INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Bahia

Pró-Reitoria de Pesquisa,
Pós-Graduação e Inovação

INSTITUTO FEDERAL DA BAHIA
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

ÁREA DO CONHECIMENTO: **(X) EXATAS** ( )HUMANAS ( )BIOLÓGICAS

PROGRAMA: ( ) PIBIC **(X) PIVIC**

Título do Projeto: Estudo e Aplicação de Big Data e Machine Learning em Ciência de Dados.

Nome do Grupo de Pesquisa Cadastrado no Diretório de Grupos de Pesquisa do CNPq: SSMART Research Group.

Orientador: Lauro Cássio Martins de Paula.

Unidade Acadêmica/Departamento: Departamento de Pesquisa Científica / Instituto Federal da Bahia – Campus Santo Antônio de Jesus.

PLANO DE TRABALHO

Edital 06/2017 – Fluxo Contínuo

Período 2019/2020

Título do Plano de Trabalho: Programação em Python para Simulação de Modelos Epidemiológicos.

Aluno: Felipe Xavier Passos
Matrícula: 20191TADSSAJ0003 – Aluno do curso TADS do IFBA-SAJ.

1. Introdução

A epidemiologia é uma área de estudo da medicina que investiga os fatores que determinam a frequência e a distribuição de diferentes tipos de doenças em grupos de pessoas. Por exemplo, tal investigação possibilita o estudo, o planejamento e a tomada de decisões para a prevenção de proliferação de doenças contagiosas. Nesse sentido, a Epidemiologia Computacional (EC) consiste no desenvolvimento e uso de modelos computacionais para compreender a proliferação de doenças do ponto de vista dinâmico. A EC utiliza técnicas da ciência da computação, matemática, geografia e saúde pública para desenvolver ferramentas e modelos para auxiliar epidemiologistas no estudo da propagação de enfermidades. Seu objetivo consiste em estudar como as doenças se espalham, mas não a doença propriamente dita (ANDERSON e MAY, 1992).

Utilizando modelos matemáticos e computacionais, torna-se possível simular o comportamento de uma determinada epidemia e seus efeitos na população de uma região específica. Com isso, pode-se desenvolver estratégias de controle de certas enfermidades. Um dos modelos mais tradicionais na EC consiste na classificação de indivíduos como *suscetíveis*, *infectados* e *recuperados*: SIR. O modelo SIR considera a distribuição homogênea de indivíduos no espaço e no tempo. Entretanto, ele não é capaz de explicar a persistência ou a erradicação de doenças infecciosas. O modelo SIR é utilizado para estudar a propagação de doenças infecciosas por meio do rastreamento de três fatores diferentes: um número (**S**) de pessoas suscetíveis à doença; um número (**I**) de pessoas infectadas com a doença; e um número (**R**) de pessoas que se recuperaram de tal doença após terem sido infectadas ou que faleceram por causa da doença. No SIR, supõe-se que a população total é fixa: N = S(t) + I(t) + R(t), onde t representa um certo período de tempo. A equação (1) mostra a fórmula utilizada para o modelo SIR, onde ∂ é a derivada parcial em relação ao tempo:

$$\frac{\partial N}{\partial t} = \frac{\partial S}{\partial t} + \frac{\partial \mathrm{I}}{\partial t} + \frac{\partial \mathrm{R}}{\partial \mathrm{t}}, \qquad \forall t > 0.$$

Um outro modelo bastante utilizado é baseado no indivíduo (MBI). O MBI é uma estratégia baseada no modelo SIR que utiliza fundamentos estocásticos (não-determinísticos) para modelar alterações nas características do indivíduos. Alguns trabalhos da literatura demonstram que o MBI tende a possuir um tempo computacional reduzido e que a taxa de crescimento com o aumento do tamanho da população é consideravelmente menor quando comparado ao modelo SIR (PEREIRA, 2008; FILHO et al., 2011). No MBI, cada indivíduo da população pode ser descrito por um conjunto de características relevantes do ponto de vista epidemiológico. Por exemplo, essas características podem ser consideradas como: idade, localização espacial e

tempo de infecção. Tais detalhes podem ser incluídos nos processos de natalidade, contágio e recuperação/morte durante a simulação do modelo. A equação (2) mostra que o *d*-ésimo indivíduo da população é associado a um vetor de *n* características, onde *I* é um indivíduo, cada *C* é uma característica diferente, e *t* é o instante de tempo considerado:

$$I_d(t) = [C_{d1}(t)\ C_{d2}(t) \ldots\ C_{dn}(t)]^T.$$

Em comparação com modelos determinísticos, modelos estocásticos como o baseado em indivíduos oferecem uma maior flexibilidade para a incorporação de características epidemiológicas específicas de uma determinada população. No entanto, a simulação desses modelos normalmente implica em um custo computacional elevado quando aplicados em grandes conjuntos de dados (Big Data). Big Data é um termo muito empregado na ciência de dados que consiste na análise e interpretação de grandes volumes de dados com vasta variedade. A ciência de dados é uma área de estudo que vai desde à concepção e obtenção dos dados até à visualização das informações extraídas desses dados. Um dos principais objetivos consiste em extrair insights significativos dos dados e auxiliar na tomada de decisões por meio de análises preditivas. Portanto, o estudo, a implementação, simulação e comparação de modelos epidemiológicos que envolvem grandes volumes de dados compreendem uma tarefa importante a ser cientificamente investigada.

2. Objetivos

Pesquisar e selecionar modelos epidemiológicos (possivelmente os modelos SIR e MBI) para a implementação na linguagem de programação Python.

Simular e comparar tais modelos utilizando pelo menos um grande conjunto de dados epidemiológicos e, se possível, aplicar alguma técnica de análise preditiva nos resultados da simulação para extrair informações que auxiliem na tomada de decisões;

Obter resultados consideráveis e realizar comparações com trabalhos da literatura que compõem o estado da arte, demonstrando uma certa superioridade na qualidade dos resultados obtidos tanto em termos computacionais quanto epidemiológicos.

Publicar e apresentar tais resultados em forma de artigo científico em pelo menos uma conferência nacional na área de saúde pública, além de publicar uma versão estendida do artigo em um periódico científico internacional;

Apresentar à comunidade acadêmica local os principais resultados obtidos com o objetivo de demonstrar a importância da iniciação científica na formação acadêmica do aluno.

3. Metodologia

A princípio, modelos epidemiológicos em destaque na literatura serão pesquisados e investigados. Artigos científicos, livros, teses de doutorado, dissertações de mestrado e monografias serão as principais fontes para a pesquisa científica a ser realizada neste trabalho. Por ser considerada uma linguagem robusta e bastante utilizada por cientistas de dados, a linguagem de programação Python também será estudada e utilizada para a implementação dos modelos.

4. Etapas e Cronograma de Execução

As etapas do trabalho do aluno e seu cronograma estão resumidos na tabela abaixo.

Atividade \ Período	2019					2020						
	A	S	O	N	D	J	F	M	A	M	J	J
Revisão Bibliográfica.	x	x	x	x	x							
Estudo e seleção de modelos epidemiológicos.	x	x	x	x	x							
Estudo dos algoritmos dos modelos epidemiológicos.				x	x	x	x	x				
Implementação dos modelos na linguagem Python.					x	x	x	x	x			
Pesquisa e obtenção de um conjunto de dados para simulação dos modelos						x	x	x	x	x		
Redação do Relatório de Acompanhamento.							x					
Comparação dos modelos							x	x	x	x	x	
Redação do Relatório Final										x	x	x
Finalização do trabalho e escrita de artigo a ser submetido em algum congresso.												x

5. Resultados Esperados na Execução

Os resultados deste trabalho de iniciação científica compreendem:

- obtenção de resultados significativos por meio da simulação de modelos epidemiológicos implementados em linguagem computacional adequada para lidar com grandes volumes de dados;

- a iniciação do aluno na pesquisa e na publicação de trabalhos em congressos e periódicos com o rigor acadêmico-científico apropriado;

- uma possível extensão do tema para o trabalho de conclusão de curso do aluno e uma preparação do mesmo para ingressar num futuro mestrado na mesma área de pesquisa.

6. Referências Bibliográficas

ANDERSON, R. M.; MAY, R. M. **Infectious diseases of humans: Dynamics and control**. Oxford University Press, 1992.

FILHO, A. R. G.; ARRUDA, F. D. B.; GALVAO, R. K. H.; YONEYAMA, T. **Programação paralela cuda para simulação de modelos epidemiológicos baseados em indivíduos**. Simpósio Brasileiro de Automação Inteligente (SBAI), 2011.

PEREIRA, E. B. **Modelos baseados em indivíduos para análise e controle de epidemias em populações heterogêneas e metapopulações**. Dissertação de Mestrado — Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, 2008.

PAULA, L. C. M. **O Hype da Ciência de Dados**. Texto publicado em blog. Disponível em < https://www.linkedin.com/pulse/o-hype-da-ci%C3%AAncia-de-dados-lauro-c-martins-de-paula/>. Acesso em 21 ago. 2019.

SILVA, D. **Ciência de Dados Aplicada na Educação**. Anais do XII Evento de Iniciação Científica, vol. 3, n. 1. UniBrasil Centro Universitário, 2017.