Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

ICICT
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

PLANO DE ENSINO – PPGICS

( ) Verão (X) 2018.1 ( ) Inverno ( ) 2018.2

IDENTIFICAÇÃO			
Disciplina: **Ciência de Dados aplicada à Saúde**			
Código:	Créditos: 04	Carga Horária: 120h	**Período** Início: 21/03/2018 Término: 18/07/2018 Dia da Semana: Quarta Horário: 14h - 17h30
Coordenador da Disciplina: Marcel Pedroso e Christovam Barcellos Professores convidados: Alexandre Chiavegatto (FSP/USP), Eduardo Ogasawara (CEFET), Fabio Porto (LNCC), Jefferson Lima (ICICT), Mario Filho (KAGGLE), Paulo Borges (ICICT), Raphael Saldanha (ICICT), Ricardo Dantas (ICICT) e Vanderlei Pascoal (ICICT).			
Linha 1: ( ) 1.1 ( ) 1.2 (X) 1.3 ( ) 1.4 (X) 1.5 ( ) 1.6 (X) 1.7 (X) 1.8			
Linha 2: ( ) 2.1 ( ) 2.2 ( ) 2.3 ( ) 2.4			

RELAÇÃO DOS PROFESSORES COM A TEMÁTICA DA DISCIPLINA *(opcional)*

A disciplina integra as atividades e objetivos do projeto de pesquisa e desenvolvimento tecnológico "Plataforma de Ciência de Dados aplicada à Saúde" coordenado pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz (Icict/Fiocruz) em parceria com o Laboratório Nacional de Computação Científica (LNCC). http://bigdata.icict.fiocruz.br/.

O referido projeto está inserido no Grupo de pesquisa "Ciência de Dados aplicada à Saúde", certificado pela Fiocruz e cadastrado no Diretório de Grupos de Pesquisa no Brasil. As principais repercussões dos trabalhos do grupo estão relacionadas com a coleta, processamento e análise de informações por meio da Ciência de Dados, fomentando o planejamento, monitoramento, avaliação de políticas públicas e serviços de saúde em tempo real, gerando indicadores de alerta e painéis de monitoramento bastante específicos. Tem como objetivo principal desenvolver e disponibilizar Plataforma de Ciência de Dados (PaaS – *Platform as a Service*) como um serviço de armazenamento, gestão e análise de Big Data em Saúde para pesquisadores, docentes e discentes de instituições de ensino e pesquisa, bem como, gestores governamentais.
http://dgp.cnpq.br/dgp/espelhogrupo/4230691756969719.

EMENTA

A Ciência de Dados é um conjunto de estratégias, ferramentas e técnicas que busca reunir equipes multidisciplinares formadas por pesquisadores com conhecimento substantivo do problema em análise (no nosso caso saúde pública), estatísticos, matemáticos e cientistas da computação. Trata-se de um campo de estudo bastante promissor e destaca-se pela capacidade de auxiliar a descoberta de informação útil a partir de grandes bases de dados e a tomada de decisão orientada por dados (*data-driven analysis*).

Ela combina métodos tradicionais de análise com algoritmos sofisticados para processar grandes volumes de dados em formatos diversos. O processo de análise no âmbito da Ciência de Dados envolve as fases de (i) coleta e ingestão: extração, transformação e carga (mais conhecido como ETL, do inglês *Extract Transform Load*); (ii) pré-processamento: seleção de registros, redução de dimensionalidade, normalização, criação

de subconjuntos de dados; (iii) análise exploratória e mineração de dados: principalmente análises voltadas para classificação, associação, agrupamento, detecção de anomalias e predição; (iv) pós-processamento: interpretação de padrões, filtragem, visualização e acoplamento em sistemas de apoio a decisão e plataformas online para visualização.

