A Selective and Regularized Extension of Gaussian Naive Bayes for Conditional Independence Relaxation

Felipe V. Dias[1], **Klenilmar L. Dias**[1], **Klessis L. Dias**[1], **Eonay B. Gurjão**[1]

[1] Laboratório de Inteligência Artificial, Redes Inteligentes e Educação Digital-IFPR
Caixa Postal 87.507-014 – Umuarama – PR – Brazil

Abstract. *Gaussian Naive Bayes is efficient and interpretable, but its conditional independence assumption can be restrictive in tabular data. This paper proposes a selective and regularized relaxation that combines the diagonal covariance structure of GaussianNB with a class-specific sparse covariance matrix, using covariance regularization, class-prior smoothing and adaptive covariance estimation. Experiments on eight public tabular datasets show consistent gains over GaussianNB in accuracy, macro F1-score and AUC. The proposal offers an interpretable probabilistic alternative when conditional independence is overly restrictive.*

Resumo. *O Gaussian Naive Bayes é eficiente e interpretável, mas sua hipótese de independência condicional pode ser restritiva em dados tabulares. Este artigo propõe um relaxamento seletivo e regularizado que combina a estrutura diagonal de covariância do GaussianNB com uma matriz de covariância esparsa específica por classe, usando regularização, suavização de priors e estimação adaptativa da covariância. Experimentos em oito bases tabulares públicas mostram ganhos consistentes sobre o GaussianNB em acurácia, F1-score macro e AUC. A proposta oferece uma alternativa probabilística interpretável quando a independência condicional é excessivamente restritiva.*

1. Introdução

O crescimento do volume e da complexidade dos dados tem intensificado a demanda por modelos de Aprendizado de Máquina (AM) capazes de combinar desempenho preditivo, interpretabilidade e viabilidade computacional. Em aplicações reais de Ciência de Dados e Inteligência Artificial, especialmente em domínios como saúde, educação, políticas públicas e apoio à decisão, os atributos observados frequentemente apresentam correlações, redundâncias ou dependências decorrentes de fenômenos comuns [Liu et al. 2024]. Nesses cenários, modelos probabilísticos simples continuam relevantes, desde que suas hipóteses sejam explicitamente analisadas e, quando necessário, relaxadas [Wickramasinghe and Kalutarage 2021].

O classificador Naive Bayes (NB) é um dos algoritmos probabilísticos mais conhecidos em AM. Sua formulação baseia-se no teorema de Bayes e na hipótese de independência condicional entre atributos dada a classe [Domingos and Pazzani 1997]. Essa hipótese simplifica a estimação das distribuições condicionais e reduz o número de parâmetros do modelo, favorecendo seu uso em cenários com poucos dados, necessidade de treinamento rápido ou exigência de interpretabilidade. No caso do Gaussian Naive Bayes (GNB), os atributos contínuos são modelados por distribuições Gaussianas independentes condicionadas à classe. Em termos matriciais, essa independência equivale a

assumir que a matriz de covariância de cada classe é diagonal. Embora essa simplificação seja eficiente, ela pode descartar informação relevante quando os atributos apresentam dependências estatísticas [Liu et al. 2024]. Em bases de saúde, por exemplo, variáveis fisiológicas, comportamentais, demográficas e clínicas podem apresentar relações importantes para a classificação de risco.

Este artigo propõe um classificador Bayesiano Gaussiano com relaxamento seletivo e regularizado da hipótese de independência condicional. A proposta combina a matriz diagonal do GNB com uma matriz de covariância completa/esparsa por classe, controlada por λ, e incorpora regularização por α, suavização de priors por β, limiar de correlação t e seleção adaptativa do estimador de covariância entre empírico, Ledoit-Wolf e OAS. As principais contribuições são: (i) propor uma generalização seletiva e regularizada do GNB; (ii) introduzir uma interpolação explícita entre independência condicional e dependência linear entre atributos; (iii) incorporar esparsificação, regularização SPD, suavização de priors e seleção adaptativa de covariância; e (iv) avaliar a proposta em oito bases tabulares públicas, comparando-a com modelos probabilísticos, lineares e ensemble.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta a fundamentação e os trabalhos relacionados. A Seção 3 descreve o método proposto. A Seção 4 apresenta a metodologia experimental. A Seção 5 discute os resultados. Por fim, a Seção 6 apresenta as conclusões e trabalhos futuros.

2. Fundamentação e Trabalhos Relacionados

2.1. Gaussian Naive Bayes – GNB

Considere um problema de classificação supervisionada com conjunto de classes $C = \{c_1, c_2, \ldots, c_K\}$ e vetor de atributos $x = (x_1, x_2, \ldots, x_d)$. Pela regra de Bayes, a classe estimada para uma instância x é dada por:

$$\hat{c} = \arg\max_{c_k} P(c_k \mid x).$$

Como $P(x)$ é constante para todas as classes, a regra de decisão pode ser escrita como:

$$\hat{c} = \arg\max_{c_k} P(c_k) P(x \mid c_k).$$

No GNB, assume-se que os atributos são condicionalmente independentes dada a classe:

$$P(x \mid c_k) = \prod_{j=1}^{d} P(x_j \mid c_k).$$

Quando os atributos são contínuos e modelados por distribuições Gaussianas univariadas, essa hipótese equivale a utilizar uma matriz de covariância diagonal para cada classe. Desse modo, as variâncias individuais são consideradas, mas as covariâncias entre atributos são ignoradas.

