Pós-Graduação da Faculdade de Saúde Pública da USP (FSP/USP)

Disciplina EPI 5717 Machine Learning para Predições em Saúde

2º Semestre – 2022
Créditos: 3 – 45 horas
Local: FSP/USP
Horário: quartas-feiras das 14 às 18 horas

Professor Responsável
Alexandre Chiavegatto Filho

Justificativa:
O rápido aumento na quantidade de dados tem aberto novas oportunidades para a saúde brasileira. Entre as várias novidades proporcionadas pelo big data em saúde, uma das mais promissoras é o uso de modelos preditivos de inteligência artificial, conhecidos como machine learning. A disciplina tem como objetivo apresentar essa área em rápido crescimento com foco nas suas aplicações práticas, além de discutir seus benefícios, limitações e possíveis uso na área da saúde. O foco do curso será no tipo de dado mais coletado em saúde, i.e. dados estruturados/tabulares, e será utilizada a linguagem Python.

Programa
1 – Perspectivas do uso de inteligência artificial em saúde.
2 – Pré-processamento dos dados (padronização, one-hot encoding, imputação, outliers, rebalanceamento, vazamento de informação).
3 – Sobreajuste e divisão da amostra em treino, validação e teste.
4 – Mensuração da performance de algoritmos preditivos (área abaixo da curva ROC, precisão, recall, especificidade, valor predito negativo e raiz quadrada do erro quadrático médio).
5 – Algoritmos para predição de variável dependente contínua (regressões lineares penalizadas com lasso e ridge, redes neurais, random forests, XGBoost, lightGBM e catboost).
6 – Algoritmos para predição de variável dependente binária (regressões logísticas penalizadas, redes neurais, random forests, XGBoost, lightGBM e catboost).
7 – Técnicas de otimização de hiperparêmetros.
8 – Estratégias para a seleção de variáveis preditoras (Boruta).
9 – Aprendizado federado.
10 – Aprendizado online (contínuo).
11 – Estratégias para a identificação da importância de variáveis preditoras (Shapley values).
12 – Desafios éticos do uso de machine learning em saúde.

Avaliação
A avaliação será realizada por meio de um trabalho final (60%) e exercícios realizados ao longo da disciplina (40%), sendo destes 20% referente aos entregáveis e 20% à discussão de artigos.

Discussão de artigos: todos os alunos devem entregar antes do início da aula uma revisão de menos de uma página para cada um dos dois artigos, com um parágrafo de resumo e o resto uma avaliação sobre a importância e a qualidade do artigo em questão.
Cinco alunos serão escolhidos para cada artigo, em que um irá apresentar o artigo e os outros quatro irão liderar o debate do artigo em relação à sua importância e qualidade.

Observação
Para realizar o curso é necessário ter conhecimentos pelo menos básicos de estatística e programação.

Bibliografia

Batista AFM, Chiavegatto Filho ADP. Machine Learning aplicado à Saúde. In: Artur Ziviani; Natalia Castro Fernandes; Débora Christina Muchaluat Saade. (Org.). Livro de Minicursos. Niterói, RJ: Sociedade Brasileira de Computação, 2019.

Chiavegatto Filho ADP, Batista AFM, dos Santos HG. Data leakage in health outcomes prediction with machine learning. Journal of Medical Internet Research 2021; 23(1).

Fernandes TF, de Oliveira TA, Teixeira CE, Batista AFM, Costa GD, Chiavegatto Filho ADP. A multipurpose machine learning approach to predict COVID-19 negative prognosis in Sao Paulo, Brazil. Scientific Reports 2021; 3343(11).

Geron A. Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow. Alta Books; 2019.

Raschka S, Mirjalili V. Python Machine Learning - Third Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing; 2020.

Topol E. Deep Medicine: How artificial intelligence can make healthcare human again. Basic Books; 2019.

Extra:

Metz C. Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and the World. Cornerstone; 2021.

Cronograma – 2022

Data	Tópico	Referência
13/09	Aula 1 - Discussão do conteúdo programático e apresentação da área.	Lones MA. How to avoid machine learning pitfalls: a guide for academic researchers. arXiv:2108.02497. 2021.
20/09	Aula 2 – Monitoria de introdução ao Python.	
27/09	Aula 3 – Pré-processamento dos dados.	"A Comprehensive Guide to Data Preprocessing" https://neptune.ai/blog/data-preprocessing-guide
04/10	Aula 4 – Sobreajuste, viés e variância; divisão da amostra em treino, validação e teste.	Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808. 2020.
06/10	Aula 5 – Monitoria.	
11/10	**Entregável 1: pré-processamento e divisão da amostra em treino e teste.** Aula 6 - Mensuração da performance de algoritmos preditivos e otimização de hiperparêmetros.	"Tour of Evaluation Metrics for Imbalanced Classification" https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
13/10	Aula 7 – Monitoria.	
18/10	**Aula 8** – Discussão de artigos científicos de machine learning em saúde.	Finlayson SG et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021; 385(3):283-286. Futoma J et al. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(9):e489-e492.
25/10	Aula 9– Principais algoritmos de machine learning para dados estruturados: regressões penalizadas, redes neurais e algoritmos de árvore (árvores de decisão, random forests e gradient boosting: XGBoost, lightGBM e catboost).	Al-Shari H, Saleh YA, Odabas A. Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance. 2021.

27/10	Aula 10 – Monitoria.	
01/11	Aula 11 - Estratégias para a seleção de variáveis preditoras.	Degenhardt F. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019;20(2):492-503.
3/11	Aula 12 – Monitoria.	
8/11	**Aula 13** - Discussão de artigos científicos de machine learning em saúde.	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Grinsztajn L at al. Why do tree-based models still outperform deep learning on tabular data? arXiv: 2022.
10/11	**Aula 14** - Discussão de artigos científicos de machine learning em saúde.	Pfaff ER et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digital Health 2022; 4(7):E532-541. Lo-Ciganic WH et al. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. Lancet Digital Health 2022; 4(6):E455-E465.
22/11	**Entregável 2: predição.** Aula 15 - Estratégias para a identificação da importância de variáveis preditoras.	Molnar CM et al. Interpretable machine learning -- a brief history, state-of-the-art and challenges. arXiv:2010.09337. 2020
24/11	Aula 16 – Monitoria.	
29/11	Aula 17 - Aprendizado federado (federated learning) e aprendizado online (contínuo).	T. Li et al. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing 2020;37(3):50-60. Hoi SCH et al. Online Learning: A Comprehensive Survey. arXiv:1802.02871. 2018
1/12	Aula 18 - Monitoria.	
6/12		Char DS et al. Identifying Ethical Considerations for Machine Learning

	Aula 19 - Desafios éticos do uso de machine learning em saúde.	Healthcare Applications. Am J Bioeth 2020;20(11):7-17. Rajkomar A et al. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med 2018;169(12):866-872. Novo: Ethical Machine Learning in Healthcare
8/12	Aula 20 - Monitoria.	
13/12	Aula 21 – Apresentação dos trabalhos.	