Trabalho Prático Hipertensão PNS 2019

Ettore Motta Gazzinelli
Aprendizado de Máquina, PUC Minas Lordes
Minas Gerias, Brasil
ettoremgazzinelli@gmail.com

Resumo

Este estudo propõe e avalia modelos classificatórios para a predição de hipertensão arterial na população brasileira, utilizando abordagens de aprendizado de máquina. A fundação do trabalho são os microdados da Pesquisa Nacional de Saúde (PNS) 2019. O modelo visa identificar os principais fatores preditivos — abrangendo aspectos demográficos, socioeconômicos e de estilo de vida — que discriminam indivíduos com diagnóstico prévio de hipertensão. Uma ênfase metodológica é colocada na comparação de estratégias para tratar o desequilíbrio de classes; para isso, algoritmos (Random Forest, Decision Tree, KNN) são combinados com técnicas de reamostragem (como SMOTE, ADASYN, RUS e SMOTEENN) e seus desempenhos são rigorosamente comparados.

Keywords

Mineração de Dados, Machine Learning, Hipertensão, PNS, Saúde Pública

ACM Reference Format:
Ettore Motta Gazzinelli. 2025. Trabalho Prático Hipertensão PNS 2019. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Descrição da Base de Dados

A fonte de dados primária para este trabalho é a Pesquisa Nacional de Saúde (PNS) 2019. Embora o levantamento original completo contenha 279.382 registros, uma extração processada foi utilizada como ponto de partida.

Um pipeline de preparação de dados, que incluiu engenharia de features (como a discretização de variáveis) e a imputação de valores ausentes através do algoritmo MissForest, foi aplicado. Este processo resultou na base de dados final de modelagem, composta por **88.736 observações**.

Para a tarefa de classificação binária, o modelo utiliza **21 features preditoras**, que englobam desde hábitos de consumo (alimentar, álcool, tabaco) e comorbidades até a autoavaliação da saúde. A variável dependente (alvo) é a `diag_hipertensao`, que representa o diagnóstico médico prévio da condição.

Tabela 1: Dicionário de Atributos Utilizados na Modelagem.

Atributo (Feature)	Descrição
`diag_hipertensao`	**Variável Alvo** (Diagnóstico de hipertensão)
`idade`	Idade do morador (Numérico)
`sexo`	Sexo do morador (Binário)
`autoavaliacao_saude`	Autoavaliação do estado de saúde (Ordinal)
`consumo_carne_vermelha`	Frequência de consumo de carne vermelha (O
`consumo_frango`	Frequência de consumo de frango (Ordinal)
`consumo_peixe`	Frequência de consumo de peixe (Ordinal)
`consumo_leite`	Frequência de consumo de leite (Ordinal)
`consumo_verdura_legume`	Frequência de consumo de verduras/legumes (
`consumo_frutas`	Frequência de consumo de frutas (Ordinal)
`consumo_doces`	Frequência de consumo de doces (Ordinal)
`consumo_substitui_almoco`	Frequência de substituição do almoço (Ordina
`bebidas_acucar_dias_semana`	Consumo de bebidas açucaradas (Ordinal)
`freq_bebida_alcoolica`	Frequência de consumo de álcool (Ordinal)
`fumante_atual`	Status de tabagismo (Binário)
`percepcao_consumo_sal`	Percepção do consumo de sal (Ordinal)
`diag_doenca_coracao`	Diagnóstico prévio de doença cardíaca (Binári
`diag_colesterol_alto`	Diagnóstico prévio de colesterol alto (Binário)
`freq_problemas_sono`	Frequência de problemas de sono (Ordinal)
`freq_cansaco`	Frequência de cansaço/fadiga (Ordinal)
`atividade_fisica_categoria`	Nível de atividade física (Ordinal)
`IMC_Categoria`	Faixa de Índice de Massa Corporal (Ordinal)

2 Metodologia de Pré-processamento e Modelagem

A preparação dos dados brutos para a etapa de modelagem seguiu um pipeline metodológico focado em robustez, tratamento de dados complexos e preparação para um ambiente de classificação desbalanceado.

2.1 Tratamento de Dados Ausentes e Outliers

Diferente de outras abordagens que filtram subamostras (como por faixa etária ou medições antropométricas), este estudo utilizou o conjunto de dados completo para maximizar a generalização.

O tratamento de valores ausentes (missing values) foi uma etapa crítica. Inicialmente, foi aplicada uma técnica de produto cartesiano para garantir a integridade estrutural dos registros. Subsequentemente, os dados faltantes restantes foram imputados utilizando **MissForest**. Este é um algoritmo de imputação não paramétrico, baseado em Random Forest, que demonstra alta performance ao lidar com tipos de dados mistos (categóricos e numéricos) sem assumir uma distribuição específica.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Adicionalmente, o impacto de valores extremos (outliers) não foi tratado por remoção (capping ou exclusão), mas sim mitigado através da **discretização** das variáveis contínuas. Ao agrupar valores em faixas (bins), a influência de pontos extremos é naturalmente contida.