Essa formulação torna o GNB simples, eficiente e interpretável, pois reduz o número de parâmetros estimados. Entretanto, sua principal limitação está na rigidez da hipótese de independência condicional: embora o NB possa apresentar bom desempenho mesmo quando essa hipótese é violada [Domingos and Pazzani 1997], dependências entre atributos podem carregar informação discriminativa relevante.

2.2. Modelos Gaussianos com Covariância Completa

Uma alternativa ao GNB é modelar $P(x \mid c_k)$ por uma distribuição normal multivariada:

$$P(x \mid c_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right),$$

em que μ_k é o vetor de médias da classe c_k e Σ_k é sua matriz de covariância. Quando Σ_k é completa, o modelo passa a representar dependências lineares entre atributos por meio dos termos fora da diagonal principal.

A Quadratic Discriminant Analysis (QDA) segue essa formulação ao ajustar uma distribuição Gaussiana multivariada por classe, produzindo fronteiras de decisão quadráticas [Friedman 1989, Hastie et al. 2009]. Diferentemente do GNB, que considera apenas variâncias individuais, o QDA utiliza matrizes de covariância completas, tornando-se mais flexível, porém mais sensível à qualidade da estimação.

Em bases pequenas, desbalanceadas, redundantes ou altamente correlacionadas, a matriz empírica pode se tornar singular ou mal condicionada. Esse problema motiva técnicas de regularização de covariância, como Regularized Discriminant Analysis, estimadores de *shrinkage* e OAS [Friedman 1989, Ledoit and Wolf 2004, Schäfer and Strimmer 2005, Chen et al. 2010].

2.3. Relaxamento da Independência Condicional

Estudos recentes têm investigado formas de relaxar a hipótese de independência condicional do NB. Liu et al. propõem transformar os atributos originais em variáveis aproximadamente independentes dadas as classes [Liu et al. 2024]. He et al. e Ou et al. exploram agrupamentos de atributos dependentes para capturar relações internas entre variáveis [He et al. 2025, Ou et al. 2025a]. Outras abordagens consideram ponderação de atributos ou estruturas específicas de covariância em classificadores Gaussianos [Ou et al. 2025b, Chen and Tang 2026].

Essas abordagens reconhecem que nem todos os atributos devem ser tratados como condicionalmente independentes. Entretanto, diferem do presente trabalho porque atuam principalmente por transformação, agrupamento ou ponderação de atributos. A proposta deste artigo atua diretamente na estrutura de covariância do GNB, formulando um relaxamento matricial seletivo e regularizado, com transição controlada entre independência condicional e dependência linear entre atributos.

3. Método Proposto

O método proposto estende o GNB por meio de uma matriz de covariância regularizada por classe. A ideia central é preservar a matriz diagonal do GNB como caso-base e permitir uma transição controlada para uma matriz completa/esparsa, parametrizada por λ. A proposta combina estimação de covariância por classe, esparsificação por correlação, interpolação matricial, regularização SPD, suavização de priors e seleção adaptativa do estimador de covariância.

3.1. Estimação e Esparsificação da Covariância

Para cada classe c_k, seja X_k o subconjunto de instâncias pertencentes à classe. Inicialmente, estima-se uma matriz de covariância $\widehat{\Sigma}_k$ por um estimador $E \in$

$\{\text{empírico}, \text{Ledoit-Wolf}, \text{OAS}\}$. A seleção do estimador é tratada como hiperparâmetro e realizada por validação cruzada interna.

A partir de $\widehat{\Sigma}_k$, calcula-se a matriz diagonal:

$$D_k = \operatorname{diag}(\widehat{\Sigma}_k).$$

Essa matriz representa a estrutura do GNB, pois preserva apenas as variâncias individuais dos atributos. Antes de interpolar as matrizes, aplica-se um limiar de correlação t para remover covariâncias fracas. A correlação entre dois atributos i e j é dada por:

$$\rho_{ij} = \frac{\widehat{\Sigma}_{ij}}{\left(\widehat{\Sigma}_{ii}\widehat{\Sigma}_{jj}\right)^{1/2}}.$$

A covariância $\widehat{\Sigma}_{ij}$ é preservada apenas se $|\rho_{ij}| \geq t$; caso contrário, é zerada. A diagonal é sempre preservada. Com isso, obtém-se uma matriz $\widehat{\Sigma}_k^{(t)}$ que mantém apenas dependências consideradas relevantes.