O termo "Big Data" vem despertando atenção fora dos grupos de pesquisa acadêmica que estão na fronteira do conhecimento em ciência da computação, física de partículas, genética e astronomia. Para definir Big Data, segundo nossas premissas, com certeza estamos falando de um volume de dados muito grande, mas além de grandes volumes existem outras características importantes na composição do conceito. Além do volume (que deve ser "Big") uma de suas principais características é a variedade de dados a serem processados, que podem ser dados estruturados, dados semiestruturados e dados não estruturados (comentários em redes sociais, blogs, websites, buscas no Google, etc.). Outro fator que contribui para a caracterização de Big Data é a velocidade necessária para o processamento das grandes e diversas bases de dados armazenadas, e com a possibilidade de processamento em tempo real. Em ambos os casos a inovação está na adoção do processamento paralelo e distribuído em diversas máquinas.

No âmbito do setor saúde não é difícil imaginar as possibilidades da abordagem da Ciência de Dados para análise, monitoramento, predição de eventos (casos) e situações de saúde e doença na população, bem como a associação destes com seus determinantes socioambientais. O setor saúde já produz uma quantidade enorme de dados sobre as pessoas que acessam o SUS, porém é importante também termos informações disponíveis sobre quem ainda não acessou, e isso só é possível com integração de bases externas ao setor e processamento em tempo real, como por exemplo, as redes sociais, blogs e mídia digital.

O conteúdo disciplinar será desenvolvido por meio de procedimentos que se sedimentam nos pressupostos didáticos de que "para se aprender, tem que fazer" e de que "é só fazendo, que se apreende". Serão construídos exercícios e casos de estudo abordando atividades de indexação, extração e análise visual de grandes quantidades de dados do setor saúde e seus determinantes socioambientais, bem como atividades de mineração de dados e análise preditiva utilizando a infraestrutura computacional da Plataforma de Ciência de Dados aplicada à Saúde do Icict contribuindo, dessa forma, com os pressupostos de inovação tecnológica e aprendizagem colaborativa da disciplina.

OBJETIVOS

- Apresentar e manusear os principais sistemas de informação em saúde e de interesse para a saúde;
- Capacitar os alunos em abordagens teóricas e metodológicas para a análise de grandes quantidades de dados em diferentes formatos por meio de estratégias e técnicas relacionadas a Ciência de Dados aplicada à Saúde;
- Promover uma interface entre aspectos teóricos e práticos sobre Ciência de Dados, mineração de dados, aprendizagem de máquina, análise preditiva e análise visual de grandes quantidades de dados do setor saúde e de seus determinantes socioambientais;
- Fomentar a utilização acadêmica da Plataforma de Ciência de Dados aplicada à Saúde promovendo inovação tecnológica e aprendizagem colaborativa.

BIBLIOGRAFIA BÁSICA

AGARWAL, RITU, DHAR, VASANT. Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, Sep 2014, Vol.25(3), pp.443-448.

Antunes, JLF; Cardoso, MRA. Uso da análise de séries temporais em estudos epidemiológicos. Epidemiol. Serv. Saúde, 24(3): 565-576; 2015

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Guia de Vigilância Epidemiológica. Ministério da Saúde, Secretaria de Vigilância em Saúde. 6. ed. – Brasília: Ministério da Saúde, 2005.ISBN 85-334-1047-6.

CASTRO, L. N., FERRARI, D. G. Introdução à Mineração de Dados. Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CHAN, J. An architecture for Big Data analytics. Communications of the IIMA, v. 13, n. 2, p. 1, nov. 2013.

CHIAVEGATTO FILHO, Alexandre Dias Porto. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol. Serv. Saúde, Brasília , v. 24, n. 2, p. 325-332, June 2015

COAKLEY, M. F. et al. Unlocking the Power of Big Data at the National Institutes of Health. Big Data, v. 1, n. 3, p. 4, 6 jun. 2013.

DHAR, V. Data science and prediction. Comm. ACM 56(12): 64–73. 2013.

DUARTE, Cristina Maria Rabelais et al. Regionalização e desenvolvimento humano: uma proposta de tipologia de Regiões de Saúde no Brasil. Cad. Saúde Pública, Rio de Janeiro , v. 31, n. 6, p. 1163-1174, June 2015.

FLACH, Peter; LEARNING, Machine. The Art and Science of Algorithms that Make Sense of Data. 2012.

