XXXI Congresso de Iniciação Científica Unicamp 2023

Inteligência Artificial: Suporte à Tomada de Decisão no Sistema de Saúde

Palavras-Chave: Aprendizado de Máquina, Abordagens Interpretáveis, Saúde.

Autores(as):
Thiago Miranda Brandão, FCA-UNICAMP
Profa. Dra. Priscila C. B. Rampazzo (orientadora), FCA-UNICAMP

INTRODUÇÃO

No aprendizado supervisionado, dados de entrada e saída são utilizados para construir um modelo de aprendizado automático [Theodoridis e Koutroumbas, 2008]. O objetivo desses métodos é prever a saída de novas variáveis a partir do conhecimento adquirido *a priori*. Os métodos supervisionados podem ser divididos pelas tarefas de Classificação e Regressão. Na Classificação, o objetivo do modelo de aprendizado é atribuir uma classe a um dado desconhecido a partir do conhecimento adquirido de eventos anteriores. De forma matemática, o modelo produz uma função de mapeamento $g : \mathbb{R}^n \rightarrow \{1, ..., m\}$, tal que m é a quantidade de classes das amostras, que permitirá encontrar uma saída y de um vetor de entrada $\boldsymbol{x}$, quando $y = g(\boldsymbol{x})$ [Goodfellow et al., 2016]. A saída do modelo de classificação pode ser tanto um valor discreto, associado à classe da amostra, como uma distribuição de probabilidade sobre a classe. Na tarefa de Regressão, dado um conjunto de amostras de treinamento $(y_i, \boldsymbol{x}_i), y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^d, i = 1, 2, ..., N$, a tarefa é estimar uma função g, que se adeque aos dados e consiga predizer a saída de novos dados de entrada [Theodoridis, 2015]. Diferente da classificação, a saída do modelo de regressão é contínua.

Os modelos clássicos de aprendizado supervisionado que foram escolhidos para serem estudados são: Classificação Linear (LC), Regressão Logística (LR), Florestas Aleatórias (RF) e Redes Neurais Artificiais (ANN), com o intuito de analisar a aplicabilidade de cada método, de acordo com a base de dados. Os Modelos de Aprendizado de Máquina têm sido eficientes no aprendizado de padrões complexos que os permitem fazer previsões sobre dados ainda não observados. Além de usar modelos para predição, a capacidade de interpretar o que um modelo aprendeu está recebendo cada vez mais atenção.

METODOLOGIA

Neste projeto, uma estrutura foi definida e implementada, para todos esses modelos, de acordo com o fluxograma da Figura 1. A partir das bases de dados, escolhidas dos repositórios públicos, a primeira tarefa a se fazer para a construção do modelo foi a padronização das bases. A padronização é um procedimento de pré-processamento de dados com o intuito de transformar os valores de um conjunto, de modo que eles fiquem dentro de um intervalo comum ou distribuídos em uma escala com sua média e desvio padrão. A padronização é uma etapa importante, pois alguns algoritmos são baseados na distância euclidiana, assim, ao dimensionar os dados, o tempo computacional é menor, porque os valores padronizados são mais fáceis de interpretar. Os três tipos de padronização implementados foram: min-max, z-score e escalonamento decimal. Ainda na etapa de pré-processamento, dependendo da base de dados escolhida, pode ser necessária a etapa de separação de dados. Exemplo: para prever risco e mortalidade em pacientes com insuficiência cardíaca, além de estudos com a base original completa, pode ser interessante a separação por sexo,

feminino ou masculino. E, finalmente, a fase de pré-processamento termina com a estratificação da base de dados. Neste projeto, foi escolhido o método *Stratified K-fold* para realizar esta tarefa. Este método divide a base em treino e teste com a vantagem de preservar o desbalanço das classes. A ideia é dividir os dados disponível em K partições (os "folds") e realizar K rodadas de treinamento e teste com essas combinações de dados. Assim, idealmente, minimiza-se as chances de algum dado importante para a classificação ser deixado de fora durante o treinamento. Todas as implementações foram realizadas em linguagem de programação *Python* e com utilização das bibliotecas (*Scikit-learn*) disponíveis na linguagem.