2.2 Engenharia e Codificação de Features

Para melhorar a interpretabilidade dos dados e dos resultados do modelo, os atributos com códigos mnemônicos (ex: `C006`) foram renomeados para descrições legíveis (ex: `sexo`).

A estratégia de codificação foi diretamente influenciada pela discretização. Como as variáveis contínuas foram transformadas em faixas ordinais (ex: "Baixo", "Médio", "Alto"), a necessidade de *One-Hot Encoding* foi eliminada para a maioria dos algoritmos baseados em árvore (Decision Tree, Random Forest). A codificação ordinal (Label Encoding) foi, portanto, aplicada implicitamente durante a etapa de discretização.

Para modelos sensíveis à escala, como o KNN, as features numéricas (discretas ou contínuas) foram padronizadas (média 0, desvio 1) usando `StandardScaler`.

2.3 Divisão e Reamostragem do Conjunto de Treino

O conjunto de dados final foi dividido em **80% para treino** e **20% para teste**. Foi utilizado o parâmetro `stratify=y` (onde `y` é a variável alvo `diag_hipertensao`) para assegurar que a proporção original das classes fosse mantida em ambas as divisões.

Uma análise do conjunto de treino revelou um desequilíbrio de classes significativo. Para mitigar o viés do modelo em direção à classe majoritária, quatro estratégias distintas de reamostragem da biblioteca `imblearn` foram aplicadas *apenas* ao conjunto de treino e avaliadas independentemente:

- **Técnicas de Oversampling:** SMOTE (Synthetic Minority Over-sampling Technique) e ADASYN (Adaptive Synthetic Sampling).
- **Técnicas de Undersampling:** RUS (Random Undersampling).
- **Técnicas Híbridas:** SMOTEENN (uma combinação de SMOTE e Edited Nearest Neighbours).

3 Modelagem e Otimização de Hiperparâmetros

Para a tarefa de classificação, foram escolhidos três algoritmos canônicos do `scikit-learn`, representando diferentes abordagens de aprendizado: **K-Nearest Neighbors (KNN)**, um modelo baseado em instância; **Decision Tree (DT)**, um modelo simbólico; e **Random Forest (RF)**, um método de ensemble.

Cada algoritmo foi treinado em pipelines separados, combinados com cada uma das quatro técnicas de reamostragem (SMOTE, ADASYN, RUS, SMOTEENN) descritas anteriormente.

3.1 Estratégia de Otimização Bayesiana

Para encontrar a configuração ótima de hiperparâmetros para cada pipeline (ex: `RF_SMOTE`), foi empregada a Otimização Bayesiana através do `BayesSearchCV`.

Diferente de uma busca exaustiva (Grid Search), essa abordagem trata a otimização como um problema probabilístico para encontrar de forma mais eficiente a melhor combinação de parâmetros. A função-objetivo para a otimização foi maximizar o **F1-Score (Macro Average)** durante a validação cruzada (CV) no conjunto de treino. O F1-Macro foi escolhido por ser uma métrica robusta que avalia o equilíbrio entre precisão e recall para *ambas* as classes, sendo ideal para o nosso cenário de dados desbalanceados.

3.2 Resultados no Conjunto de Teste

Após a conclusão da otimização, o melhor estimador de cada pipeline foi submetido à avaliação final contra o conjunto de teste (20% dos dados), que o modelo nunca havia visto. Os resultados consolidados são apresentados na Tabela 2.

Tabela 2: Resultados Consolidados dos Modelos no Conjunto de Teste (N=17.748). A métrica-alvo da otimização foi o F1-Macro.

Modelo	Balanceador	F1-Macro	Recall (Cls. 1)	Precisão (Cls.
RandomForest	SMOTE	**0.72**	0.66	**0.55**
RandomForest	ADASYN	0.71	0.66	0.53
RandomForest	RUS	0.71	0.79	0.51
RandomForest	SMOTEENN	0.67	0.83	0.46
DecisionTree	SMOTE	0.70	0.69	0.51
DecisionTree	ADASYN	0.69	0.71	0.50
DecisionTree	RUS	0.70	0.77	0.50
DecisionTree	SMOTEENN	0.66	**0.85**	0.45
KNN	SMOTE	0.71	0.66	0.53
KNN	RUS	0.70	0.72	0.51
KNN	SMOTEENN	0.67	0.81	0.46

3.3 Hiperparâmetros Selecionados

Os hiperparâmetros ótimos para cada pipeline foram identificados via `BayesSearchCV`. As tabelas 3, 4 e 5 sumarizam as configurações finais para cada família de algoritmo, combinadas com as diferentes estratégias de reamostragem.