GARRISON, L. P. Universal Health Coverage - Big Thinking versus Big Data. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, v. 16, n. 1 Suppl, p. S1–3, 2013.

GKOULALAS-DIVANIS, Aris; LABBI, Abderrahim. Large-scale data analytics. Springer, 2014.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor. An Introduction to Statistical Learning: With Applications in R. 2014.

LANDER, Jared P. R for everyone: advanced analytics and graphics. Pearson Education, 2014.

MACDONALD, C. Using big data to improve health: geo-medicine combines pollution and health data to better inform patients, doctors and researchers. november, p. 14–15, 2012.

MATTMANN, CA. A vision for data science: to get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers. Nature, Jan 24, 2013, Vol.493 (7433), p.473(3).

MATSUDAIRA, KATE. The Science of Managing Data Science. Association for Computing Machinery. Communications of the ACM, Jun 2015, Vol.58(6), p.44

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Elsevier Brasil, 2014.

MCKINSEY GLOBAL INSTITUTE. Big Data: The Next Frontier for Innovation, Competition, and Productivity. Technical Report, June 2011.

MEDRONHO R; BLOCH KV; Luiz RR; Werneck GL (eds.). Epidemiologia. Atheneu, São Paulo, 2009, 2ª Edição.

MELLO JORGE MHP, LAURENTI R, GOTLIEB SLD. Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e SINASC. Ciência & Saúde Coletiva 2007; 12(3): 643-654.

MENEZES, N. N. C. Introdução à programação com Python – 2ª edição: Algoritmos e lógica de programação para iniciantes. Novatec Editora, 2016.

SILVA, Leandro Nunes de Castro; FERRARI, Daniel Gomes. Introdução à Mineração de Dados. Editora Saraiva, 2016.

VIANNA, R.C.X.F. et al. Mineração de dados e características da mortalidade infantil. Cad. Saúde Pública [online]. 2010, vol.26, n.3, pp.535-542. ISSN 0102-311X. http://dx.doi.org/10.1590/S0102-311X2010000300011.

VIPIN KUMAR, MICHAEL STEINBACH, PANG-NING TAN. Introdução ao DATAMINING Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

BIBLIOGRAFIA COMPLEMENTAR *(opcional)*

CRITÉRIOS DE AVALIAÇÃO
1) Fichamento de 4 **textos de referência** (2,0 pontos, sendo 0,5 por texto); **2)** Resolução de **exercícios práticos** (EP) (4,0 pontos, sendo 1,0 por curso finalizado (até 3 cursos) e 1,00 se finalizar projeto na plataforma DataCamp), sugestão cursos e projetos: **Cursos em Python:** "Intro to Python for Data Science", "Intermediate Python for Data Science", "Supervised Learning with scikit-learn", "Unsupervised Learning in Python", "Introduction to Data Visualization with Python". **Cursos em R:** "Introduction to R", "Intermediate R", "Machine Learning Toolbox", "Data Scientist with R", "Introduction to Machine Learning ", Machine Learning with R" e "Data Visualization with ggplot2". **Projeto em Python ou R**: "Introduction to DataCamp Projects" **3)** Participar de **competição de machine learning** na plataforma Kaggle in Class (https://inclass.kaggle.com/c/icict-fiocruz-taxa-mort-brasil) (2,0 pontos para 1º colocado e 1,5 ponto para cada participante com *score* acima do mínimo estabelecido); **4)** Entrega de **trabalho final** utilizando compartilhado por meio da ferramenta Jupyter Notebook ou Colaboratory Google (2,0 pontos).

CRONOGRAMA	
Data	**Conteúdo**
(1) 21/03	Apresentação da disciplina e dos professores responsáveis **Introdução à Ciência de Dados aplicada à Saúde** **Prof. responsável**: Marcel Pedroso e professores convidados
(2) 28/03	**Sistemas de Informação em Saúde (SIS)** - Principais fontes de dados de saúde e de interesse para saúde **Introdução a Epidemiologia e Demografia** - Conceitos principais - Principais medidas e métricas - Exemplos de uso Textos de referência: Jannuzzi, PM. Indicadores Sociais no Brasil: conceitos, fontes de dados e aplicações - 1a edição. 1. ed. Campinas: Editora Alínea/PUC-Campinas, 2001. v. 1. 141p. Costa, AJL; Kale, PL. Medidas de frequência de doenças. In: Medronho, RA et al. Epidemiologia. São Paulo: Editora Atheneu, 2009, p. 13-30.

