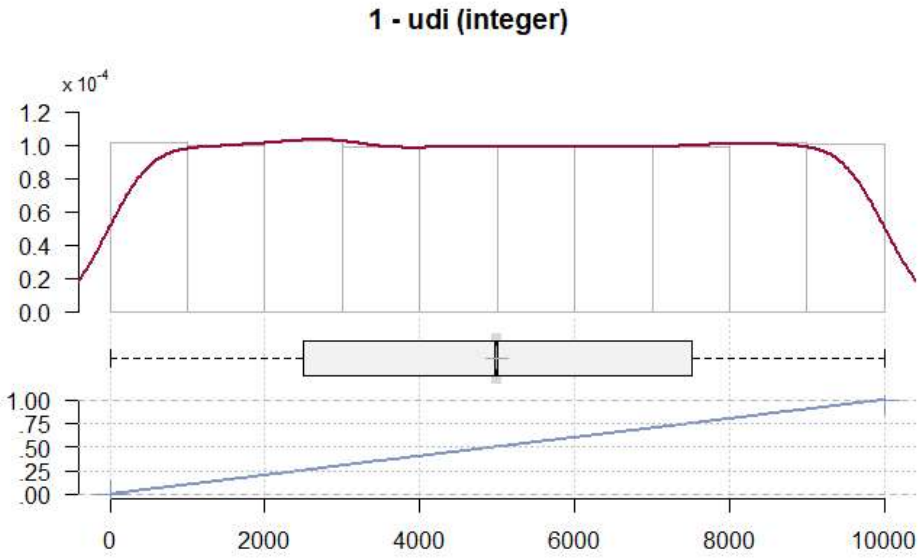
Para visualizar os dados disponíveis, podemos utilizar gráficos de distribuição de frequências e percentuais para dados qualitativos e curvas de distribuição e boxplots para dados quantitativos.

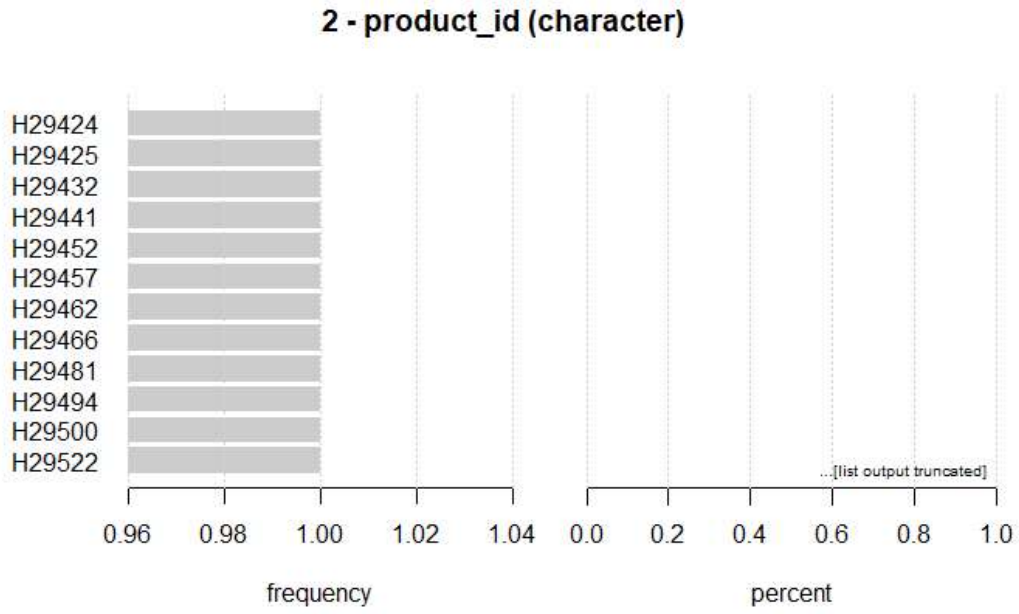
No caso dos dados qualitativos, a distribuição de frequências permite ver a quantidade de cada categoria presente na amostra e o percentual de cada uma em relação ao total. Já nos dados quantitativos, a curva de distribuição ajuda a entender a forma como os dados estão distribuídos, enquanto o boxplot fornece informações sobre as principais medidas estatísticas, como mínimo, máximo, intervalo interquartil, média e mediana.

