ANÁLISE E VISUALIZAÇÃO DE DADOS COM DATA SCIENCE E OPEN DATA

Guilherme de Cleva Farto

guilherme.farto@gmail.com

João Guilherme Fernandes

Joaoguilherme89@hotmail.com.br

RESUMO: Com o grande crescimento da automatização da área da indústria, educação e agricultura a quantidade de dados é cada vez maior com o decorrer do tempo. A análise de dados virou o pote de ouro para todos os setores comercias e científicos, pois com ela, é possível descobrir padrões positivos e negativos e com isso auxiliar seus utilizadores. Esta pesquisa visa explorar os contextos de análise e visualização de dados com a utilização de ferramentas open-source juntamente com data sets abertos (Open data).

PALAVRAS-CHAVE: Data Science, Analise de dados, Open data, Visualização de dados, Python.

ABSTRACT: With the rapid growth of automation in industry, education and agriculture, the amount of data is increasing over time. A data analysis has turned or pot of gold for all commercial and scientific sectors, because with it it is possible to find positive and negative patterns and with these aids used. This research aims to explore data analysis and response contexts using open source tools and open data sets.

KEYWORDS: Data Science, Data Analysis, Open Data, Data Visualization, Python.

1. INTRODUÇÃO

Com a indústria 4.0, a quantidade de dados gerados é cada vez maior. De acordo com uma pesquisa desenvolvida pela EMC [1], o volume de dados saltará de 130 exabytes para 40.000, ou 40 trilhões de gigabytes até 2020; após está data, a tendência é que o volume de dados dobre a cada dois anos. Este grande aglomerado de dados recebeu o nome de *BIG DATA.*

Desta forma, o termo Data Science veio a tona. Ciência de dados ou, Data Science, envolve o uso de métodos para analisar grandes quantidades de dados (Big Data) e extrair o conhecimento que eles contém (CIELEN, 2016). Muitas empresas ao redor do globo perceberam isso e começaram a realizar a análise de dados, de modo que poderiam tomar decisões com mais segurança usando gráficos e relatórios.

Para realizar a análise de dados, é necessário o uso de algumas ferramentas de desenvolvimento como, por exemplo, a linguagem *Python.* Embora Python não tenha um conjunto de pacotes e bibliotecas tão abrangente quanto os disponíveis para a linguagem R, a combinação de *Python* com ferramentas como *Pandas*, *Numpy, Scipy, Scikit-learn e Seaborn* torna a linguagem uma das principais escolhas entre os Cientistas de Dados. A linguagem *Python* também está lentamente se tornando útil para tarefas em Machine Learning e de base para o trabalho estatístico intermediário (anteriormente apenas sob o domínio de R) (MATOS, 2017).

Precisa-se, também, de um conjunto de dados para que seja possível realizar a análise exploratória. Na internet é possível encontrar uma grande diversidade de data sets abertos ao público. Dados abertos (Open data) são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa (ISOTANI, 2015). O governo federal brasileiro, por exemplo, possui grande gama de data sets abertos, disponíveis em http://dados.gov.br/.

2. DATA SCIENCE

Data Science surgiu com o crescimento de dados gerados a partir da indústria 4.0 e com o surgimento do *Big Data.* A indústria percebeu que com grandes volumes de dados, era possível realizar estudos sobre eles e sendo assim obter conclusões e idéias para os negócios, podendo diminuir perdas e aumentar o rendimento de diversos setores.Nos dias de hoje diversas áreas fazem o uso da ciência de dados para obter melhores rendimentos como: Educação, Agricultura e Saúde.Um exemplo de uso da ciência de dados na Educação é a Educação 2.0, com ela é possível entender em quais pontos de uma matéria um aluno está tendo mais dificuldades e proporcionar atividades focadas nisso. (HEKIMA, 2016)

