CLASSIFICADOR ENSEMBLE: UMA ABORDAGEM NÃO PARAMÉTRICA APLICADO À DETECÇÃO DE DIABETES

Adriana dos Santos Lima[1], Silvio Cabral Patricio[2], Leonara Alves Cesario da Silva[3] e Renato Valladares Panaro[4]

Introdução

Se tratando de dados dicotômicos, isto é, que admitem somente duas respostas possíveis, os modelos supervisionados de aprendizado de máquina comumente utilizados são: regressão logística, árvores aleatórias e *K–Nearest neighbor* (KNN). No entanto, a depender dos dados, tais predições podem não apresentar uma boa acurácia (> 70%). Surge então, como alternativa aos modelos usuais, os classificadores ensemble (Gul & Perperoglou, 2018).

O método ensemble é uma técnica de aprendizado de máquina que combina o resultado de múltiplos modelos em busca de produzir um melhor modelo preditivo. Existem vários algoritmos pré-fixados de classificadores *ensemble*, tais como: *bagging, boosting, bayesian averaging*, entre outros. No entanto, a escolha dos modelos preditivos base e a maneira que estes resultados serão combinados são livres (Opitz & Maclin, 1999).

Os classificadores ensemble são uma classe de métodos utilizados para aumentar a acurácia do modelo com a junção de modelos mais fracos, quando o emprego de métodos mais simples, separadamente, não apresentam o resultado desejado. Para a aplicação de métodos mais sofisticados para a predição dos dados, faz-se necessário o uso de técnicas iniciais que permitam o correto aproveitamento do modelo.

Objetivos

O trabalho proposto tem por objetivo apresentar um classificador *ensemble* como uma alternativa aos modelos usuais de Aprendizado de Máquina, onde a abordagem apresentada será aplicada a dados de detecção de diabetes.

[1] Universidade Federal de Minas Gerais (UFMG), adrianalima@ufmg.br

[2] Universidade Federal de Minas Gerais (UFMG), silviocp@ufmg.br

[3] Universidade Federal do Rio de Janeiro (UFRJ), leonara@dme.ufrj.br

[4] Universidade Federal de Minas Gerais (UFMG), renatovp@ufmg.br

Material e Método

Os dados utilizados foram retirados do Instituto Nacional de Diabetes e Doenças Digestivas e Renais (National Institute of Diabetes and Digestive and Kidney Diseases). Nesse caso, quer-se prever, com base em medidas de diagnóstico, se um paciente tem diabetes. Usou-se diversas restrições na seleção dessas instâncias de um banco de dados maior. Em particular, todos os pacientes escolhidos são do sexo feminino com pelo menos 21 anos de idade. As variáveis presentes no conjunto de dados são: número de gestações, concentração plasmática de glicose em teste oral de tolerância à glicose, pressão sanguínea, grossura da pele, insulina, índice de massa corporal (IMC), índice de histórico de diabetes, idade e classe (0 – desenvolveu a doença, 1 – não desenvolveu a doença).

Em seguida, fez-se uma análise exploratória dos dados para uma melhor visualização das variáveis. Constatou-se que existiam erros de medição nas variáveis de glicose, pressão sanguínea, grossura da pele, insulina e índice de massa corporal, pois apresentavam valores nulos que, de fato, não tem coerência com o que está sendo medido. Além disso, observou-se a presença de *outliers* em três das variáveis, sendo elas: glicose, pressão sanguínea e grossura da pele. Para o pré-processamento dos dados, efetuou-se o método de imputação de dados de média ponderadas para o tratamento dos valores faltantes e dos outliers e, para a normalização, realizou-se o procedimento de escala por máximos e mínimos.

Após a verificação de inconsistência e limpeza do banco de dados, dividiu-se 75% dos dados para treinamento e 25% para teste. Aplicou-se, então, os modelos de regressão logística, árvores aleatórias e KNN. Para o método ensemble, 25% dos dados de treino foram para selecionados para treinamento dos classificadores mais fracos (previsões do knn com acurácia menor que 65%) e suas predições foram passadas para o modelo de árvores aleatórias onde este, com base nos resultados obtidos na modelagem anterior, é responsável por prever se o indivíduo é acometido pela diabetes. Por fim, comparou-se a acurácia de cada um deles.

A análise de dados foi feita utilizando o software estatístico R. O pacote *data.table* foi utilizado para a devida leitura do banco de dados (Dowle & Srinivasan, 2019). Para manipulação e tratamento inicial dos dados foi utilizado o pacote *dplyr* (Wickham, 2019). Ao constatar inconsistência no banco de dados, tais variáveis observadas com erro foram consideradas como valor faltante e foram imputadas utilizando o pacote *mice* (Buuen & Groothuis-Oudshoorn, 2011). Os dados de treino e teste foram alocados segundo a saída do pacote *caret* (Kuhn at all, 2019). As modelagens de predição foram feitas com os seguintes pacotes: *randomForest* (Liaw & Wiener, 2002) para o algoritmo de floresta aleatória e *class*

(Venables & Ripley, 2002) para o KNN. Por fim, a matriz de confusão e a acurácia dos modelos foram calculadas a partir do pacote *caret*.