Essas estatísticas foram escolhidas de acordo com o tipo de dado, visando a melhor compreensão da distribuição dos dados e das principais medidas estatísticas.

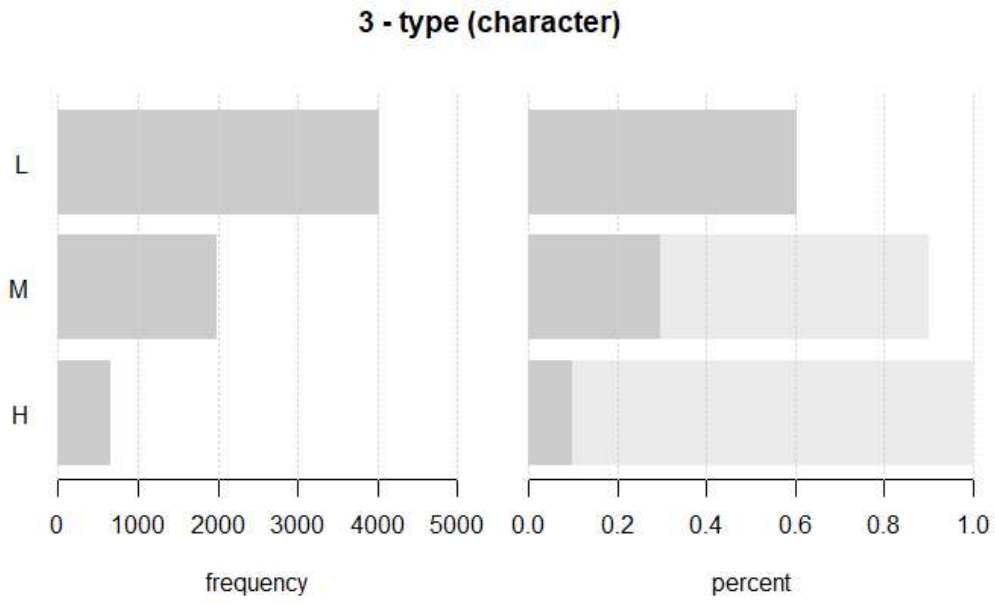
udi:
número sequencial da produção, um dos identificadores disponíveis.




product_id:
número de série dos produtos, identifica sua variante, e seu número de série.




type:
Identifica o tipo de produto, se L, M ou H.



```
3 - type (character)

  length      n    NAs unique levels  dupes
   6'667  6'667      0      3      3      y
         100.0%   0.0%

   level   freq   perc  cumfreq  cumperc
1      L  4'022  60.3%   4'022    60.3%
2      M  1'987  29.8%   6'009    90.1%
3      H    658   9.9%   6'667   100.0%
```

air_temperature_k:
Distribuição da temperatura do ar durante o processo de fabricação:

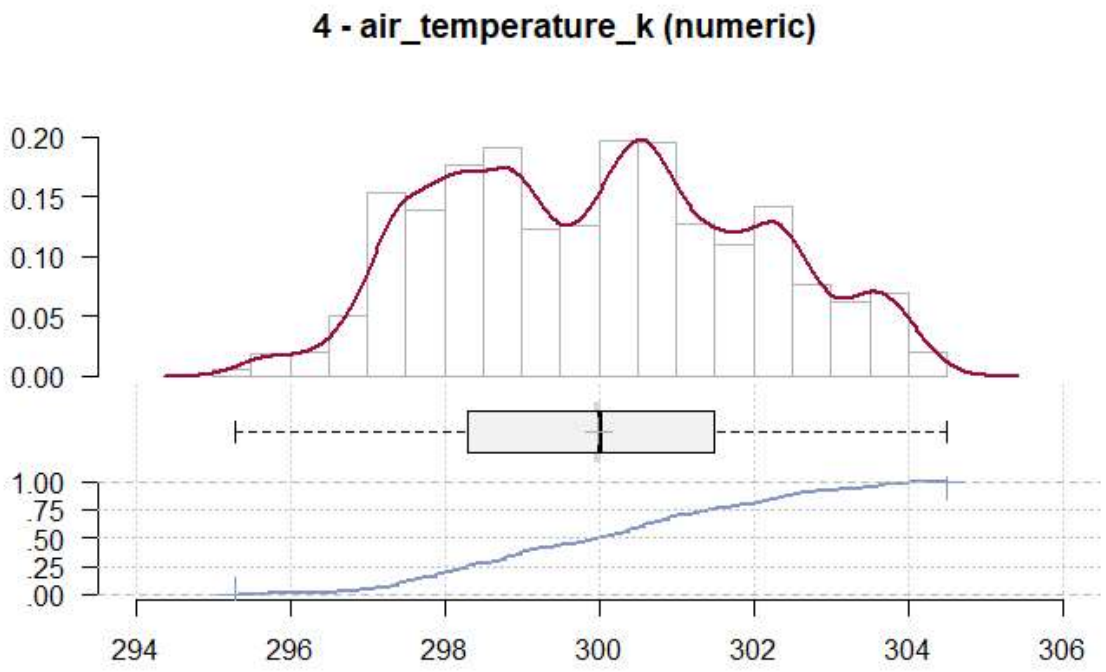
```
4 - air_temperature_k (numeric)

  length       n     NAs  unique      0s    mean  meanCI'
   6'667   6'667       0      93       0  299.99  299.94
          100.0%    0.0%            0.0%          300.04

     .05     .10     .25  median     .75     .90     .95
  297.10  297.40  298.30  300.00  301.50  302.70  303.50

   range      sd   vcoef     mad     IQR    skew    kurt
    9.20    1.99    0.01    2.37    3.20    0.13   -0.82

lowest : 295.3 (2), 295.4 (2), 295.5 (15), 295.6 (22), 295.7 (14)
highest: 304.1 (29), 304.2 (25), 304.3 (8), 304.4 (5), 304.5
```




process_temperature_k:
Distribuição da temperatura do processo de fabricação.

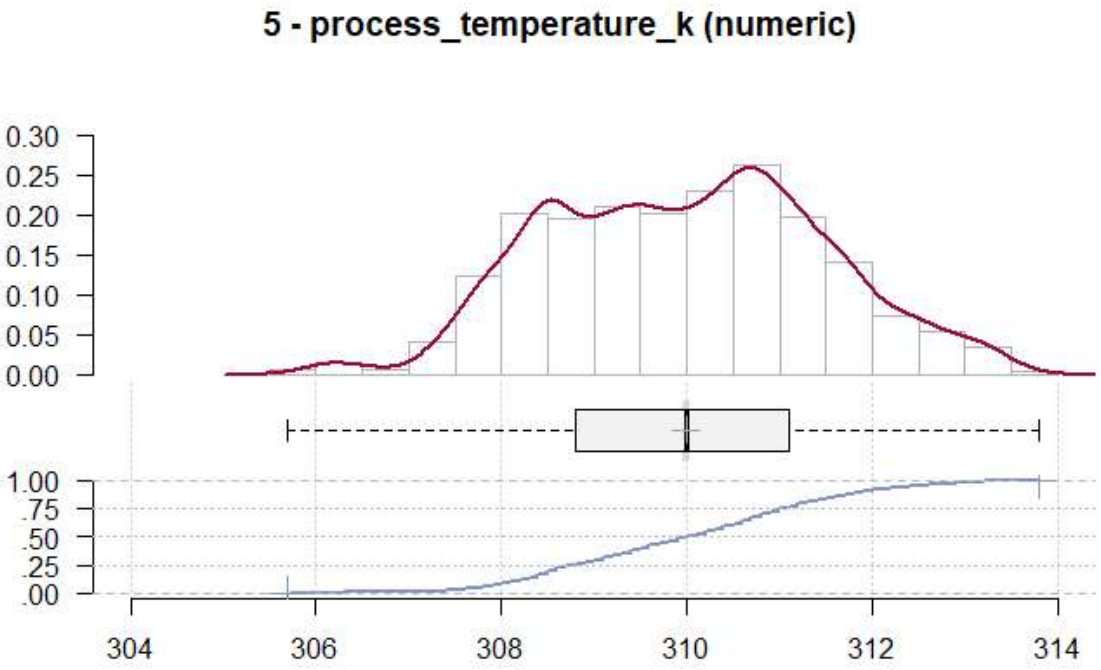
```
5 - process_temperature_k (numeric)

  length       n     NAs  unique      0s    mean  meanCI'
   6'667   6'667       0      82       0  309.99   309.96
          100.0%    0.0%            0.0%           310.03

     .05     .10     .25  median     .75     .90     .95
  307.70  308.10  308.80  310.00  311.10  311.90  312.50

   range      sd   vcoef     mad     IQR    skew    kurt
    8.10    1.49    0.00    1.63    2.30    0.02   -0.51

lowest : 305.7, 305.8 (2), 305.9 (5), 306.0 (11), 306.1 (10)
highest: 313.4 (18), 313.5 (14), 313.6 (8), 313.7 (3), 313.8
```