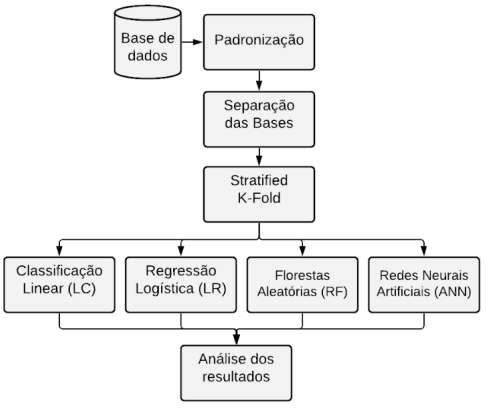

Figura 1: Metodologia do Projeto.

Para a análise dos resultados foram obtidas medidas de desempenho dos modelos de previsão relacionadas à matriz de confusão e à curva ROC. Uma matriz de confusão compara as previsões do modelo aos verdadeiros status padrão para as observações de treinamento no conjunto de dados. Os elementos na diagonal da matriz representam indivíduos cujos status padrão foram previstos corretamente (TPR e TNR), enquanto os elementos fora da diagonal representam indivíduos que foram classificados incorretamente (FPR e FNR) [James et al., 2013]. A Acurácia representa a proporção de classificações corretas dentre todas as classificações do modelo. As classificações positivas corretas dentre todas as classificações positivas realizadas pelo modelo são representadas pela Precisão. Recall é a capacidade de um modelo para detectar todas as amostras positivas. A métrica F1-score combina Precisão e Recall de modo a trazer um número único que indique a qualidade geral do modelo e trabalha bem até com conjuntos de dados que possuem classes desproporcionais. A curva ROC é um gráfico que relaciona a taxa de verdadeiro positivo com a taxa de falso positivo. O desempenho geral de um classificador, resumido sobre todos os limites possíveis, é dado pela área sob a curva (AUC). Uma curva ROC ideal abrangerá o canto superior esquerdo, portanto, quanto maior o AUC, melhor o classificador [James et al., 2013].

RESULTADOS E DISCUSSÃO

Os resultados foram avaliados em uma base de dados de Insuficiência Cardíaca (Repositório UCI). A base conta com 299 amostras e 13 atributos, sendo 105 mulheres e 194 homens. A base de dados é desbalanceada: 68% dos dados são de sobreviventes e 32% dos dados são de pessoas que morreram. Foram realizados dois tipos de experimentos. No primeiro experimento foi considerada a base original e estabelecidos 5 *folds*. Os resultados dos métodos Classificação Linear

(LC), Regressão Logística (LR), Florestas Aleatórias (RF) e Redes Neurais Artificiais (ANN) são apresentados na Tabela 1. Os resultados referem-se à média dos 5 *folds* e indicam boa generalização nos dados de teste, com RF como o classificador com os melhores resultados. Em relação à Precisão, RF e LR apresentaram os melhores resultados.

	LC	LR	RF	ANN
Acurácia	0,833	0,839	0,853	0,776
AUC	0,875	0,876	0,901	0,821
TPR	0,677	0,67	0,729	0,666
TNR	0,906	0,916	0,911	0,827
FPR	0,094	0,084	0,089	0,173
FNR	0,323	0,323	0,271	0,334
F1-score	0,721	0,730	0,759	0,657
Precisão	0,777	0,795	0,795	0,657

Tabela 1: Resultados, base original, teste, média dos 5 *folds*.

No segundo experimento, as bases foram separadas por gênero (Tabelas 2 e 3). Com as bases separadas, o RF continua como o classificador com os melhores resultados; com a base homem o classificador obteve melhor desempenho do que com a base mulher. Considerando a base mulher, a ANN apresentou melhor resultado para a métrica TPR. Não foi possível observar fatores que justificassem a diferença de desempenho.

