Aplicação de Modelos de Aprendizado Supervisionado para Regressão e Classificação de Dados Biológicos

Danilo Silva Pessoa de Araújo
UNIFOR - 2218928
Turma - T292-40

Cássio Tiago Holanda de Castro
UNIFOR - 2316229
Turma - T296-40

March 29, 2025

Abstract

Este relatório apresenta os resultados de um estudo que investiga a aplicação de modelos de aprendizado supervisionado para tarefas de regressão e classificação, utilizando dados biológicos. Na primeira etapa, exploramos a regressão linear e suas variações (MQO tradicional e regularizado) para prever a atividade enzimática com base em temperatura e pH. Na segunda etapa, aplicamos modelos gaussianos bayesianos e MQO para classificar expressões faciais a partir de sinais eletromiográficos (EMG). Avaliamos o desempenho dos modelos por meio de simulações de Monte Carlo, utilizando a soma dos desvios quadráticos (RSS) para regressão e acurácia para classificação. Os resultados demonstram a eficácia dos modelos em ambas as tarefas, fornecendo insights valiosos sobre a relação entre variáveis biológicas e seus efeitos observáveis.

1 Introdução

O presente trabalho é composto por duas etapas em que deve-se utilizar os conceitos de IA baseados em modelos preditivos que realizam seu processo de aprendizagem através da minimização de uma função custo (loss function). Em ambas etapas do trabalho, tais modelos utilizam o paradigma supervisionado para aprender a partir dos pares, vetor de características (variáveis regressoras) e variável dependente. Contudo, a tarefa da primeira etapa trata-se do desenvolvimento de um sistema que faz predições quantitativas (regressão), ao passo que a segunda etapa é caracterizada pelo desenvolvimento de um sistema que realiza predições qualitativas (classificação).

2 Metodologia

2.1 Regressão

Os dados são carregados a partir de um arquivo CSV denominado `atividade_enzimatica.csv`, contendo três colunas: temperatura, pH e atividade enzimática. Eles são organizados em três vetores:

- $Eixo\_x\_temp$: Temperatura

- $Eixo\_x\_ph$: pH
- $Eixo\_y$: Atividade Enzimática

Para modelagem, uma matriz de variáveis explicativas (X_matriz) é construída adicionando um termo de intercepto (coluna de 1s) aos vetores de temperatura e pH. Para reduzir problemas de sobreajuste, utiliza-se a regularização com múltiplos valores de λ (0, 0.25, 0.5, 0.75, 1). Matrizes identidade escaladas por λ são pré-computadas para otimizar os cálculos. Os coeficientes da regressão são calculados para cada valor de λ usando a equação:

$$\beta = (X^T X + \lambda I)^{-1} X^T Y \tag{1}$$

Os valores preditos são obtidos e representados em um gráfico 3D junto aos dados reais, com diferentes superfícies de regressão coloridas para cada λ. Para avaliar o desempenho do modelo, um processo de validação cruzada é realizado: Os dados são embaralhados e divididos em 80% para treino e 20% para teste. O modelo é ajustado aos dados de treino e testado no conjunto de teste. O erro quadrático residual (RSS) é calculado para cada λ e também para um modelo de referência baseado na média dos valores de treino. Esse processo é repetido 500 vezes para reduzir variabilidade. Os valores médios, desvio padrão, mínimo e máximo do RSS são calculados e apresentados em tabelas comparativas para cada λ, juntamente com os resultados do modelo baseado na média.

2.1.1 Modelos

Foram implementados os seguintes modelos:

- Mínimos Quadrados Ordinários (MQO) tradicional:

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

  Calcula-se a inversa de $X^T X$, multiplicando-a pela transposta de X e pelo vetor de resultados observados y, gerando assim os coeficientes β que melhor descrevem a relação entre as variáveis.