rotational_speed_rpm:
Distribuição das rotações por minuto durante o processo de fabricação.

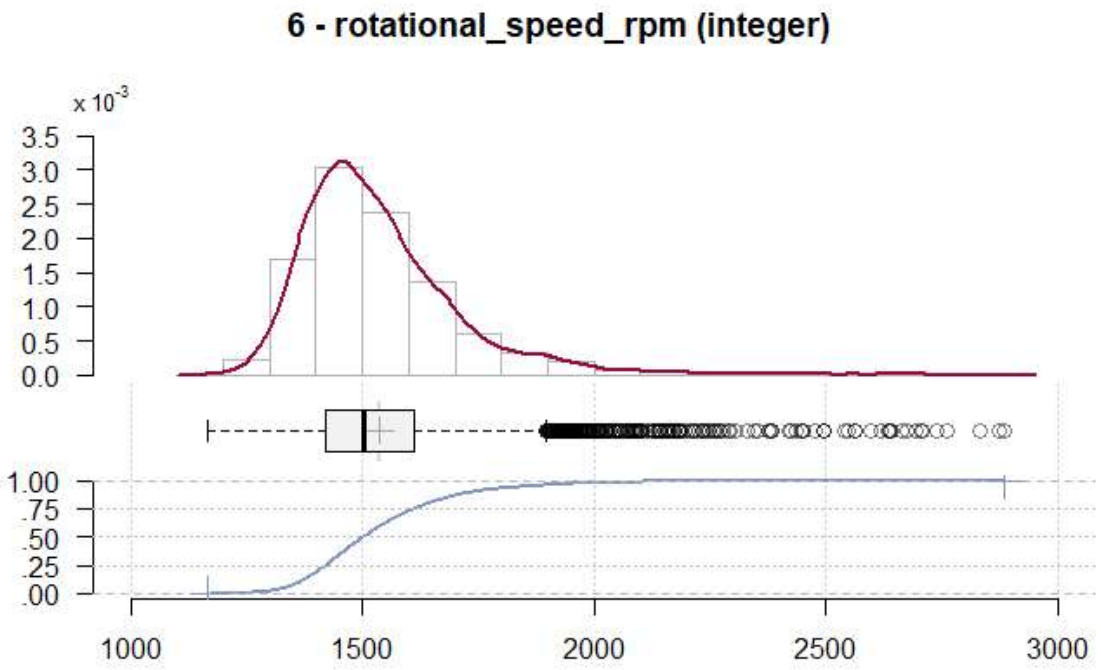
```
6 - rotational_speed_rpm (integer)

    length         n       NAs    unique        0s      mean    meanCI'
     6'667     6'667         0       840         0  1'537.42  1'533.17
              100.0%      0.0%             0.0%            1'541.67

       .05       .10       .25    median       .75       .90       .95
  1'332.00  1'363.00  1'422.50  1'503.00  1'612.00  1'742.00  1'867.70

     range        sd      vcoef       mad       IQR      skew      kurt
  1'718.00    177.18       0.12    134.92    189.50      1.95      7.30

lowest : 1'168, 1'181, 1'183, 1'192, 1'200
highest: 2'737, 2'760, 2'833, 2'874, 2'886
```




