XXXIII SIC SALÃO INICIAÇÃO CIENTÍFICA

Evento	Salão UFRGS 2021: SIC - XXXIII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2021
Local	Virtual
Título	Algoritmos de Ensemble para Identificação de Biomarcadores da COVID-19
Autor	PEDRO HENRIQUE MENDES DUARTE
Orientador	MARCO AURELIO PIRES IDIART

Algoritmos de Ensemble para Identificação de Biomarcadores da COVID-19

Aluno: Pedro Henrique Mendes
Professor: Marco A. P. Idiart

Porto Alegre, Agosto de 2021

Algoritmos de aprendizado de máquina (AM) tem se tornado uma importante ferramenta para geração de modelos acurados para sistemas complexos que envolvem várias variáveis. Uma das vantagens destes algoritmos é exemplificar o comportamento a partir dos dados, sem necessidade de teorizarmos a relevância de cada variável do conjunto de dados. Um dos principais uso de AM é para categorização/predição do comportamento de um sistema, ou seja, como prever o comportamento final do sistema a partir de valores de um conjunto de dados de entrada. Aqui neste estudo, estes dados de entrada serão valores de medidas quantitativas de testes clínicos coletados a partir de amostras de sangue de pacientes internados por problemas associados a COVID-19, e o desfecho que estamos interessados em predizer é se o paciente terá alta ou irá a óbito. Os algoritmos que utilizamos são da classe chamada de algoritmos de *ensembles* que, ao invés de produzirem um único modelo, produzem muitos modelos preditivos que serão usados em conjuntos para resolver o problema. Durante o processo de treinamento a importância das variáveis de entrada na capacidade preditivas dos modelos é avaliada a cada iteração, e só serão guardadas as variáveis mais relevantes. Baseado na Ref. [1] foi analisado um conjunto de dados obtido no Hospital Tongji, em Wuhan, China, entre 10/01/2020 e 18/02/2020. Esses dados contém 375 casos de COVID-19 com 201 pacientes recuperados e 174 óbitos. O principal objetivo de realizar esse estudo é identificar quais biomarcadores que indicam os casos mais graves de COVID-19. Obtemos que os biomarcadores mais importantes são: nível de Desidrogenase Lática, percentual de Neutrófilos, percentual de Linfócitos, contagem de Neutrófilos e nível de Procalcitonina. As acurácias obtidas foram 82% para *Random Forest* e 81% para *Extreme Gradient Boosting*, dois algoritmos de *ensemble*.

Referências

[1] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, and Y. Yuan, "An interpretable mortality prediction model for covid-19 patients," *Nature Machine Intelligence*, vol. 2, pp. 283–288, May 2020.