- MQO regularizado (Tikhonov) com $\lambda = \{0, 0.25, 0.5, 0.75, 1\}$:

  $$\mathbf{W} = (X^T X + \lambda \mathbf{I})^{-1} X^T Y$$

  O termo λI adiciona uma regularização à matriz $X^T X$, onde I é a matriz identidade e λ controla a intensidade da regularização. A inversa de $(X^T X + \lambda I)$ é então multiplicada pela transposta de X e pelo vetor de respostas Y, gerando os coeficientes regularizados W.

- Média dos valores observáveis:

  $$\mathrm{RSS}_{\text{média}} = \sum_{i=1}^{n} (\bar{y} - y_i)^2$$

  A fórmula mede o desvio entre a média das observações e cada valor observado, fornecendo uma medida de quão bem o modelo está representando os dados. Quanto menor o valor de $\mathrm{RSS}_{\text{média}}$, melhor o modelo se ajusta aos dados.

2.2 Classificação

Os dados utilizados no experimento consistem em sinais de eletromiografia (EMG) coletados a partir de diferentes expressões faciais. A base de dados foi processada para remover ruídos e normalizar os valores. O pré-processamento dos sinais EMG incluiu: Filtragem para remoção de ruídos; Normalização dos dados para escala padrão; Separação dos dados em conjunto de treino e teste. Foram utilizados diferentes classificadores para prever as expressões faciais:

- **Regressão MQO**: Modelo de regressão linear baseado em mínimos quadrados;
- **Modelo Gaussiano com matriz agregada**: Com uma matriz de covariância agregada, ponderada por cada classe

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln P(w_i) \tag{2}$$

- **Covariância Igual**: Variante do modelo gaussiano onde a matriz de covariância é compartilhada entre as classes;
- **Covariância Agregada**: Usa uma única matriz de covariância para todas as classes;
- **Naive Bayes**: Assume independência condicional entre as características;
- **Friedman**: Extensão do modelo Naive Bayes que incorpora dependências entre as variáveis.

3 Avaliação dos Modelos

Para medir o desempenho dos classificadores, foram utilizadas as seguintes métricas:

- Acurácia média;
- Desvio padrão da acurácia;
- Valores máximo e mínimo obtidos.

A avaliação foi realizada utilizando validação cruzada para reduzir viés nos resultados. A fronteira de decisão dos classificadores foi plotada para ilustrar as regiões de decisão dos modelos em um espaço bidimensional. Isso permitiu a análise visual do comportamento de cada classificador. A metodologia apresentada permite comparar diferentes abordagens para classificação de expressões faciais a partir de sinais EMG. A análise das métricas e das fronteiras de decisão auxilia na escolha do modelo mais adequado para o problema em questão.

3.0.1 Modelos

Foram implementados os seguintes modelos:

- MQO tradicional representado pela formula;

$$w = (X^T X)^{-1} X^T y$$

- Classificador Gaussiano Tradicional;

$$D_{\text{Mahalanobis}}(x, \mu_i, \Sigma_i^{-1}) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)$$

- Classificador Gaussiano com covariâncias iguais;

$$p(x \mid \mu_i, \Sigma^{-1}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \right)$$

- Classificador Gaussiano com matriz de covariância agregada;

$$\Sigma_{\text{agregada}} = \sum_{i=1}^{k} \frac{n_i}{N} \Sigma_i$$

  Onde n_i é o número de amostras da classe i, N é o número total de amostras, e Σ_i é a matriz de covariância de cada classe.