torque_nm:
distribuição do torque em Nm.

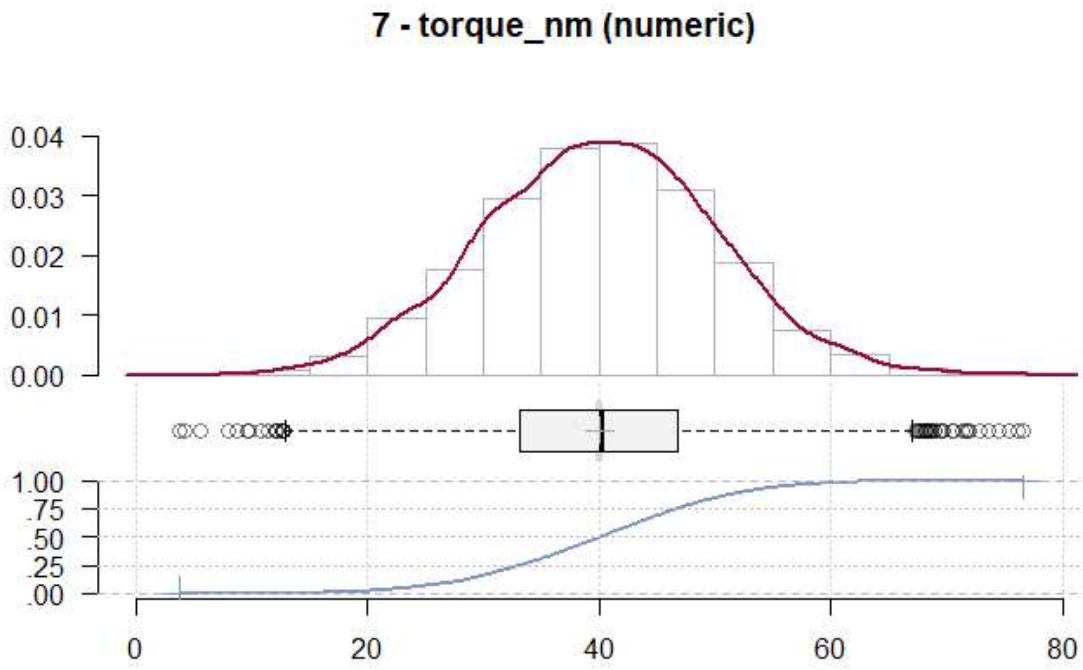
```
7 - torque_nm (numeric)

  length       n    NAs  unique     0s   mean  meanCI'
   6'667   6'667      0     547      0  40.06    39.82
          100.0%   0.0%          0.0%          40.30

     .05     .10     .25  median    .75    .90     .95
   23.50   27.40   33.20   40.20  46.80  52.64   56.20

   range      sd  vcoef     mad    IQR   skew    kurt
   72.80    9.95   0.25   10.08  13.60   0.01    0.01

lowest : 3.8, 4.2, 5.6, 8.0, 8.8
highest: 73.6, 74.5, 75.4, 76.2, 76.6
```




tool_wear_min:
Distribuição do tempo de desgaste da ferramenta.