3.2. Relaxamento e Regularização

O relaxamento entre a matriz diagonal e a matriz completa/esparsa é definido por:

$$\Sigma_k^{(\lambda)} = (1-\lambda)D_k + \lambda\widehat{\Sigma}_k^{(t)},$$

em que $\lambda \in [0,1]$. Quando $\lambda = 0$, o modelo equivale ao GNB. Quando $\lambda = 1$, utiliza-se integralmente a matriz completa/esparsa. Valores intermediários representam graus parciais de relaxamento da independência condicional.

Para aumentar a estabilidade numérica, aplica-se uma regularização em direção à matriz identidade escalada:

$$\Sigma_k^{*} = (1-\alpha)\Sigma_k^{(\lambda)} + \alpha\tau_k I,$$

em que:

$$\tau_k = \frac{\operatorname{tr}(\Sigma_k^{(\lambda)})}{d}.$$

Nessa formulação, d é o número de atributos, I é a matriz identidade e $\alpha \in [0,1]$ controla a intensidade da regularização em direção à matriz identidade escalada $\tau_k I$. Quando $\alpha = 0$, preserva-se integralmente a matriz relaxada $\Sigma_k^{(\lambda)}$; quando $\alpha = 1$, utiliza-se apenas a identidade escalada $\tau_k I$.

Após essa etapa, a matriz é simetrizada e projetada para o cone das matrizes simétricas positivas definidas. Para isso, realiza-se a decomposição espectral da matriz regularizada e aplica-se o limite inferior positivo ϵ aos seus autovalores.

Seja $\Sigma_k^{*} = Q\Lambda Q^T$ a decomposição espectral de Σ_k^{*}, em que $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. A projeção SPD é definida por:

$$\Sigma_k^{SPD} = Q \operatorname{diag}\left(\max(\lambda_1, \epsilon), \ldots, \max(\lambda_d, \epsilon)\right) Q^T.$$

3.3. Suavização dos Priors e Classificação

O prior empírico de uma classe é dado por:

$$P_{emp}(c_k) = \frac{n_k}{n},$$

em que n_k é o número de instâncias da classe c_k e n é o número total de instâncias. O prior uniforme é:

$$P_{uni}(c_k) = \frac{1}{K}.$$

O prior suavizado é definido por:

$$P_\beta(c_k) = (1-\beta)P_{emp}(c_k) + \beta P_{uni}(c_k).$$

Quando $\beta = 0$, utiliza-se o prior empírico; quando $\beta = 1$, utiliza-se o prior uniforme. Valores intermediários permitem controlar o efeito do desbalanceamento de classes.

Após estimar μ_k, $P_\beta(c_k)$ e Σ_k^{SPD}, a classificação de uma nova instância x é realizada por:

$$\hat{c} = \arg\max_{c_k} \left[\log P_\beta(c_k) + \log P(x \mid c_k)\right].$$

A densidade Gaussiana multivariada é calculada na forma logarítmica. A implementação evita a inversão explícita de Σ_k^{SPD}, utilizando fatoração de Cholesky para resolver o termo quadrático de Mahalanobis de forma numericamente mais estável [Golub and Van Loan 2013].

3.4. Fluxo Algorítmico

O Algoritmo 1 resume o treinamento e a predição do método proposto. As operações de esparsificação, relaxamento, regularização, projeção SPD e suavização de priors seguem as formulações apresentadas nas Seções 3.1–3.3.

Algorithm 1 Treinamento e predição do método proposto

Entrada: Conjunto de treino X, y; hiperparâmetros $\lambda, \alpha, \beta, t$; estimador E
Saída: Parâmetros $\{\mu_k, \Sigma_k^{SPD}, L_k, P_\beta(c_k)\}_{k=1}^K$

para cada classe $c_k \in C$ **faça**
  Selecione $X_k = \{x_i \in X : y_i = c_k\}$
  Estime μ_k e $\widehat{\Sigma}_k$ usando E
  Defina $D_k = \mathrm{diag}(\widehat{\Sigma}_k)$
  Obtenha $\widehat{\Sigma}_k^{(t)}$ zerando covariâncias com $|\rho_{ij}| < t$, preservando a diagonal
  Calcule $\Sigma_k^{(\lambda)} = (1-\lambda)D_k + \lambda\widehat{\Sigma}_k^{(t)}$
  Calcule $\tau_k = \mathrm{tr}(\Sigma_k^{(\lambda)})/d$
  Regularize $\Sigma_k^* = (1-\alpha)\Sigma_k^{(\lambda)} + \alpha\tau_k I$
  Projete Σ_k^* para Σ_k^{SPD} e calcule seu fator de Cholesky L_k
  Calcule $P_\beta(c_k) = (1-\beta)P_{\text{emp}}(c_k) + \beta P_{\text{uni}}(c_k)$
fim para
para cada instância de teste x **faça**
  Calcule $s_k = \log P_\beta(c_k) + \log P(x \mid c_k)$, para todo $c_k \in C$
  Prediga $\hat{c} = \arg\max_{c_k \in C} s_k$
fim para

4. Metodologia Experimental

4.1. Etapas de Ciência de Dados

As bases foram carregadas do UCI Machine Learning Repository com o pacote `ucimlrepo`. Linhas com valores ausentes foram removidas; atributos categóricos foram convertidos por *one-hot encoding*; e rótulos textuais foram normalizados antes da codificação numérica, evitando a criação artificial de classes por diferenças de formatação.