- Classificador gaussiano regularizado (Friedman) com $\lambda = \{0.25, 0.5, 0.75\}$;

$$\Sigma_i^{\lambda} = \frac{(1 - \lambda)(n_i \cdot \Sigma_i) + (\lambda \cdot N \cdot \Sigma_{\text{agregada}})}{(1 - \lambda) n_i + \lambda \cdot N}$$

  onde λ varia de 0 a 1, com incrementos de 0,25. No caso em que $\lambda = 1$, a função discriminante utilizada assume a seguinte forma:

- Classificador Naive Bayes;

$$p(y = i \mid x) = \frac{p(x \mid y = i) p(y = i)}{p(x)}$$

  Onde a função densidade de probabilidade $p(x \mid y = i)$ para a classe i é:

$$p(x \mid y = i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right)$$

3.0.2 Validação

A validação foi realizada por meio de simulações de Monte Carlo (500 rodadas), com particionamento de 80% dos dados para treinamento e 20% para teste.

3.0.3 Análise

O MQO tradicional apresenta uma média de 72.365, o que é razoável, mas o desvio padrão de 28 indica uma grande variação nos resultados. Isso sugere que o modelo pode ser inconsistente, com resultados bastante dispersos, o que pode indicar a necessidade de ajustes ou a utilização de outro modelo para maior estabilidade e precisão. O Gaussiano Regularizado (Friedman $= 0.25$) parece ser o modelo mais estável e com o melhor desempenho geral, dado que tem a maior média e o menor desvio padrão, indicando consistência nos resultados. O Gaussiano (Covariância Agregada) e o Gaussiano (Covariância de todo o conjunto de treino) têm boas médias, mas os altos desvios padrão indicam que esses modelos têm variações significativas nos resultados, o que pode ser problemático em tarefas que exigem precisão. O Naive Bayes apresenta uma boa média com um desvio padrão razoável, sendo um modelo equilibrado em termos de desempenho e estabilidade. O Classificador Gaussiano Tradicional tem a menor média, o que indica que não é adequado para essa tarefa específica, apesar de sua consistência.

4 Resultados da Regressão

4.1 Tabela 1

Modelos	Média	Desvio Padrão	Maior Valor	Menor Valor
Média da variável dependente	22.876	1.2385	19.4423	27.0581
MQO tradicional	4.33	0.43	5.76	3.29
MQO regularizado (0,25)	4.33	0.43	5.75	3.3
MQO regularizado (0,5)	4.33	0.43	5.74	3.3
MQO regularizado (0,75)	4.33	0.42	5.74	3.31
MQO regularizado (1)	4.33	0.42	5.75	3.32

Table 1: Resultados da Regressão

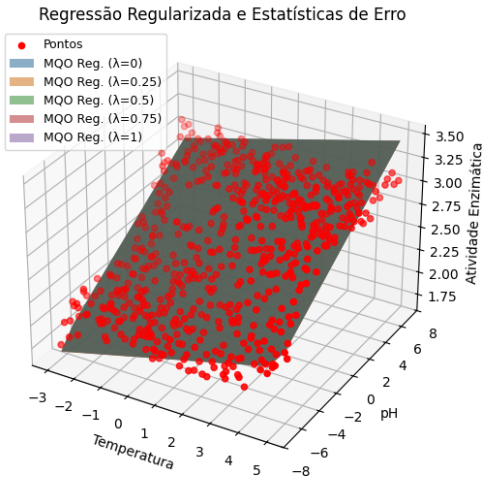

Figure 1: Gráfico de Resultados da Regressão

4.2 Classificação

Modelos	Média	Desvio Padrão	Maior Valor	Menor Valor
MQO tradicional	72.365 28	0.620	74.48	70.43
Classificador Gaussiano Tradicional	19.97	0.355	21.0	18.7
Gaussiano (Cov. de todo cj. treino)	94.819 48	0.2099	95.46	94.24
Gaussiano (Cov. Agregada)	96.238 92	0.175	96.76	94.24
Bayes Ingênuo (Naive Bayes Classifier)	95.749 72	0.188	96.31	96.0099
Gaussiano Regularizado (Friedman $\lambda = 0,25$)	97.46	0.1470	98.0	97.06
Gaussiano Regularizado ($\lambda = 0,5$)	96.782 76	0.1622	97.289	96.3
Gaussiano Regularizado ($\lambda = 0,75$)	96.4696	0.1680	96.9	96.009