4 Avaliação e Discussão dos Resultados

Os estimadores ótimos, derivados do processo de `BayesSearchCV`, foram então aplicados ao conjunto de teste (20% dos dados), que foi mantido isolado durante todo o treinamento. A Tabela 2 (apresentada anteriormente) resume o desempenho de cada pipeline.

4.1 Métricas Focais para Saúde

Em um problema de diagnóstico de saúde, como a hipertensão, a acurácia global pode ser enganosa. O foco da avaliação deve ser em métricas específicas para a classe positiva (Classe 1: Hipertensão):

- **Precisão (Precision):** Responde à pergunta: "Dos indivíduos que o modelo classificou como hipertensos, quantos realmente possuem a condição?"Uma alta precisão é crucial para minimizar falsos positivos e evitar encaminhamentos desnecessários.

Tabela 3: Hiperparâmetros Otimizados (BayesSearchCV) para KNN.

Balanceador	N neighbors	Metric	Weights
SMOTE	31	manhattan	uniform
ADASYN	31	manhattan	distance
RUS	31	manhattan	distance
SMOTEENN	27	manhattan	distance

Tabela 4: Hiperparâmetros Otimizados (BayesSearchCV) para Árvore de Decisão.

Balanceador	Criterion	Max depth	Min samples leaf	Min samples split
SMOTE	entropy	6	1	19
ADASYN	entropy	8	20	19
RUS	gini	6	1	8
SMOTEENN	entropy	5	7	2

Tabela 5: Hiperparâmetros Otimizados (BayesSearchCV) para Random Forest.

Balanceador	N estimators	Criterion	Max depth	Min samples leaf	Min samples split
SMOTE	300	entropy	28	1	20
ADASYN	170	gini	23	1	20
RUS	181	entropy	13	12	16
SMOTEENN	300	entropy	50	1	2

- **Recall (Sensibilidade):** Responde à pergunta: "De todos os indivíduos que são de fato hipertensos, quantos o modelo conseguiu identificar?"Um alto recall é vital para minimizar falsos negativos e garantir que os pacientes que necessitam de atenção sejam encontrados.
- **F1-Score (Macro Average):** A média harmônica entre Precisão e Recall. Foi utilizada a média **Macro Average**, que calcula a métrica para cada classe e tira a média aritmética. Isso dá peso igual tanto à classe minoritária (hipertensos) quanto à majoritária (não-hipertensos), fornecendo uma visão holística do equilíbrio do modelo.

4.2 Análise Comparativa dos Modelos

A análise da Tabela 2 não aponta um único modelo "vencedor"absoluto. Em vez disso, ela revela um ***trade-off*** claro entre equilíbrio geral (F1-Macro) e capacidade de detecção (Recall).

O pipeline **RandomForest_SMOTE** alcançou o melhor desempenho equilibrado. Ele atingiu o **F1-Macro mais alto (0.72)** e, de forma notável, a **maior Precisão (0.55)** para a classe positiva. Isso o posiciona como o modelo mais confiável: suas previsões positivas têm a maior probabilidade de estarem corretas, gerando menos falsos positivos.

Em contrapartida, o pipeline **DecisionTree_SMOTEENN** atingiu o **Recall mais alto (0.85)** de forma disparada. Este modelo foi o mais eficaz em "encontrar"pacientes hipertensos, minimizando falsos negativos. No entanto, esse ganho em sensibilidade teve um custo significativo em sua precisão (0.45), que foi a mais baixa de todos os testes, indicando um alto volume de falsos positivos.

Os modelos baseados em KNN apresentaram um desempenho competitivo e equilibrado (ex: `KNN_SMOTE` com 0.71 de F1-Macro), mas não superaram o Random Forest em equilíbrio, nem as Árvores de Decisão em sensibilidade de detecção.

5 Conclusão e Próximos Passos

A análise comparativa dos pipelines de modelagem não aponta para um único estimador ótimo de forma absoluta. Em vez disso, os resultados (Tabela 2) revelam um **trade-off estratégico** claro, que depende diretamente do objetivo de implementação clínica.

Se o objetivo primário for um **diagnóstico de alta confiabilidade e equilíbrio**, o pipeline **RandomForest_SMOTE** demonstrou ser a escolha mais robusta. Ele alcançou o maior F1-Macro (0.72) e a melhor Precisão (0.55) entre todas as combinações. Na prática, isso se traduz em um modelo que, ao prever "Hipertensão", tem a maior taxa de acerto, minimizando assim o número de falsos positivos (pacientes saudáveis encaminhados desnecessariamente).

Por outro lado, se o objetivo for uma **ferramenta de triagem de máxima sensibilidade** — onde a prioridade é "encontrar"o maior número possível de indivíduos doentes, mesmo ao custo de mais alarmes falsos — o pipeline **DecisionTree_SMOTEENN** é o mais indicado. Embora sua precisão tenha sido baixa (0.45), sua capacidade de Recall (0.85) foi inigualável, garantindo a minimização de falsos negativos (pacientes doentes não identificados).