Os dados foram divididos de forma estratificada em 70% para treinamento e 30% para teste, em 30 repetições independentes. Modelos probabilísticos e lineares foram avaliados em *pipelines* com `StandardScaler`, com remoção de atributos constantes por `VarianceThreshold`. A seleção de hiperparâmetros foi feita por validação cruzada interna apenas no conjunto de treinamento.

4.2. Bases de Dados

Foram utilizadas oito bases públicas de classificação tabular, contemplando diferentes domínios, tamanhos, números de atributos, cardinalidades de classe e níveis de desbalanceamento. A Tabela 1 apresenta suas principais características. A base Wine Quality foi mantida como cenário multiclasse desbalanceado.

Tabela 1. Bases de dados utilizadas na avaliação experimental.

Base	n	d	K	Min. classe	Domínio
CDC Diabetes	253680	21	2	35346	Saúde
Maternal Health	1014	6	3	272	Saúde
Breast Cancer	569	30	2	212	Saúde
Wine Quality	6497	11	7	5	Qual. físico-química
Bank Marketing	45211	47	2	5289	Marketing
Adult Income	48842	105	2	11687	Socioeconômico
Ionosphere	351	34	2	126	Sinais
Spambase	4601	57	2	1813	Spam

4.3. Modelos Comparados

O método proposto foi comparado com quatro modelos. O **GaussianNB Tuned** foi utilizado como baseline principal, pois representa a formulação que a proposta busca generalizar. O **QDA Tuned** foi incluído como competidor conceitual próximo, pois também modela distribuições Gaussianas por classe com matrizes de covariância específicas. Diferentemente do QDA, porém, o método proposto preserva a matriz diagonal do GaussianNB como referência e controla a transição para uma matriz completa ou esparsa por meio de λ. A **LogisticRegression Tuned** foi usada como baseline discriminativo linear, e o **HistGradientBoosting Tuned** como referência ensemble não linear [Friedman 2001].

4.4. Seleção de Hiperparâmetros e Métricas

A seleção de hiperparâmetros foi realizada por validação cruzada estratificada interna com três partições, embaralhamento dos dados e semente fixa igual a 42. Esse procedimento foi aplicado exclusivamente ao conjunto de treinamento de cada repetição, evitando o uso do conjunto de teste durante o ajuste dos modelos. Para GaussianNB, QDA, Regressão Logística e HistGradientBoosting, foi utilizada busca em grade. Para o método proposto,

foi utilizada busca aleatória (`RandomizedSearchCV`) com 40 iterações. Em todos os casos, a função objetivo da seleção de hiperparâmetros foi o F1-score macro médio obtido na validação cruzada interna:

$$\theta^* = \arg\max_{\theta \in \Theta} \frac{1}{K} \sum_{i=1}^{K} F1^{(i)}_{macro}(\theta).$$

A escolha do F1-score macro se justifica porque as bases podem apresentar desbalanceamento entre classes. Diferentemente da acurácia simples, o F1-score macro atribui igual importância média às classes [Sokolova and Lapalme 2009]. Os modelos foram avaliados por acurácia, acurácia balanceada, F1-score macro, AUC e tempo de treinamento. Os resultados são reportados no formato média $\pm$ desvio-padrão, considerando 30 repetições independentes.

Para comparar o método proposto ao GaussianNB Tuned, foi aplicado o teste de Wilcoxon pareado com correção de Holm para múltiplas comparações [Wilcoxon 1945, Holm 1979]. Além da análise por base, foram calculados o ranking médio dos modelos e a contagem de vitórias, empates e derrotas do método proposto em relação aos baselines, considerando as oito bases avaliadas e as quatro métricas principais: acurácia, acurácia balanceada, F1-score macro e AUC.

5. Resultados e Discussão

Esta seção apresenta os resultados obtidos nas oito bases de dados utilizadas na avaliação experimental. As tabelas foram produzidas a partir dos arquivos gerados pelo experimento computacional. Os resultados completos incluem acurácia, acurácia balanceada, F1-score macro, AUC e tempo de treinamento. A análise enfatiza a comparação entre o método proposto e o GaussianNB Tuned, pois a proposta foi concebida como uma generalização seletiva e regularizada desse classificador.

5.1. Resultados Gerais

As Tabelas 2 e 3 apresentam os resultados médios dos modelos avaliados nas oito bases. Os valores representam média $\pm$ desvio-padrão em 30 repetições independentes.

