XXXI ENCONTRO ANUAL DE INICIAÇÃO CIENTÍFICA

PREVISÃO DE RISCO DE DIABETES EM ESTÁGIO INICIAL USANDO TÉCNICA DE APRENDIZADO DE MÁQUINA

Willian Penteado (PIBIC/Fundação Araucária-UNICENTRO),
Sandra Mara Guse Scós Venske (Orientadora), Carolina Paula de Almeida.
e-mail: willianpenteado5@gmail.com
Universidade Estadual do Centro-Oeste - UNICENTRO,
Departamento de Ciência da Computação. Guarapuava, PR.

Engenharias - Engenharia de Produção

Palavras-chave:
Aprendizado supervisionado, problemas de classificação, seleção de características.

Resumo

O Aprendizado de Máquina (AM) é o campo de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programado. O objetivo deste trabalho foi o estudo de técnicas de AM aplicadas no problema de classificação de *Risco de Diabetes em Estágio Inicial*. O Diabetes é uma doença causada pela produção insuficiente ou má absorção da insulina e atinge mais de 13 milhões de pessoas no Brasil. Onze classificadores foram testados no problema: Regressão Logística, K-ésimo Vizinho mais Próximo, Máquinas de Vetores de Suporte, Processo Gaussiano, Árvores de Decisão, Florestas Aleatórias, *Perceptron* Multicamadas, *AdaBoost*, *Naive Bayes* e Análise Discriminante Quadrática. Além disso, foram estudadas duas técnicas de seleção de características e utilizadas duas métricas de avaliação dos modelos. O modelo de Florestas Aleatórias obteve melhor desempenho, enquanto o *Naive Bayes*, o pior dentre todos os classificadores.

Introdução

O Aprendizado de Máquina (AM) é o processo de indução de hipóteses a partir de experiências anteriores (FACELI et al., 2021). Neste trabalho foi usado um conjunto de dados (*dataset*) público sobre diabetes do *National Institute of Diabetes and Digestive and Kidney Diseases*. O *dataset* prediz, a partir de características de diagnóstico, se um paciente possui diabetes.

O objetivo deste trabalho foi o estudo de técnicas de AM supervisionado aplicadas no problema de classificação de *Risco de Diabetes em Estágio Inicial*. No pré-processamento dos dados foi utilizada a técnica de *holdout* (FACELI et al., 2021), dividindo o *dataset* em conjunto de treinamento e teste. Nesta fase foram

aplicadas técnicas de seleção de características usando *chi2* e *Recursive Feature Elimination* (RFE) (FACELI et al., 2021). Em seguida, onze classificadores foram testados (GÉRON, 2019): Regressão Logística (RL), K-ésimo Vizinho mais Próximo (KNN), Máquinas de Vetores de Suporte (SVM) núcleo Linear e RBF, Processo Gaussiano (PG), Árvores de Decisão (AD), Florestas Aleatórias (FA), *Perceptron* Multicamadas (MLP), *AdaBoost*, *Naive Bayes* (NB) e Análise Discriminante Quadrática (QDA). Para avaliação dos modelos, foi utilizada validação cruzada com as métricas acurácia e F_1-*score*.

Materiais e métodos

O fluxograma do algoritmo proposto é apresentado na Figura 1.

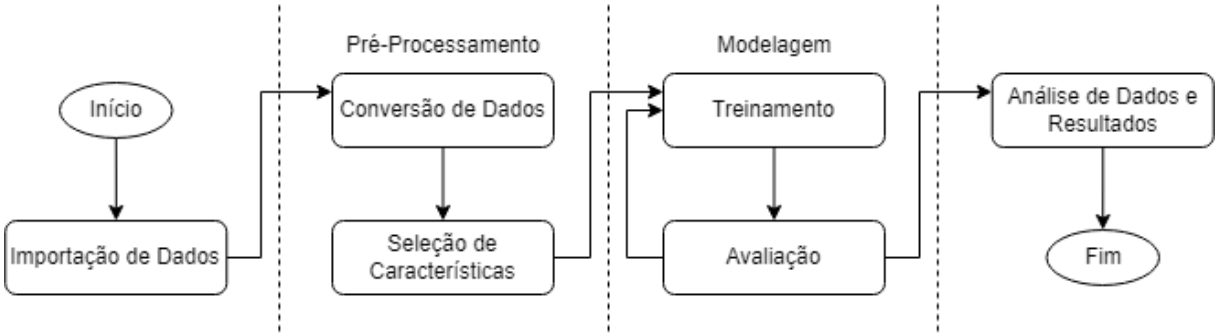

Figura 1 - Fluxograma do algoritmo desenvolvido. Fonte: o autor.

O algoritmo inicia com a *Importação de Dados*, onde carrega o *dataset*. No *Pré-Processamento* foram feitas as conversões dos tipos de dados e implementadas as técnicas de *Seleção de Características chi2* e RFE. Esta seleção usou *holdout*, que é a divisão do *dataset* em conjuntos de treinamento (67%) e teste (33%). Dentro da *Modelagem*, no *Treinamento*, foram utilizados onze classificadores. A *Avaliação* testou a eficácia do modelo, aplicando validação cruzada *k-fold* com as métricas acurácia e F_1-*score*. Na etapa final, foi feita análise dos resultados dos modelos, com auxílio de tabelas.