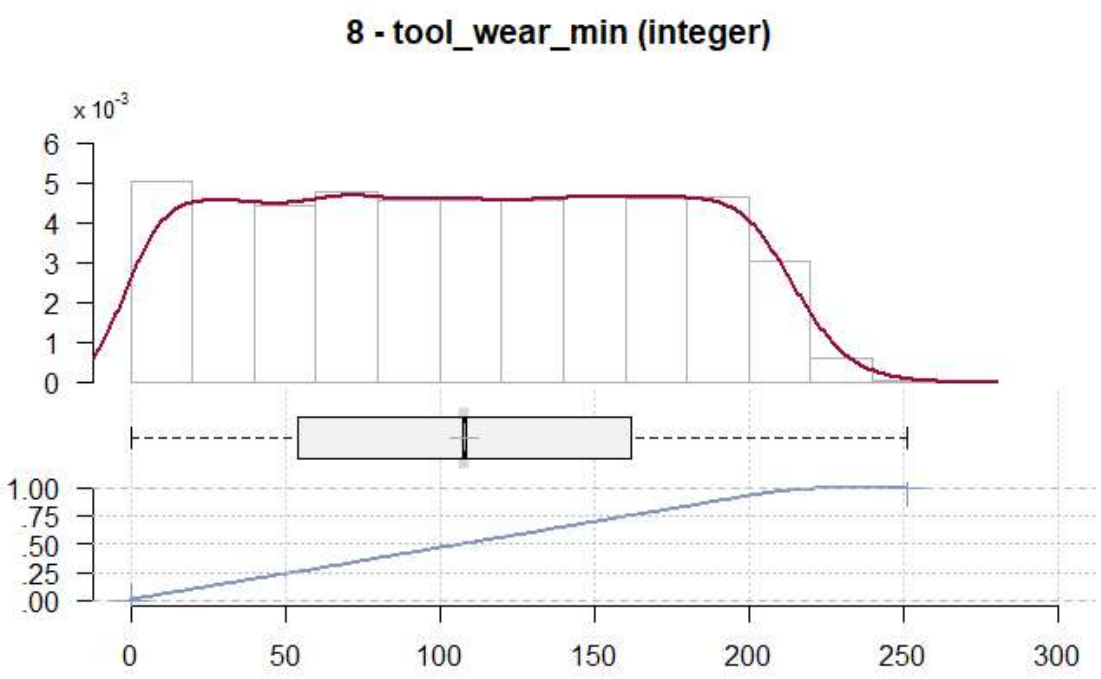
```
8 - tool_wear_min (integer)

  length       n     NAs  unique      0s    mean  meanCI'
   6'667   6'667       0     243      82  108.10  106.58
          100.0%   0.0%          1.2%          109.62

     .05     .10     .25  median     .75     .90     .95
    9.00   20.00   54.00  108.00  162.00  195.00  206.00

   range      sd  vcoef     mad     IQR    skew    kurt
  251.00   63.36   0.59   80.06  108.00    0.02   -1.16

lowest : 0 (82), 2 (50), 3 (23), 4 (22), 5 (42)
highest: 241, 242, 244 (3), 246 (2), 251
```