Os resultados mostram que o método proposto superou o GaussianNB Tuned em acurácia nas oito bases avaliadas, com ganhos mais expressivos em Adult Census Income, Ionosphere e Maternal Health Risk. Esse comportamento indica que a incorporação seletiva de dependências entre atributos pode beneficiar o classificador quando a hipótese de independência condicional é restritiva. Entretanto, o método proposto não obteve o melhor desempenho absoluto na maioria das bases. O HistGradientBoosting Tuned apresentou a maior acurácia em seis das oito bases, enquanto LogisticRegression Tuned foi superior em Breast Cancer Wisconsin e QDA Tuned em Ionosphere. Assim, os resultados posicionam a proposta como uma generalização probabilística interpretável do GaussianNB, e não como substituta universal de modelos mais flexíveis.

Tabela 2. Resultados médios dos modelos nas bases Adult Census Income, Bank Marketing, Breast Cancer Wisconsin e CDC Diabetes.

Base	Modelo	Acc.	Bal. Acc.	F1	AUC
Adult	GaussianNB	0.5851 ± 0.0299	0.7050 ± 0.0173	0.5769 ± 0.0266	0.8063 ± 0.0308
Adult	QDA	0.7805 ± 0.0048	0.7956 ± 0.0031	0.7421 ± 0.0041	0.8700 ± 0.0024
Adult	LogReg	0.8525 ± 0.0025	0.7666 ± 0.0038	0.7835 ± 0.0037	0.9066 ± 0.0022
Adult	HGB	**0.8741 ± 0.0023**	**0.7983 ± 0.0044**	**0.8160 ± 0.0037**	**0.9287 ± 0.0018**
Adult	Proposto	0.6882 ± 0.0184	0.7598 ± 0.0100	0.6658 ± 0.0157	0.8474 ± 0.0066
Bank	GaussianNB	0.8621 ± 0.0032	0.7023 ± 0.0051	0.6885 ± 0.0049	0.8109 ± 0.0056
Bank	QDA	0.8813 ± 0.0024	0.7070 ± 0.0058	0.7094 ± 0.0053	0.8703 ± 0.0039
Bank	LogReg	0.8451 ± 0.0030	**0.8320 ± 0.0042**	0.7291 ± 0.0037	0.9092 ± 0.0025
Bank	HGB	**0.9084 ± 0.0017**	0.7251 ± 0.0067	**0.7514 ± 0.0056**	**0.9349 ± 0.0020**
Bank	Proposto	0.8668 ± 0.0030	0.7163 ± 0.0057	0.7005 ± 0.0052	0.8365 ± 0.0058
Breast	GaussianNB	0.9353 ± 0.0208	0.9280 ± 0.0226	0.9305 ± 0.0222	0.9868 ± 0.0077
Breast	QDA	0.9643 ± 0.0126	0.9563 ± 0.0167	0.9614 ± 0.0138	0.9933 ± 0.0069
Breast	LogReg	**0.9741 ± 0.0124**	**0.9707 ± 0.0150**	**0.9722 ± 0.0135**	**0.9949 ± 0.0045**
Breast	HGB	0.9612 ± 0.0142	0.9563 ± 0.0154	0.9584 ± 0.0151	0.9910 ± 0.0062
Breast	Proposto	0.9437 ± 0.0178	0.9292 ± 0.0222	0.9383 ± 0.0197	0.9913 ± 0.0061
CDC	GaussianNB	0.7731 ± 0.0014	0.6879 ± 0.0017	0.6356 ± 0.0015	0.7836 ± 0.0019
CDC	QDA	0.8144 ± 0.0012	0.6693 ± 0.0017	**0.6512 ± 0.0015**	0.8041 ± 0.0017
CDC	LogReg	0.7313 ± 0.0014	**0.7455 ± 0.0020**	0.6327 ± 0.0014	0.8223 ± 0.0017
CDC	HGB	**0.8669 ± 0.0006**	0.5705 ± 0.0017	0.5887 ± 0.0024	**0.8297 ± 0.0017**
CDC	Proposto	0.7854 ± 0.0012	0.6805 ± 0.0018	0.6394 ± 0.0015	0.7862 ± 0.0019

Tabela 3. Resultados médios dos modelos nas bases Ionosphere, Maternal Health Risk, Spambase e Wine Quality.

