Modelo Matemático para Previsão da Cor da Cerveja

1 Descrição do Modelo

O desafio consiste em prever a cor da cerveja logo após a etapa de resfriamento no processo de fabricação. O framework segue a metodologia CRISP-DM com as seguintes etapas: 1. Entendimento do Negócio; 2. Entendimento e Preparação dos Dados; 3. Segmentação das Etapas de Fabricação; 4. Modelagem; 5. Avaliação.

2 Formulação Matemática

2.1 1. Pré-processamento dos Dados

2.1.1 1.1 Interpolação Linear

$$x_i = \frac{x_{i+1} + x_{i-1}}{2}$$

onde:

- x_i é o valor interpolado para a posição i onde há dados ausentes.
- x_{i+1} e x_{i-1} são os valores adjacentes a x_i utilizados para interpolação.

2.1.2 1.2 Transformação de Escala

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

onde:

- x' é o valor normalizado.
- x é o valor original da variável.
- $x_{\min}$ e $x_{\max}$ representam os valores mínimo e máximo de x na escala original.

2.2 2. Segmentação de Etapas com KMeans

Utilizando o algoritmo KMeans com $k = 10$, a função objetivo é minimizar:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} 1(c_i = j)\|x_i - \mu_j\|^2$$

onde:

- n é o número total de pontos de dados.
- k é o número de clusters (aqui definido como 10).
- c_i é o cluster atribuído ao ponto de dados x_i.
- μ_j é o centroide do cluster j.
- $\|x_i - \mu_j\|^2$ representa a distância quadrada entre o ponto x_i e o centroide μ_j.

2.3 3. Modelagem com Rede Neural LSTM

2.3.1 Equações da LSTM

1. Esquecimento:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f)$$

2. Atualização:

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, X_t] + b_C)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{C}_t$$

3. Saída:

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o)$$
$$h_t = o_t \cdot \tanh(c_t)$$

4. Camada de saída:

$$\hat{y} = W_y \cdot h_t + b_y$$

onde:

- f_t, i_t, o_t são as portas de esquecimento, entrada e saída, respectivamente.
- W_f, W_i, W_C, W_o e W_y são matrizes de pesos.
- b_f, b_i, b_C, b_o, b_y são os vetores de bias.

- h_{t-1} é o estado oculto anterior.
- X_t é a entrada no instante t.
- c_t é o estado da célula.
- $\hat{y}$ é a predição final.

2.4 4. Função de Custo

Para avaliar o erro, utiliza-se a função de erro quadrático médio (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

onde:

- y_i é o valor observado para o ponto i.
- $\hat{y}_i$ é o valor previsto para o ponto i.
- n é o número total de pontos.

2.5 5. Avaliação do Modelo

2.5.1 Erro Médio Absoluto (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

onde:

- y_i é o valor observado.
- $\hat{y}_i$ é o valor previsto.

2.5.2 Coeficiente de Determinação (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

onde:

- y_i é o valor observado.
- $\hat{y}_i$ é o valor previsto.
- $\bar{y}$ é a média dos valores observados.

2.5.3 Raiz do Erro Quadrático Médio (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

onde:

- y_i é o valor observado.
- $\hat{y}_i$ é o valor previsto.
- n é o número total de pontos.