

F classif

É o cálculo do F-value da estatística. Basicamente ele verifica o quão distante da média de um grupo está do outro, caso seja distante o F-value é maior, e muito próximo será menor.

Exemplo:

The image shows a handwritten calculation of the F-value for three groups (A, B, and C). The data is as follows:

Group	Sample 1	Sample 2	Sample 3
A	2	1	3
B	6	4	5
C	9	9	7

The mean for each group is calculated as $\mu_A = 2$, $\mu_B = 5$, and $\mu_C = 8$. The overall mean is $\mu = 5$.

The Sum of Squares Due to Error (SSE) is calculated as:

$$SSE = (2-5)^2 + (1-5)^2 + (3-5)^2 + (6-5)^2 + (4-5)^2 + (5-5)^2 + (9-8)^2 + (9-8)^2 + (7-8)^2 = 54$$

The Sum of Squares Due to Treatment (SST) is calculated as:

$$SST = (2-2)^2 + (1-2)^2 + (3-2)^2 + (6-5)^2 + (4-5)^2 + (5-5)^2 + (8-8)^2 + (9-8)^2 + (7-8)^2 = 6$$

The F-value is then calculated as:

$$F\text{-VALUE} = \frac{\frac{SSE}{m-1}}{\frac{SST}{m(m-1)}} = \frac{\frac{54}{3-1}}{\frac{6}{3(3-1)}} = 27$$

Primeiro é calculado a média de cada grupo A, B e C:

$$A = 2$$

$$B = 5$$

$$C = 8$$

Média geral: 5

SQD = Somatório de cada grupo, sendo o valor da amostra de um grupo X - a média do grupo X elevado ao quadrado

$$SQD = (2-2)^2 + (1-2)^2 + (3-2)^2 + (6-5)^2 + (4-5)^2 + (5-5)^2 + (8-8)^2 + (9-8)^2 + (7-8)^2 = 6$$

SQD = Vai utilizar a média geral e a média do grupo repetindo por amostras de cada grupo:

$$SQE = (2-5)^2 + (2-5)^2 + (2-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + (8-5)^2 + (8-5)^2 + (8-5)^2 = 54$$

F-VALUE = 27

Vai se repetindo, pois há 3 valores de cada grupo

Por fim é só calcular o F-value:

$$F-VALUE = \frac{\frac{54}{3-1}}{\frac{6}{3(3-1)}} = 27$$

Sendo M o número de amostras = 3

e N o número de linhas = 3

Com isso conseguimos utilizar o conceito do F-value para datasets na seleção de features para problemas de classificação. Basta dividir uma coluna do dataframe em N grupos, sendo grupos a quantidade de classes, 0,1 ou N classes. Assim podemos extrair a média de cada grupo e verificar se elas são diferentes, caso sejam, a variável contribui bastante para a classificação do dado, caso sejam próximas, ou seja, um valor baixo de F-value, não é uma variável boa para a classificação. No fim queremos ver se a média dos dados para cada classe é diferente, para haver uma melhor divisão na hora de classificar os dados.