Base	Modelo	Acc.	Bal. Acc.	F1	AUC
Ionosphere	GaussianNB	0.8943 ± 0.0291	0.8662 ± 0.0370	0.8795 ± 0.0344	0.9305 ± 0.0239
Ionosphere	QDA	**0.9377 ± 0.0213**	**0.9203 ± 0.0272**	**0.9303 ± 0.0241**	0.9785 ± 0.0109
Ionosphere	LogReg	0.8695 ± 0.0259	0.8348 ± 0.0369	0.8492 ± 0.0327	0.9004 ± 0.0282
Ionosphere	HGB	0.9157 ± 0.0245	0.8935 ± 0.0344	0.9047 ± 0.0293	0.9559 ± 0.0219
Ionosphere	Proposto	0.9267 ± 0.0236	0.9017 ± 0.0318	0.9165 ± 0.0280	**0.9787 ± 0.0115**
Maternal	GaussianNB	0.6017 ± 0.0189	0.5836 ± 0.0209	0.5641 ± 0.0211	0.7797 ± 0.0153
Maternal	QDA	0.6522 ± 0.0223	0.6391 ± 0.0234	0.6339 ± 0.0237	0.8078 ± 0.0159
Maternal	LogReg	0.6151 ± 0.0283	0.6142 ± 0.0279	0.6111 ± 0.0288	0.7941 ± 0.0139
Maternal	HGB	**0.7929 ± 0.0240**	**0.7962 ± 0.0242**	**0.7965 ± 0.0240**	**0.9210 ± 0.0152**
Maternal	Proposto	0.6475 ± 0.0220	0.6355 ± 0.0235	0.6289 ± 0.0247	0.8070 ± 0.0175
Spambase	GaussianNB	0.8164 ± 0.0095	0.8411 ± 0.0082	0.8157 ± 0.0094	0.9275 ± 0.0183
Spambase	QDA	0.9033 ± 0.0062	0.8970 ± 0.0071	0.8983 ± 0.0066	0.9466 ± 0.0045
Spambase	LogReg	0.9253 ± 0.0072	0.9226 ± 0.0081	0.9219 ± 0.0076	0.9709 ± 0.0032
Spambase	HGB	**0.9527 ± 0.0048**	**0.9499 ± 0.0061**	**0.9504 ± 0.0051**	**0.9875 ± 0.0028**
Spambase	Proposto	0.8262 ± 0.0145	0.8487 ± 0.0118	0.8252 ± 0.0141	0.9430 ± 0.0063
Wine	GaussianNB	0.4458 ± 0.0128	0.2581 ± 0.0193	0.2423 ± 0.0144	0.7046 ± 0.0443
Wine	QDA	0.4896 ± 0.0131	0.2648 ± 0.0137	0.2566 ± 0.0129	0.6935 ± 0.0248
Wine	LogReg	0.5275 ± 0.0621	0.2327 ± 0.0438	0.2204 ± 0.0113	0.7766 ± 0.0138
Wine	HGB	**0.6299 ± 0.0131**	**0.3309 ± 0.0133**	**0.3590 ± 0.0160**	**0.8283 ± 0.0168**
Wine	Proposto	0.4462 ± 0.0253	0.3141 ± 0.0801	0.2451 ± 0.0166	0.7412 ± 0.0408

5.2. Comparação com GaussianNB

A comparação direta com o GaussianNB Tuned é central, pois o método proposto foi concebido como uma generalização seletiva e regularizada do GNB. A Tabela 4 apresenta as diferenças médias entre o método proposto e o GaussianNB Tuned. Valores positivos

indicam vantagem do método proposto.

Tabela 4. Diferença média entre o método proposto e o GaussianNB Tuned.

Base	Acurácia	Acurácia Bal.	F1-macro	AUC
Adult Census Income	+0.1031	+0.0547	+0.0889	+0.0411
Bank Marketing	+0.0047	+0.0139	+0.0120	+0.0256
Breast Cancer Wisconsin	+0.0084	+0.0013	+0.0078	+0.0045
CDC Diabetes	+0.0122	-0.0075	+0.0038	+0.0026
Ionosphere	+0.0324	+0.0355	+0.0369	+0.0482
Maternal Health Risk	+0.0458	+0.0520	+0.0648	+0.0272
Spambase	+0.0098	+0.0076	+0.0095	+0.0155
Wine Quality	+0.0004	+0.0561	+0.0027	+0.0367

A comparação direta mostra que o método proposto obteve ganhos sobre o GaussianNB Tuned em 31 das 32 combinações entre bases e métricas. A única exceção ocorreu na acurácia balanceada da CDC Diabetes, na qual o GaussianNB Tuned apresentou vantagem de 0.0075. Mesmo nessa base, o método proposto superou o GaussianNB em acurácia, F1-score macro e AUC. Os maiores ganhos ocorreram em Adult Census Income, Maternal Health Risk e Ionosphere. Em Wine Quality, apesar dos ganhos pequenos em acurácia e F1-score macro, houve melhora em acurácia balanceada e AUC, aspecto relevante em cenário multiclasse desbalanceado. Esses resultados reforçam que o relaxamento seletivo da covariância pode incorporar dependências úteis sem abandonar a estrutura probabilística do GNB.

5.3. Hiperparâmetros e Estimadores de Covariância

A Tabela 5 resume os hiperparâmetros médios e a frequência de seleção dos estimadores de covariância. O parâmetro λ controla o grau de relaxamento da independência condicional; α controla a regularização em direção à identidade escalada; β controla a suavização do prior; e t define o limiar de correlação para manter covariâncias relevantes.

