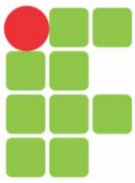

INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
BAHIA

ESTUDO E APLICAÇÃO DE BIG DATA E MACHINE LEARNING EM CIÊNCIA DE DADOS

Equipe executora:

Lauro Cássio Martins de Paula

SANTO ANTÔNIO DE JESUS – BA

2019

PROJETO: Estudo e Aplicação de Big Data e Machine Learning em Ciência de Dados

INTRODUÇÃO

Nos últimos anos, a Ciência de Dados vem se destacando e chamando cada vez mais a atenção tanto no Brasil quanto no mundo. Um dos principais motivos é devido ao fato do crescimento exponencial do volume de dados gerados diariamente no mundo todo. Por exemplo, empresas de tecnologia como Google, Facebook e Whatsapp têm tido um aumento no volume de dados gerados na escala de milhões por segundo (PAULA, 2017). Enquanto o Google teve um aumento de 2,4 milhões para 3,8 milhões de pesquisas em sua máquina de busca num intervalo de 60 segundos, o Whatsapp teve um aumento de 12,5 milhões para 29 milhões de mensagens enviadas a cada minuto, em 2016. Esse último caso representa um aumento de 43% no volume de dados gerados.

Nesse contexto, a Ciência de Dados surge como 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 é extrair insights dos dados e auxiliar na tomada de decisões por meio de análises descritivas e, principalmente, preditivas. Para isso, torna-se necessária a união da Ciência de Dados com outras áreas de mesma importância: Engenharia de Big Data, Machine Learning, Business Analytics e Visualização de Dados (Story Telling).

Big Data consiste na análise e interpretação de grandes volumes de dados com vasta variedade. Soluções específicas para Big Data, que permitam os profissionais de TI trabalhar com informações não-estruturadas, tornam-se necessárias. A Engenharia de Big Data é a área responsável por toda a parte de infraestrutura. Ela é fundamental para o armazenamento e obtenção dos dados e para a realização das análises pelo analista de dados.

A aprendizagem de máquina (Machine Learning) é uma área de estudo da computação que utiliza a matemática e a estatística para desenvolver programas de computador capazes de aprender sozinhos (MILLER, 2014). Utilizando um bom modelo de treinamento, o software torna-se capaz de realizar predições com base no conhecimento acumulado. Diversas empresas têm utilizado Machine Learning para aprender as preferências de seus clientes. Por exemplo, a Netflix faz recomendações de filmes e séries com base em suas avaliações por parte dos usuários.

Nesse sentido, percebe-se que a Ciência de Dados deixou de se restringir apenas a uma aplicação teórica e tem sido comumente aplicada em diferentes áreas, tanto acadêmicas quanto comerciais. Com isso, a sua utilização está difundida e implícita no nosso cotidiano de tal forma que a implementação de algoritmos eficientes para analisar e extrair informações relevantes dos dados torna-se cada vez mais necessária. Portanto, este projeto visa estimular o desenvolvimento de pesquisas científicas com o intuito de implementar novos algoritmos de Machine Learning para automatizar o processo de análise (descritiva e preditiva) de grandes volumes de dados em geral e extrair insights significativos que auxiliam na tomada de decisões.

JUSTIFICATIVA

Considerando o fato de que o mercado de trabalho, especialmente na área de tecnologia, tem necessitado cada vez mais de profissionais competentes e capacitados na utilização de ferramentas computacionais para a análise de dados, torna-se necessário um estudo eficiente e adequado para o aprimoramento constante dessas ferramentas.

O pesquisador que apresenta este projeto possui experiência acadêmica significativa na área de computação e pretende estimular o estudo da Ciência de Dados por parte dos alunos interessados em realizar iniciação científica e trabalhos de conclusão de curso nesse tema. Com esta proposta, objetiva-se propiciar um espaço de discussão e aprofundamento do tema por meio de seminários e apresentações semanais no nosso grupo de pesquisa (http://dgp.cnpq.br/dgp/espelhogrupo/461437). Adicionalmente, este projeto propõe-se a iniciar o aluno no universo da pesquisa científica com o rigor acadêmico apropriado, visando à preparação técnico-científica do aluno tanto para a carreira acadêmica quanto para o mercado de trabalho.

OBJETIVOS

O objetivo geral deste trabalho consiste no estudo, na implementação e aplicação de técnicas de Machine Learning e Big Data para a análise de diferentes volumes de dados provenientes do mundo real. Dentre os objetivos específicos, destacam-se:

- Aplicação das principais etapas e ferramentas em Ciência de Dados desde à obtenção, limpeza, preparação e análise dos dados até às conclusões ou formulação de hipóteses a partir dos mesmos (storytelling);
- Implementação de algoritmos a partir de técnicas de Machine Learning para automatizar o processo de análise de dados, aumentando a eficiência computacional;

METODOLOGIA

Para o desenvolvimento deste projeto de pesquisa, as seguintes etapas poderão ser aplicadas:

- Revisão bibliográfica dos livros e trabalhos científicos que compõem o estado da arte em Machine Learning;
- Estudo dos principais trabalhos científicos que descrevem técnicas de Big Data e Machine Learning;
- Elaboração de planos de trabalho para alunos de iniciação científica e de trabalho de conclusão de curso a partir deste projeto de pesquisa;
- Documentação e arquivamento de todas as ferramentas utilizadas, todos os dados gerados e resultados obtidos para utilização futura;
- Escrita e publicação de artigos científicos em conferências e periódicos relacionados ao tema deste projeto;
- Divulgação e apresentação à comunidade acadêmica local/regional de todos os resultados obtidos ao final de cada plano de trabalho.

VIABILIDADE TÉCNICA

Para a execução deste projeto, estão disponíveis os laboratórios de informática do campus com diferentes softwares livres. Caso seja necessária a utilização de algum software comercial, os trâmites legais serão providenciados para a aquisição e instalação.

REFERÊNCIAS

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

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 25 abr. 2019.

MILLER, S. **Collaborative Approaches Needed to Close the Big Data Skills Gap**. Journal of Organization Design, vol. 3, n. 1, 2014.

DARROW, B. **Data Science is Still White Hot, But Nothing Lasts Forever**. Fortune. Disponível em < http://fortune.com/2015/05/21/data-science-white-hot/>. Acesso em 25 abr. 2019.

DE MAURO, **A. Human Resources for Big Data Professions: A Systematic Classification of Job Roles and Required Skills Sets**. Information Processing and Management, 2017.