Resultados e Discussão

Para o desenvolvimento do algoritmo proposto foi usado o *framework Scikit-Learn*[1], baseado na linguagem Python[2] (GÉRON, 2019). Os dados utilizados

[1] https://scikit-learn.org/stable
[2] https://www.python.org

foram do *dataset* público "*Early stage diabetes risk prediction*" (previsão de risco de diabetes em estágio inicial)[3] que foram coletados utilizando questionários de pacientes do *Sylhet Diabetes Hospital* em Sylhet, Bangladesh. O *dataset* é composto por 520 instâncias (casos) e 17 atributos (características): sexo, se o paciente possui poliúria, polidipsia, perda repentina de peso, fraqueza, polifagia, aftas genitais, desfoque visual, coceira, irritabilidade, cicatrização retardada, paresia parcial, rigidez muscular, alopecia e obesidade. A classificação das instâncias pode ser positiva, caso o paciente tenha a doença, ou negativa, caso contrário.

O primeiro teste realizado foi relacionado à seleção de quais características do paciente são mais importantes para o diagnóstico. Duas técnicas de AM foram utilizadas para fazer esta seleção: *chi2* e RFE. A técnica RFE foi a melhor.

No segundo teste, foi utilizada a RFE para testar os 11 classificadores, com 3 divisões de base com validação cruzada k-*fold*, com k = {3, 5, 10}. Eles foram avaliados segundo as métricas de acurácia e F1-*score*[4] (equações a seguir). A validação cruzada 10-*fold* obteve melhores resultados, apresentados na Tabela 1.

$$Precisao = \frac{VP}{VP + FP} \qquad Sensibilidade = \frac{VP}{VP + FN}$$

$$F_1 = 2 \times \frac{Precisao \times Sensibilidade}{Precisao + Sensibilidade} = \frac{VP}{VP + \frac{FN+FP}{2}}$$

Nas equações acima, VP são Verdadeiros Positivos, VN são Verdadeiros Negativos; FP são Falsos Positivos; e FN são Falsos Negativos.

Na primeira coluna são listados os classificadores, a coluna "k = 10", mostra o resultados sem seleção de características utilizando validação cruzada. As colunas seguintes apresentam os resultados utilizando a seleção de características com as c características mais importantes. Os resultados com c < 6 foram removidos da tabela devido a questão de espaço e considerando que obtiveram piores resultados.

A Tabela 1 aponta que o melhor classificador é o de FA com 99% de F1-*score*, seguido pelo KNN e PG, ambos com 97%. O KNN demonstra sensibilidade à seleção de características. Esses classificadores também tiveram destaque considerando a métrica de acurácia. O classificador NB obteve o pior desempenho dentre todos os classificadores, com 90% de F1-*score*.

[3] Disponível em: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset
[4] Por questão de espaço somente são apresentados os resultados para a métrica F1-*score*.

Tabela 1: Resultados com RFE e 10-*fold* em termos de F1-*score*. Os melhores valores estão destacados em uma escala de cinza, sendo o mais escuro o melhor.

Classificador	K = 10	c = 15	c = 14	c = 13	c = 12	c = 11	c = 10	c = 9	c = 8	c = 7
RL	0.9386	0.9385	0.9402	0.9385	0.9355	0.9321	0.9354	0.9287	0.9265	0.9087
KNN	0.9402	0.9796	0.9730	0.9668	0.9748	0.9733	0.9639	0.9534	0.9548	0.9225
SVM Linear	0.9397	0.9355	0.9399	0.9401	0.9331	0.9317	0.9301	0.9300	0.9286	0.9179
SVM RBF	0.9633	0.9428	0.9427	0.9416	0.9401	0.9431	0.9412	0.9268	0.9303	0.9179
PG	0.9781	0.9781	0.9765	0.9733	0.9700	0.9669	0.9670	0.9596	0.9531	0.9236
AD	0.9648	0.9504	0.9587	0.9570	0.9587	0.9605	0.9513	0.9572	0.9328	0.9251
FA	0.9907	0.9814	0.9814	0.9765	0.9749	0.9719	0.9674	0.9565	0.9457	0.9299
MLP	0.9335	0.9458	0.9427	0.9477	0.9465	0.9388	0.9401	0.9404	0.9364	0.9073
AdaBoost	0.9384	0.9272	0.9272	0.9238	0.9167	0.9253	0.9141	0.9175	0.9287	0.9089
NB	0.9097	0.9157	0.9174	0.9136	0.9143	0.9137	0.9085	0.9116	0.9116	0.9116
QDA	0.9525	0.9524	0.9525	0.9375	0.9376	0.9282	0.9212	0.9193	0.9259	0.9253

Portanto, dentre todos os classificadores o FA foi o que obteve melhor desempenho, em ambas as métricas aplicadas, com 98% de acurácia e 99% de F1-*score*, sem seleção de características e com validação cruzada de 10-*fold*.

Considerações Finais

O objetivo deste trabalho de pesquisa foi o estudo de técnicas de Aprendizado de Máquina (AM) supervisionado a serem aplicadas em um problema de classificação - a previsão de riscos de diabetes em estágio inicial.

Como seleção de características, a RFE foi a melhor técnica comparada com a *chi2*. O modelo de Florestas Aleatórias foi o classificador que obteve melhor desempenho utilizando validação cruzada 10-*fold*.

Como trabalhos futuros, pode-se fazer uma análise dos parâmetros dos modelos, além de estudar outras técnicas de seleção de características.

Agradecimentos

Os autores agradecem o apoio da Fundação Araucária.

Referências

FACELI, K., LORENA, A. C., GAMA, J., E CARVALHO, A. C. P. D. L. F. D. (2021). **Inteligência artificial: uma abordagem de aprendizado de máquina.** LTC, 2 ed.

GÉRON, A. (2019). **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow.** Alta Books, 1 ed.