Tabela 5. Hiperparâmetros médios e estimadores de covariância selecionados pelo método proposto.

Base	λ	α	β	t	Empírico	Ledoit-Wolf	OAS
Adult Census Income	0.7800	0.0267	0.0667	0.2000	0/30	4/30	26/30
Bank Marketing	0.7000	0.0001	0.5000	0.0000	0/30	0/30	30/30
Breast Cancer Wisconsin	0.2967	0.0937	0.6333	0.1500	12/30	5/30	13/30
CDC Diabetes	0.2000	0.2000	0.0000	0.2000	30/30	0/30	0/30
Ionosphere	0.8800	0.1777	0.3167	0.1117	26/30	1/30	3/30
Maternal Health Risk	0.7200	0.1617	0.6250	0.1067	6/30	8/30	16/30
Spambase	0.8333	0.0133	0.7833	0.1800	28/30	2/30	0/30
Wine Quality	0.7433	0.0859	0.2417	0.0833	8/30	15/30	7/30

Na CDC Diabetes, o valor médio de $\lambda = 0.2000$ indica baixo relaxamento da independência condicional, mantendo a matriz de covariância próxima da diagonal. O valor $\beta = 0.0000$ indica preservação do prior empírico, enquanto o limiar $t = 0.2000$ mostra que o método descartou covariâncias fracas. A covariância empírica foi selecionada em todas as repetições, o que é coerente com o grande número de instâncias da base. Na

Maternal Health Risk, o valor médio de $\lambda = 0.7200$ indica maior incorporação das covariâncias entre atributos. O valor médio de $\beta = 0.6250$ sugere que a suavização do prior foi importante, aproximando parcialmente os priors de uma distribuição uniforme. Além disso, estimadores regularizados foram selecionados com maior frequência: OAS em 16 das 30 repetições e Ledoit-Wolf em 8 das 30. Esse resultado sugere que a regularização da covariância foi relevante para melhorar a estabilidade do modelo nessa base. Nas bases Adult Census Income, Bank Marketing, Spambase, Wine Quality e Ionosphere, os valores médios de λ foram iguais ou superiores a 0.7000, indicando maior relaxamento da independência condicional. Esse comportamento sugere que, nesses cenários, as covariâncias entre atributos carregam informação discriminativa útil. Em Adult Census Income e Bank Marketing, estimadores regularizados foram predominantes, com destaque para OAS. Em Ionosphere e Spambase, a covariância empírica foi selecionada na maior parte das repetições, sugerindo que, nessas bases, a estrutura empírica das covariâncias foi suficientemente estável para o método proposto.

Esses resultados reforçam o caráter adaptativo da proposta. O método não impõe um grau fixo de relaxamento para todas as bases; ao contrário, ajusta λ, α, β, t e o estimador de covariância conforme a estrutura estatística observada em cada conjunto de dados.

5.4. Análise Estatística

Para verificar se as diferenças entre o método proposto e o GaussianNB Tuned foram consistentes ao longo das 30 repetições, foi aplicado o teste de Wilcoxon pareado com correção de Holm para múltiplas comparações. A Tabela 6 apresenta, para cada base e métrica, a diferença média entre o método proposto e o GaussianNB Tuned, acompanhada do p-valor ajustado pela correção de Holm. Valores positivos indicam vantagem do método proposto.

Tabela 6. Teste de Wilcoxon pareado entre o método proposto e o GaussianNB Tuned. Em cada célula, apresenta-se diferença média / p-valor de Holm. Valores positivos favorecem o método proposto.

Base	Acc.	Bal. Acc.	F1	AUC
Adult	+0.1031 / 3.11×10^{-5}	+0.0547 / 5.96×10^{-8}	+0.0889 / 5.96×10^{-8}	+0.0411 / 3.54×10^{-7}
Bank	+0.0047 / 3.11×10^{-5}	+0.0139 / 5.96×10^{-8}	+0.0120 / 5.96×10^{-8}	+0.0256 / 5.96×10^{-8}
Breast	+0.0084 / 1.03×10^{-2}	+0.0013 / 1.00	+0.0078 / 1.77×10^{-2}	+0.0045 / 2.42×10^{-4}
CDC	+0.0122 / 3.11×10^{-5}	-0.0075 / 5.96×10^{-8}	+0.0038 / 5.96×10^{-8}	+0.0026 / 5.96×10^{-8}
Ionosphere	+0.0324 / 6.26×10^{-5}	+0.0355 / 6.27×10^{-5}	+0.0369 / 6.26×10^{-5}	+0.0482 / 3.11×10^{-5}
Maternal	+0.0458 / 3.11×10^{-5}	+0.0520 / 5.96×10^{-8}	+0.0648 / 5.96×10^{-8}	+0.0272 / 5.96×10^{-8}
Spambase	+0.0098 / 2.31×10^{-4}	+0.0076 / 2.31×10^{-4}	+0.0095 / 3.11×10^{-5}	+0.0155 / 5.96×10^{-8}
Wine	+0.0004 / 1.00	+0.0561 / 1.03×10^{-2}	+0.0027 / 9.85×10^{-1}	+0.0367 / 5.96×10^{-8}