	LC	LR	RF	ANN
Acurácia	0,804	0,809	0,846	0,762
AUC	0,874	0,874	0,907	0,853
TPR	0,633	0,650	0,729	0,825
TNR	0,886	0,886	0,901	0,825
FPR	0,114	0,114	0,099	0,175
FNR	0,367	0,350	0,271	0,372
F1-score	0,672	0,681	0,751	0,630
Precisão	0,724	0,726	0,776	0,634

Tabela 2: Resultados, base homem, teste, média dos 5 *folds*.

	LC	LR	RF	ANN
Acurácia	0,762	0,781	0,790	0,771
AUC	0,842	0,840	0,904	0,843
TPR	0,552	0,552	0,552	0,576
TNR	0,859	0,888	0,902	0,859
FPR	0,141	0,112	0,098	0,141
FNR	0,448	0,448	0,448	0,424
F1-score	0,599	0,616	0,626	0,596
Precisão	0,685	0,700	0,727	0,637

Tabela 3: Resultados, base mulher, teste, média dos 5 *folds*.

Com o classificador que obteve o melhor resultado (RF), foi realizada a análise de importância de atributos, de acordo com cada base: original, homem e mulher (Figura 2).

Foram estudados três métodos de interpretabilidade [Molnar, 2021]: *Partial Dependence Plot*, *Permutation Importance*, *Shapley Values*. Estes métodos são *model-agnostic*, o que significa que eles podem ser aplicados em qualquer modelo de aprendizado de máquina. Optou-se pela técnica *Shapley Values* para análise da importância dos atributos. O propósito é obter informações sobre o comportamento do modelo de acordo com as variáveis de entrada e entender a importância de cada variável durante a predição [FISHER et al., 2018]. Na técnica *Shapley Values*, o ato de realizar uma predição é como se fosse um jogo: os valores das variáveis são jogadores e a predição do modelo é o resultado final da participação de todos os jogadores. A influência de cada variável é calculada comparando como o modelo se sairia caso não tivesse acesso a este valor. Se a influência é próximo de zero, então temos que o *Shapley Value* influencia pouco na predição; se o *Shapley*

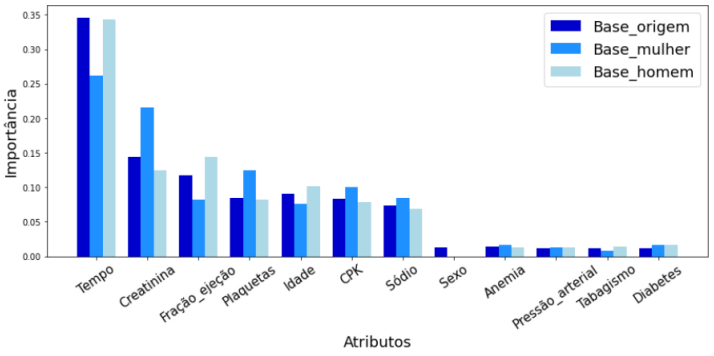

Figura 2: Importância dos atributos, RF, média dos 5 *folds*.

Value é grande, a variável associada tem muita influência na predição do modelo, seja positiva ou negativa.

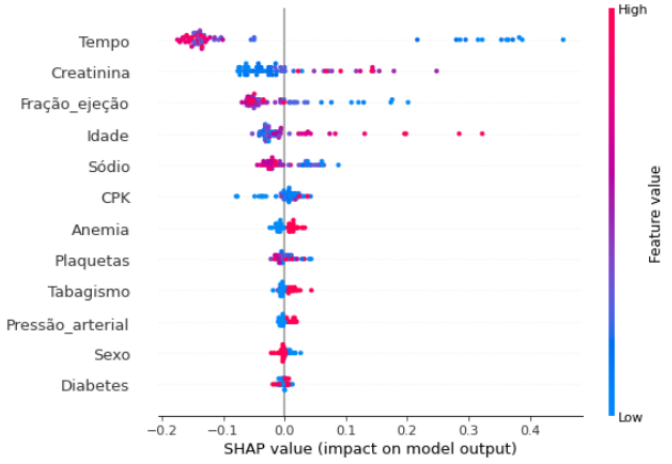

Figura 3: *Shapley Values*, RF, base original, 1 *fold*.