		Costa, AJL; Kale, PL; Vermelho, LL. Indicadores de Saúde. In: Medronho, RA et al. Epidemiologia. São Paulo: Editora Atheneu, 2009, p. 31-82. **Prof. Responsável**: Ricardo Dantas
(3)	04/04	**Introdução ao R** Texto de referência: LANDER, Jared P. R for everyone: advanced analytics and graphics. Pearson Education, 2014. **Prof. responsável**: Eduardo Ogasawara
(4)	11/04	**Aplicações de Ciência de Dados à Saúde** Textos de referência: IANNA, Rossana Cristina Xavier Ferreira et al. Mineração de dados e características da mortalidade infantil. Cad. Saúde Pública [online]. 2010, vol.26, n.3, pp.535-542. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2010000300011&lng=en&nrm=iso David Karlsson, Joakim Ekberg, Armin Spreco, Henrik Eriksson, Toomas Timpka. Visualization of Infectious Disease Outbreaks in Routine. Series Studies in Health Technology and Informatics, 2013. **Prof. responsável:** Christovam Barcellos
(5)	18/04	**Introdução ao Python** Texto de referência: Menezes, N. N. C. Introdução à programação com Python – 2ª edição: Algoritmos e lógica de programação para iniciantes. Novatec Editora, 2016. **Prof. responsável:** Jefferson Lima
(6)	25/04	**Introdução a análise estatística de dados** - Tipos de dados e atributos, estatísticas descritivas, medidas de posição e dispersão Texto de referência: PINHEIRO RS; TORRES TZG. Análise Exploratória de Dados. Seção III, Capítulo 18, p. 323-42. In: MEDRONHO R; BLOCH KV; Luiz RR; Werneck GL (eds.). Epidemiologia. Atheneu, São Paulo, 2009, 2ª Edição. Texto de referência: LUIZ RR. Associação Estatística em Epidemiologia: Análise Bivariada. Seção III, Capítulo 24, p. 429-456. In: MEDRONHO R; BLOCH KV; Luiz RR; Werneck GL (eds.). Epidemiologia. Atheneu, São Paulo, 2009, 2ª Edição. **Prof. responsável**: Paulo Borges
(7)	02/05	**Introdução a mineração de dados e machine learning** Texto de referência: CASTRO, L. N., FERRARI, D. G. Introdução à Mineração de Dados. Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016. Páginas 1 - 25 **Prof. responsável**: Marcel Pedroso

(8) 09/05	**Laboratório em Ciência de Dados aplicada à Saúde** Uso livre do laboratório para atividades da disciplina
(9) 16/05	**Regressão aplicada à inferência estatística** Texto de referência: LUIZ RR. Associação Estatística em Epidemiologia: Análise Bivariada. Seção III, Capítulo 24, p. 429-56. LUIZ RR. Associação Estatística em Epidemiologia: Análise com Múltiplas Variáveis. Seção III, Capítulo 25, p. 457-84 In: MEDRONHO R; BLOCH KV; Luiz RR; Werneck GL (eds.). Epidemiologia. Atheneu, São Paulo, 2009, 2ª Edição. **Prof. responsável**: Paulo Borges
(10) 23/05	**Regressão aplicada à análise preditiva** - Testes de validação e principais métricas de desempenho dos modelos de predição Texto de referência: JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor. An Introduction to Statistical Learning: With Applications in R. 2014. ("Linear Regression and Classification" Páginas 59 – 137). http://www-bcf.usc.edu/~gareth/ISL/ **Prof. responsável**: Alexandre Chiavegatto
(11) 30/05	**Machine Learning (1) com R – Aprendizagem Supervisionada** - Tarefas de classificação - Kaggle **Prof. responsável:** Raphael Saldanha / Vanderlei Pascoal
(12) 06/06	**Machine Learning (2) com Python – Aprendizagem não Supervisionada** **- Redução de dimensionalidade e análise de agrupamento** Texto de referência: CASTRO, L. N., FERRARI, D. G. Introdução à Mineração de Dados. Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016. DUARTE, CMR et al. Regionalização e desenvolvimento humano: uma proposta de tipologia de Regiões de Saúde no Brasil. Cad. Saúde Pública, Rio de Janeiro, v. 31, n. 6, p. 1163-1174, June 2015. http://www.scielo.br/scielo.php?pid=S0102-311X2015000601163&script=sci_abstract&tlng=pt **Prof. responsável**: Marcel Pedroso
(13) 13/06 No LNCC?	**Machine Learning (3) com Python - Árvores de decisão** Texto de referência: FLACH, Peter. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. 2012. Páginas 129 – 156 e 331 – 333 **Prof. responsável**: Fábio Porto