Resultados e Discussão

A amostra compôs-se de 768 indivíduos do sexo feminino com idade compreendida entre 21 e 66 anos, e IMC médio de 32,24 Kg/m². Verificou-se, também, que somente 34,67% dos indivíduos apresentam diagnóstico de diabetes. A fim de diagnosticar um indivíduo como diabético (ou não) a partir de suas características físicas e fisiológicas, foram ajustados diversos modelos de predição.

A Figura 1 mostra a acurácia do modelo KNN comparando diversos valores de números de vizinhos (k). É possível observar que não há um padrão consistente em proferir que ao aumentar o numéro de vizinhos, aumenta-se a acurácia. Além disso, nota-se que os modelos obtiveram acurácia variando de 60% a 80%.

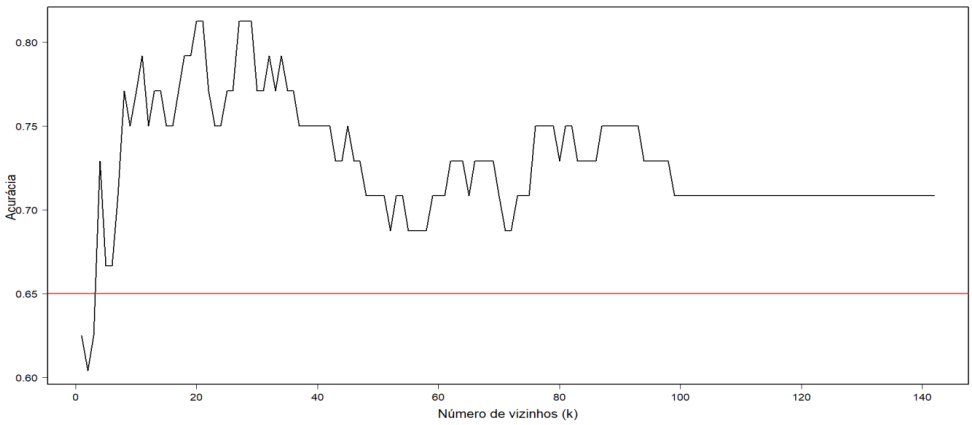

Figura 1 : Acurácia do Modelo KNN ao aumentar o número de vizinhos (k).

Após o ajuste e seleção daqueles modelos que fariam parte do ajuste via árvores aleatórias, fez-se a matriz de confusão para a comparação dos resultados obtidos com os observados para os dados de teste. A Tabela 1 apresenta os resultados finais do método ensemble.

Tabela 1: Matriz de Confusão do Método Ensemble.

		Valor de referência	
		Não Diabético	Diabético
Predição	Não Diabético	105	6
	Diabético	21	58

Fonte: Elaborado pelos autores.

Nota-se que 27 indivíduos foram classificados incorretamente, resultando em uma sensibilidade de 91% e especificidade de 83%. Além disso, ouve um aumento de aproximadamente 5,59% na acurácia do modelo proposto em relação ao modelo com maior acurácia via KNN.

A seguir, ilustra-se, na Tabela 2, a comparação entre todos os modelos ajustados neste trabalho. O modelo de árvore aleatória apresentado foi ajustado utilizando 200 árvores e, no mínimo, cinco nós terminais. Já para o modelo KNN foi utilizado k = 23 vizinhos. Percebe-se que o modelo ensemble tem melhor desempenho na predição de diabetes com acurácia de 85,79%. O modelo de árvore aleatória apresentou a segunda melhor performance com acurácia de 78,70%, seguida pelo KNN (78,26%) e, logo após, pelo modelo de regressão logística (77,83%).

Tabela 2: Comparação da Acurácia do Método Ensemble com outros Modelos de Predição.

Modelo	**Acurácia (%)**
Regressão Logística	77,83%
Árvore Aleatória	78,70%
KNN	78,26%
Ensemble	85,79%

Fonte: Elaborado pelos autores.

Conclusão

A utilização de métodos ensemble no software R como uma ferramenta para a previsão de dados de diabetes, mostrou-se uma opção promissora que pode ser aplicável a outros conjuntos de dados como alternativa para os métodos usuais encontrados em modelos de aprendizagem de máquina. Para projetos futuros, pretende-se ampliar a aplicação desses modelos para outros bancos de dados e aproveitar-se de outras metodologias, tais como: redes neurais, máquinas de vetores de suportes, dentre outras,

como maneiras de agregar formas robustas de predição para as variáveis objetivas no modelo.

Referências

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Gul, A. & Perperoglou, A. (2018) Ensemble of a subset of kNN classifiers. Second Edition. Springer, New York. ISBN 12:827–840

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr

Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.0. https://CRAN.R-project.org/package=data.table

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-82. https://CRAN.R-project.org/package=caret

Opitz, D. & Maclin, R. (1999) Popular Ensemble Methods: An Empirical Study. Second Edition. Journal Of Artificial Intelligence Research, USA. ISBN 11: 169-198

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0