Portanto, a seleção final do modelo não é puramente estatística, mas uma decisão funcional sobre qual tipo de erro é mais aceitável para o problema de saúde pública da hipertensão.

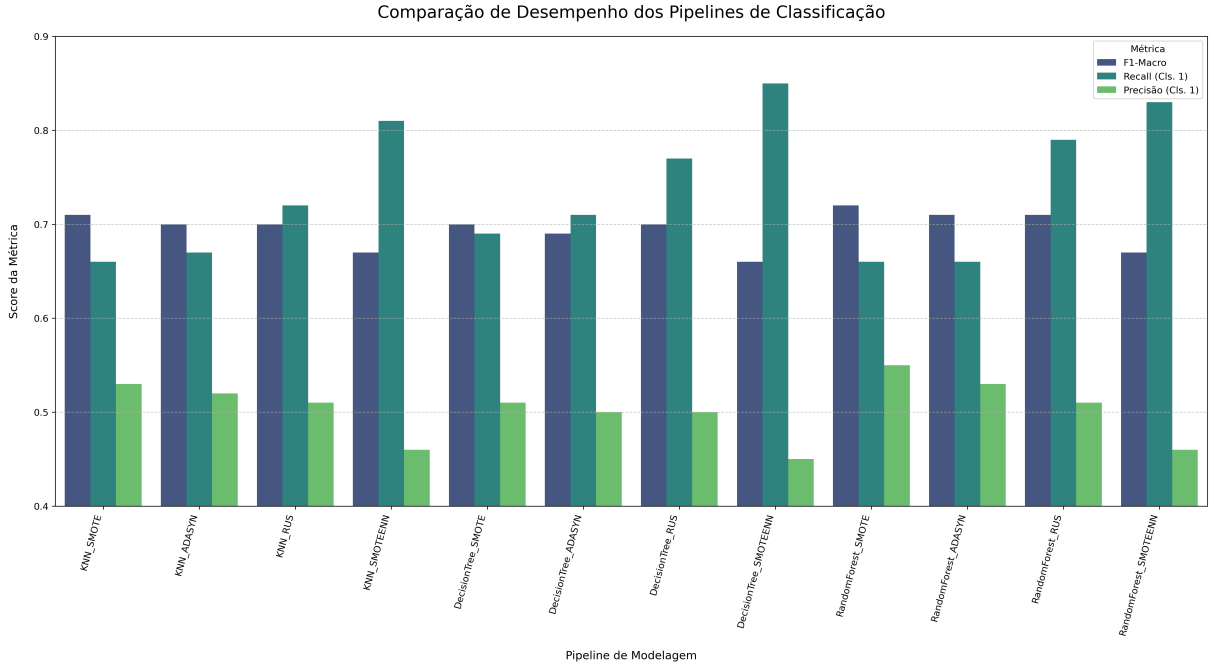

Figura 1: Comparação visual do desempenho dos pipelines no conjunto de teste. O gráfico ilustra o trade-off entre o equilíbrio geral (F1-Macro, melhor no RF_SMOTE) e a sensibilidade de detecção (Recall Classe 1, melhor no DT_SMOTEENN).

Disponibilidade de Código

Todo o código-fonte para este projeto está disponível publicamente:

- **Pré-processamento:** https://github.com/Ettorew/DataMining/tree/main/Codigos
- **Modelagem e Avaliação (ML):** https://github.com/Ettorew/Aprendizado_de_Maquina_I/tree/main/pns

Referências

[1] Y. Li, X. Wang, L. Zhang, Y. Wang, e H. Li. 2022. Application of machine learning to predict hypertension based on dietary and lifestyle risk factors. *BMC Medical Informatics and Decision Making*, 22, 145.

[2] L. J. Appel, T. J. Moore, E. Obarzanek, et al. (DASH Collaborative Research Group). 1997. A Clinical Trial of the Effects of Dietary Patterns on Blood Pressure. *New England Journal of Medicine*, 336, 1117-1124.

[3] E. M. Clañg, et al. 2023. An Open-Source Dataset on Dietary Approaches to Stop Hypertension (DASH) Adherence for Hypertension and Cardiovascular Disease Risk Factor Management. *Nutrients*, 15(12), 2795.

[4] M. Al-Tamimi, et al. 2021. Machine Learning Approaches for Hypertension Prediction Using Framingham Heart Study Data. Em *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2901-2905.

[5] G. M. A. Hasan, et al. 2020. Prediction of Hypertension in the National Health and Nutrition Examination Survey (NHANES) Population Using Machine Learning. *Journal of Clinical Medicine*, 9(4), 1144.