(14) 20/06	**Visualização de dados (Kibana)** Texto de referência: **Prof. responsável:** Raphael Saldanha / Jefferson lima
(15) 27/06 No LNCC?	**Modelagem e aplicações em Big Data** Texto de referência: GKOULALAS-DIVANIS, Aris; LABBI, Abderrahim. Large-scale data analytics. Springer, 2014. **Prof. responsável**: Fábio Porto
(16) 04/07 **Sala 209**	Apresentação (em grupos) do projeto/proposta de trabalho final utilizando a Plataforma de Ciência de Dados aplicada à Saúde **Prof. responsável**: Marcel, Christovam, Jefferson, Raphael e Pascoal
(17) 11/07	Apresentação (em grupos) do projeto/proposta de trabalho final utilizando a Plataforma de Ciência de Dados aplicada à Saúde **Prof. responsável**: Marcel, Christovam, Jefferson, Raphael e Pascoal

Rio de Janeiro, 07 de março de 2018.

Linha 1: "Produção, Organização e Uso da Informação em Saúde"

Dedica-se à análise das políticas, modelos, processos e práticas de produção, organização, avaliação e uso da informação e do conhecimento no campo da saúde coletiva. A partir de múltiplas perspectivas teórico-metodológicas, prioriza-se o estudo de:

1.1. regimes de produção, regulação e novas dinâmicas de pesquisa científica em saúde;

1.2. inquéritos e pesquisas nacionais de saúde;

1.3. repositórios, ambientes virtuais, redes sociais e sistemas de informação;

1.4. práticas culturais, técnicas e tecnologias;

1.5. linguagens, padrões e indicadores;

1.6. prospecção e estudos métricos em ciência e tecnologia;

1.7. adequação de métodos que utilizem informações dos sistemas nacionais de informação para avaliar situações de saúde;

1.8. sistematização e análise das informações para a formulação de políticas públicas e monitoramento da situação de saúde brasileira e seus determinantes socioambientais.

Linha 2: "Informação, Comunicação e Mediações"

Tomando o direito à comunicação como inerente ao direito à saúde, estuda as relações entre instituições, profissionais de saúde e de comunicação e a população, em suas diversas formas de organização, em seus processos de produção, circulação e apropriação dos sentidos sociais. Dedica-se à discussão conceitual e ao desenvolvimento de metodologias que levem à melhor compreensão da natureza e das características das mediações culturais, sociais, políticas, institucionais e tecnológicas envolvidas em tais processos. Seus projetos priorizam:

2.1. a análise de produtos, práticas, processos e sistemas de comunicação, bem como de políticas públicas nesses domínios;

2.2. o estudo das relações entre mídia e saúde, em suas múltiplas formas discursivas;

2.3. a análise sobre a produção de sentidos nos novos espaços e ambientes de comunicação, com ênfase nos que se desenvolvem a partir de tecnologias virtuais;

2.4. estudos que evidenciem e ampliem a compreensão do lugar da comunicação nos processos sociais e nas relações de poder na sociedade, bem como a relação entre comunicação e produção das desigualdades sociais em saúde.