Table 2: Resultados da Classificação

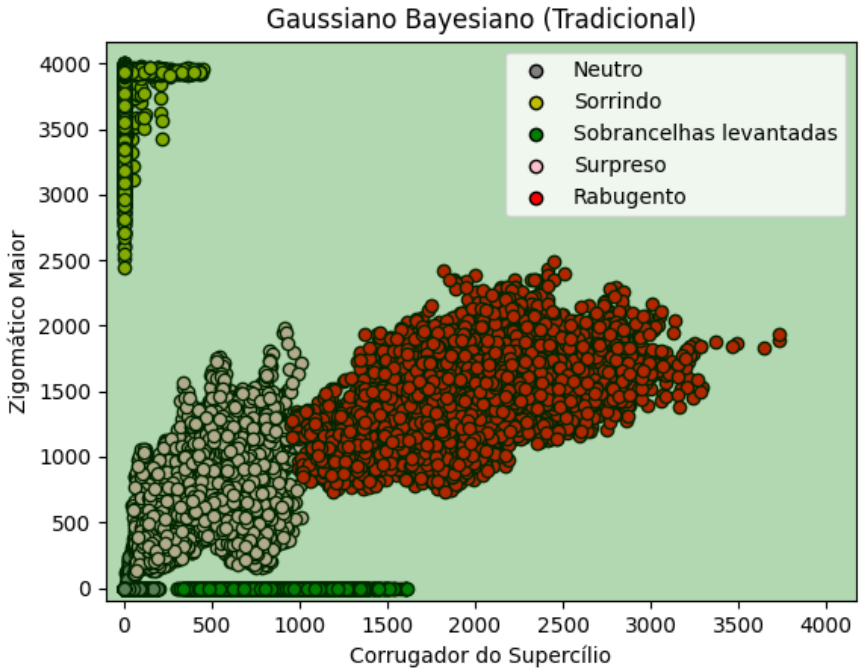

Figure 2: Gráfico de Resultados da Classificação

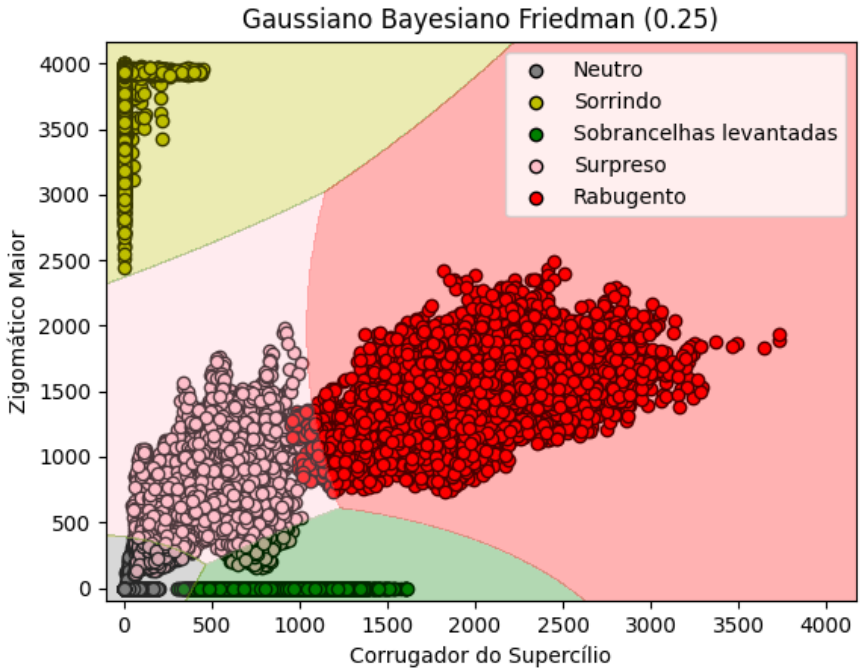

Figure 3: Gráfico de Resultados da Classificação

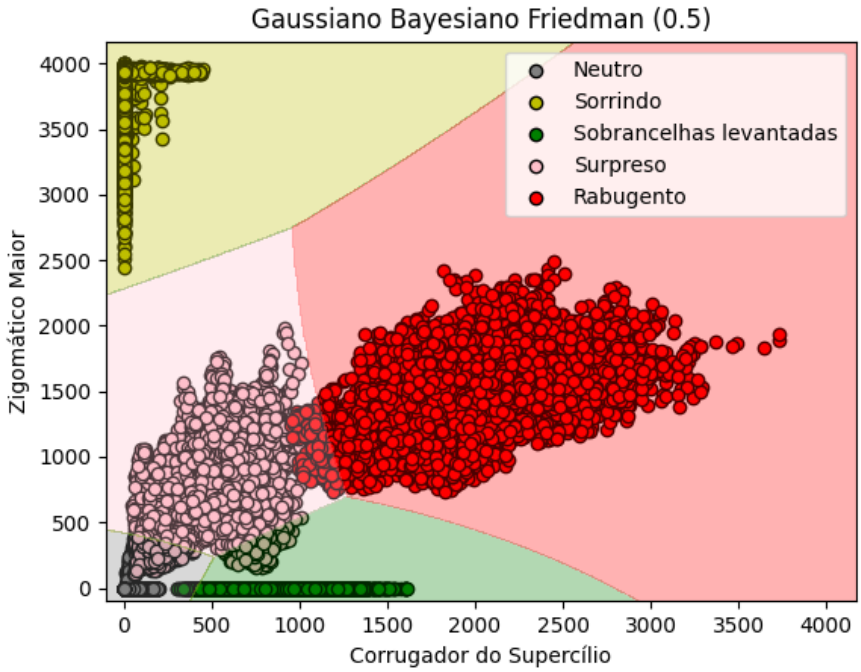

Figure 4: Gráfico de Resultados da Classificação

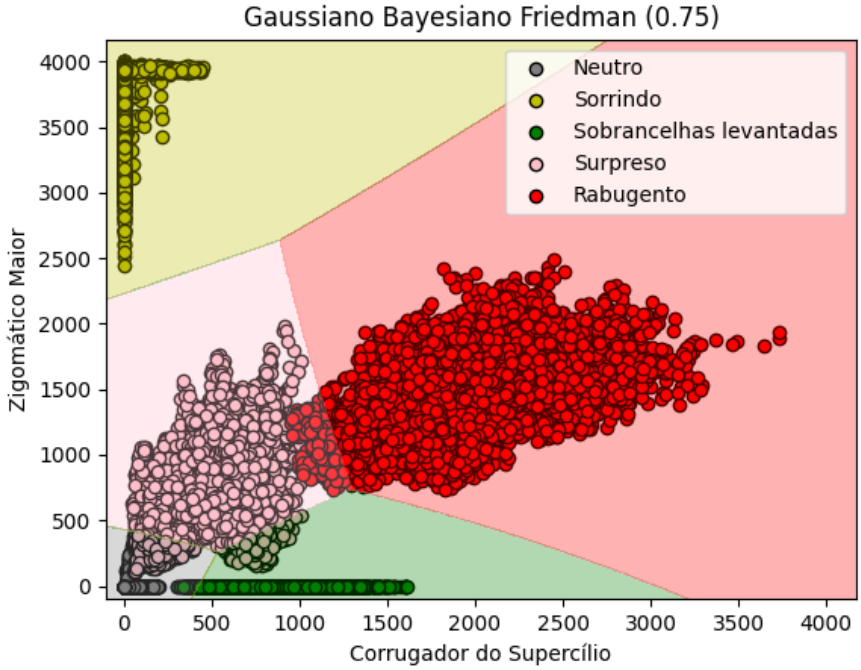

Figure 5: Gráfico de Resultados da Classificação

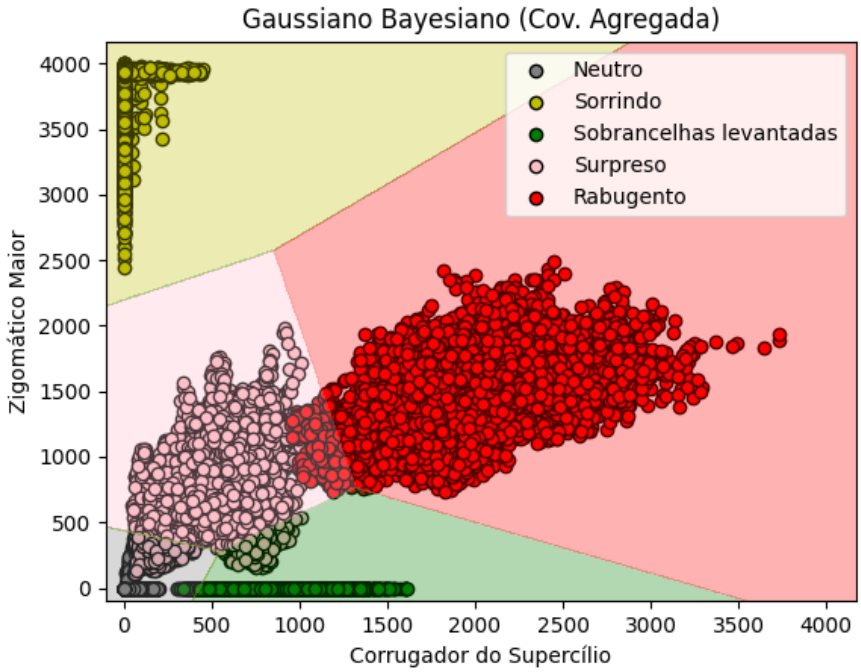

Figure 6: Gráfico de Resultados da Classificação

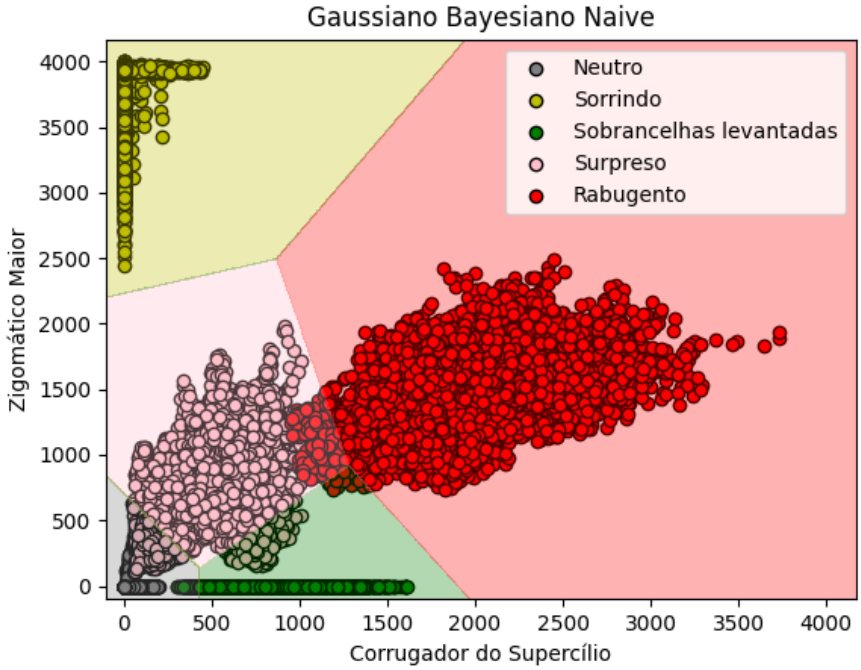

Figure 7: Gráfico de Resultados da Classificação

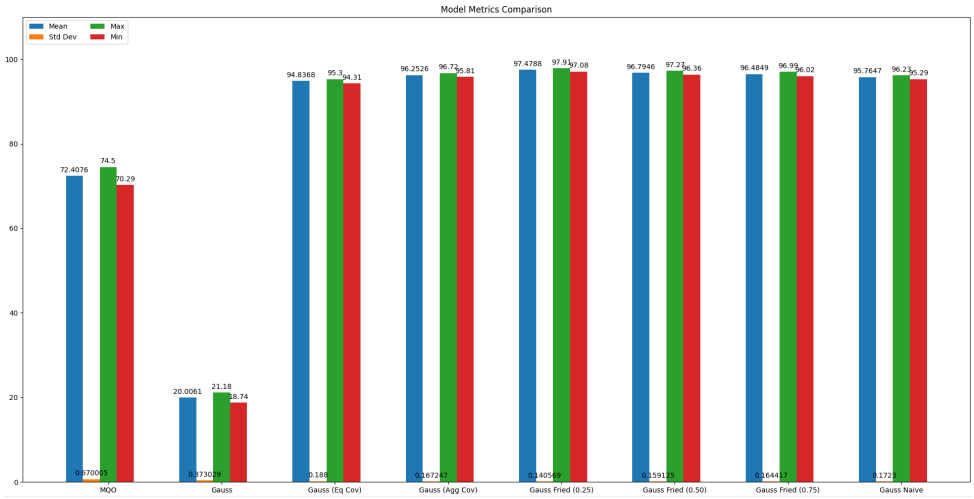

Figure 8: Gráfico de Resultados da Classificação

5 Conclusões