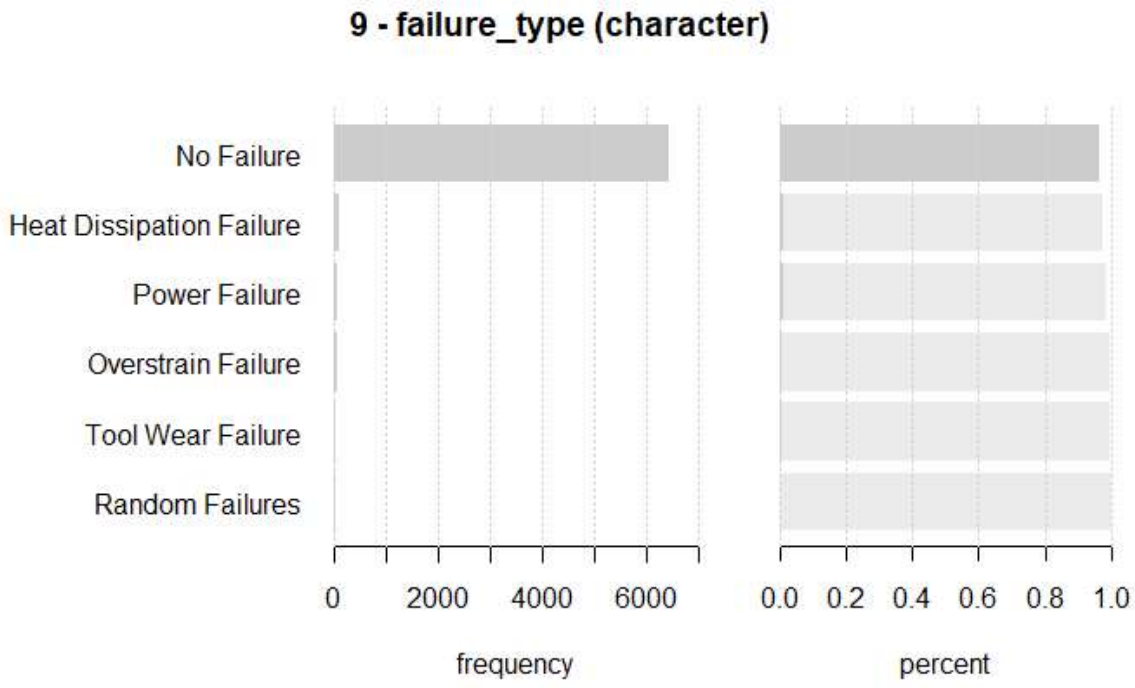



failure_type:
Descrição dos tipos de falha que ocorreram.

**9 - failure_type (character)**

| length | n | NAs | unique levels | dupes |
|---|---|---|---|---|
| 6'667 | 6'667 | 0 | 6 | 6 | y |
| | 100.0% | 0.0% | | |

| | level | freq | perc | cumfreq | cumperc |
|---|---|---|---|---|---|
| 1 | No Failure | 6'435 | 96.5% | 6'435 | 96.5% |
| 2 | Heat Dissipation Failure | 75 | 1.1% | 6'510 | 97.6% |
| 3 | Power Failure | 63 | 0.9% | 6'573 | 98.6% |
| 4 | Overstrain Failure | 52 | 0.8% | 6'625 | 99.4% |
| 5 | Tool Wear Failure | 30 | 0.4% | 6'655 | 99.8% |
| 6 | Random Failures | 12 | 0.2% | 6'667 | 100.0% |




Modelando a previsão:

Para prever o tipo de falha a partir dos dados, seria possível utilizar um modelo de regressão logística, que permite trabalhar com um resultado categórico, como "falha" ou "não falha". No entanto, essa abordagem tem a limitação de trabalhar apenas com respostas binárias.

Foram criados alguns dados a partir de transformações da base de dados tanto para o conjunto de treino quanto para o conjunto de teste. As transformações incluíram:

"failure": onde todos os tipos de falha foram transformados em um único resultado ("1"), enquanto as não falhas foram transformadas em outro resultado ("0");
"heat": sendo a diferença entre a temperatura do processo e a temperatura do ar;
"power": resultado do produto entre o torque e a rotação do motor em rad/s;
"overstrain": o produto entre o torque e o desgaste da ferramenta.
Esse é um problema de classificação, em que estamos tentando prever se algo vai falhar ou não. A regressão logística é uma boa opção para esse tipo de problema, pois permite lidar com variáveis categóricas e pode ser treinado com os dados disponíveis. Seus prós incluem a simplicidade e interpretabilidade do modelo, além de ser relativamente rápido para treinar e aplicar.

A medida de performance escolhida para avaliar a predição foi a curva ROC teórica com base na modelagem sobre os dados de treino para encontrar o valor que traga a maior Sensibilidade e a maior Especificidade. Essa medida permite avaliar a capacidade do

modelo de classificar corretamente os casos positivos e negativos, considerando um trade-off entre a sensibilidade e a especificidade.