Analisando os resultados da base original, apresentados na figura 3 destacam-se: CREATININA: altos níveis podem ter relação direta com risco cardiovascular; EF e SÓDIO: níveis baixos podem indicar insuficiência cardíaca; IDADE está relacionada com o envelhecimento da população; ANEMIA pode ser um forte marcador que aumenta a taxa de mortalidade. Os resultados destas análises consideram *Shapley Values* de 1 *fold*. Apenas pequenas variações foram observadas nos demais *folds*. Analisando os resultados das bases separadas, homem e mulher, respectivamente Figuras 4 e 5, não foi possível observar diferenças significativas de fatores de riscos associados ao gênero.

CONCLUSÕES

Métodos de Aprendizado de Máquina se apresentam como ferramentas eficientes para trabalhar com dados da área de saúde. Para a base de dados escolhida para análise, o método de Florestas Aleatórias (RF) obteve o melhor desempenho. E não foi possível identificar a causa da diferença de desempenho dos classificadores em bases separadas.

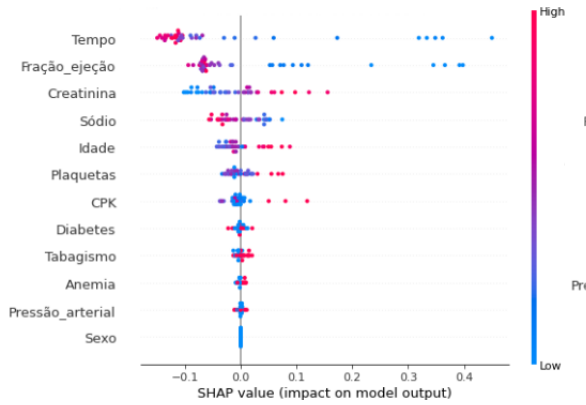

Figura 4: *Shapley Values*, RF, base homem.

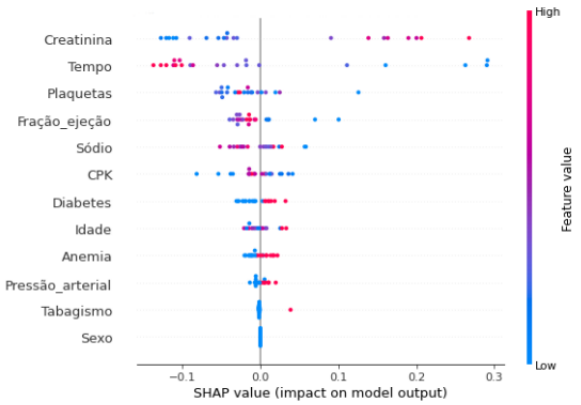

Figura 5: *Shapley Values*, RF, base mulher.

Em relação à análise da importância dos atributos, não foi possível observar diferentes fatores de riscos associados aos gêneros. É importante destacar que as interpretações e análises apresentadas não significam causalidade. Estas análises podem auxiliar na tomada de decisão e no diagnóstico, mas é necessário um melhor entendimento dos dados e do problema, através dos conhecimentos de um especialista.

Referências

FISHER, A., RUDIN, C., e DOMINICI, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.

Goodfellow, I., Bengio, Y., e Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

James, G., Witten, D., e Tibshirani, T. H. R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Molnar, C. (2021). Interpretable machine learning a guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/.

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic Press, USA.

Theodoridis, S. e Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press, USA.