O MQO tradicional apresentou um desempenho sólido, com um erro médio (RSS) de 4.32 e uma pequena variação no desvio padrão (1,21), o que sugere uma boa estabilidade no ajuste dos dados. A fórmula clássica do MQO é amplamente utilizada devido à sua simplicidade e eficácia. No entanto, como esperado, o desempenho do MQO tradicional pode ser superado por modelos que implementam técnicas de regularização, especialmente quando há a necessidade de melhorar a generalização do modelo para novos dados ou cenários mais complexos. Os demais modelos reguladores apenas exibem pequenas variações conforme aumenta o lambda, indicando uma influência moderada da regularização na amplitude dos resultados, que se repetiu nos valores máximo e mínimo também. Os modelos de regressão operam em uma escala menor, mostrando como ainda é necessários alguns ajustes para previsão ser mais assertiva. Os modelos Gaussianos (Tradicional, com Covariância Igual e Agregada) demonstraram desempenho superior, com erros médios que variaram de 19.98 a 96.24. Entre os modelos gaussianos, o modelo Gauss com covariância agregada obteve o melhor desempenho, alcançando um erro médio de 96.24 e apresentando uma baixa variabilidade no desvio padrão. Este modelo se destacou principalmente pela sua capacidade de capturar correlações entre as variáveis de entrada, ajustando melhor a previsão e proporcionando um modelo mais robusto. A regularização da covariância, ao considerar as dependências entre as variáveis, foi um fator essencial para esse desempenho superior. Quando há uma maior correlação entre as variáveis, como no caso dos dados utilizados, essa abordagem melhora significativamente a estabilidade e a precisão das previsões.

Em conclusão, a regularização desempenha um papel crucial na melhoria da performance dos modelos, especialmente em cenários com dados complexos. Modelos como o Gauss Friedman, com regularização controlada por lambda, demonstram como a complexidade do modelo pode ser ajustada para obter uma previsão mais precisa e robusta, o que depende da escolha adequada dos parâmetros de regularização.

6 Referências

Prof.Paulo Cirilo Souza Barbosa - Inteligencia artificial computacional (Slides 5 e 6) ; UNIFOR