Os resultados indicam diferenças estatisticamente significativas na maioria das combinações entre bases e métricas. Em Adult Census Income, Bank Marketing, CDC Diabetes, Ionosphere, Maternal Health Risk e Spambase, as diferenças foram significativas em todas as métricas. Na Breast Cancer Wisconsin, apenas a acurácia balanceada não foi significativa; na Wine Quality, acurácia e F1-score macro não foram significativas, embora acurácia balanceada e AUC tenham apresentado ganhos significativos. A única diferença média negativa ocorreu na acurácia balanceada da CDC Diabetes, indicando

vantagem do GaussianNB Tuned nessa métrica específica. Como análise complementar, o ranking médio geral posicionou o HistGradientBoosting Tuned em primeiro lugar, seguido por QDA Tuned, LogisticRegression Tuned, método proposto e GaussianNB Tuned. Assim, o método proposto ocupa uma posição intermediária: supera de forma consistente o GaussianNB Tuned, mas não supera, em desempenho absoluto, modelos de maior flexibilidade.

5.5. Limitações e Aplicabilidade

Os resultados indicam que o método proposto é uma alternativa probabilística regularizada ao GNB, mas não um substituto universal de modelos de maior capacidade. HistGradientBoosting Tuned obteve melhor desempenho absoluto na maior parte das bases, e QDA Tuned foi superior em algumas métricas. A contribuição da proposta está em preservar a matriz diagonal do GNB como referência e controlar, por meio de λ, a transição entre independência condicional e dependência linear entre atributos. O método apresenta custo computacional superior ao GaussianNB, devido à busca de hiperparâmetros, à estimação regularizada das covariâncias e à fatoração de Cholesky. Ainda assim, mantém formulação probabilística interpretável. Seus hiperparâmetros permitem analisar o grau de relaxamento, a intensidade de regularização, a suavização dos priors e a filtragem de dependências fracas, sendo aplicável a problemas tabulares em que simplicidade e interpretação probabilística são relevantes.

6. Conclusões e Trabalhos Futuros

Este artigo apresentou um classificador Bayesiano Gaussiano com relaxamento seletivo e regularizado da hipótese de independência condicional do GNB. A proposta combina a matriz diagonal do modelo ingênuo com uma matriz de covariância completa/esparsa por classe, controlada por λ, além de incorporar regularização, suavização de priors e seleção de covariâncias relevantes por limiar de correlação. A avaliação em oito bases tabulares públicas mostrou que o método proposto superou o GaussianNB Tuned em acurácia, F1-score macro e AUC em todas as bases, e em acurácia balanceada em sete delas. Esses resultados indicam que a independência condicional estrita pode ser restritiva em diferentes problemas tabulares e que seu relaxamento seletivo pode melhorar o desempenho do GaussianNB. O método não obteve o melhor desempenho absoluto frente a modelos mais flexíveis, como HistGradientBoosting, QDA e LogisticRegression. Assim, sua contribuição está em oferecer uma alternativa probabilística interpretável para cenários em que o GaussianNB é atrativo por sua simplicidade, mas a independência condicional precisa ser relaxada.

Como trabalhos futuros, pretende-se comparar a proposta com classificadores Bayesianos estruturais, como Tree-Augmented Naive Bayes e redes Bayesianas ajustadas para classificação, investigar testes de independência para análise das dependências preservadas, explorar visualmente as matrizes de covariância aprendidas e estudar estratégias mais eficientes de seleção dos hiperparâmetros.

Declaração sobre Uso de IA Generativa

O ChatGPT/OpenAI foi utilizado apenas como apoio editorial.

Referências

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. (2010). Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029.

Chen, Y.-J. and Tang, M. (2026). Classification of high-dimensional data with spiked covariance matrix structure. *Transactions on Machine Learning Research*.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. Johns Hopkins University Press, 4 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

He, Y., Ou, G., Fournier-Viger, P., and Huang, J. Z. (2025). Attribute grouping-based naive bayesian classifier. *Science China Information Sciences*, 68(3):132106.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Liu, H., Wang, S., and Jiang, L. (2024). Alleviating conditional independence assumption of naive bayes. *Statistical Papers*, 65:2767–2797.

Ou, G., He, Y., Cheng, Y., and Huang, J. Z. (2025a). Relaxed naïve bayesian classifier based on maximum dependent attribute groups. *Information Sciences*, 689:122013.

Ou, G., He, Y., Fournier-Viger, P., and Huang, J. Z. (2025b). A novel multi-source weighted naive bayes classifier. *Information Sciences*, 721:122568.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Wickramasinghe, I. and Kalutarage, H. (2021). Naive bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing*, 25:2277–2293.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.