2.1.OPENDATA

Muitas organizações e setores industrias utilizam ou realizam a distribuição de dados de forma aberta para que haja interoperabilidade entre os diversos setores da indústria. Para que um conjunto de dados seja discriminado como “aberto” é necessário que ele siga algumas obrigações: Disponibilidade e Acesso, Reutilização e Redistribuição e Participação Universal, ou seja, para que um dado seja “aberto” ele deve poder ser livremente usado, reutilizado e redistribuído por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras. (OPEN DATA HANDBOOK, 2018)

2.2VISUALIZAÇÃO DE DADOS

A visualização de dados é muito importante durante o processo de análise e o processamento de dados, pois é nessa etapa que acaba ficando visível as conclusões e ideias que estão sendo extraídas do grande volume de dados analisado, ou seja, a visualização de dados consiste na representação gráfica de informações e dados. (TABLEAU, 2018) A visualização de dados é realizada de diversas formas como a representação dos dados em gráficos,

diagramas, infográficos, tabelas e painéis. A visualização é realizada desta forma para facilitar a compreensão do usuário.

3 . PROPOSTA DE DESENVOLVIMENTO DO TRABALHO

3.1. OBJETIVOS

O objetivo geral desta pesquisa é realizar e demonstrar o estudo sobre data Science utilizando open data. Com isso, pretende-se incentivar e aprimorar pesquisas relacionadas à big data, data Science e open data com a ajuda de ferramentas como Python e suas bibliotecas para a análise de dados.Como resultados, espera-se contribuir com o material teórico sobre ANÁLISE E VISUALIZAÇÃO DE DADOS COM DATA SCIENCE E OPEN DATA para a comunidade científica de forma que sirva como base para futuras pesquisas.

3.2. TECNOLOGIAS E RECURSOS

3.2.1. PYTHON

Python é uma linguagem dinâmica, interpretada, robusta, multiplataforma, multiparadigma (orientação à objetos, funcional, refletiva e imperativa) e está preparada para rodar em JVM e .NET Framework. Lançada em 1991 por Guido van Rossum, é uma linguagem livre (até para projetos comerciais) e hoje pode-se programar para desktops, web e mobile (DEVMEDIA, 2016). Com tudo, a linguagem será usada para o processamento, análise e visualização dos dados juntamente com um conjunto de bibliotecas específicas para os procedimentos citados.

3.2.2. PANDAS

Pandas é uma biblioteca de código livre feita para a linguagem de programação *Python*. Esta biblioteca é utilizada para realizar operações de análise e processamento de dados, sendo para muitos, a espinha dorsal da maioria dos projetos de ciência de dados e machine Learning.

3.2.3. MATPLOTLIB

Matplotlib é uma biblioteca da linguagem python que tem como o objetivo realizar plotagens 2D. Com ela é possível realizar a construção de gráficos, histogramas, espectros de potência, gráficos de barras, gráficos de erros, gráficos de dispersão etc. com apenas algumas linhas de código. (MATPLOTLIB, 2012). Com esta biblioteca será possível a criação dos gráficos após o processo de análise e processamento de dados juntamente com a linguagem *python.*

3.3. JUPYTER NOTEBOOK

Jupyter Notebook é um ambiente web e interativo destinado para a criação de documentos “Notebooks”. Esse ambiente é construído utilizando somente tecnologias open-source. Além do Python, pode conectar-se a linguagens como o R, Julia, Ruby, Scala e Haskell. Atualmente, são suportadas mais de 40 linguagens de programação.

Essa ferramenta será utilizada para a construção do projeto, onde será inserido os códigos necessários para a análise, processamento e visualização dos dados.

3.4. TESTES E RESULTADOS

Com a utilização dos recursos e ferramentas citadas anteriormente, foram realizados diversos testes no contexto proposto. A realização de análise e visualização de dados obteve sucesso, juntamente com a utilização de tecnologias open source e dados abertos. Com isso, obteve uma análise capaz de gerar insights através de gráficos, mapas e histogramas.

3.5. TRABALHOS FUTUROS

Pretende-se desenvolver como trabalho futuro, criar modelos de *machine learning* utilizando tecnologias e ferramentas *open source* juntamente com um conjunto de dados abertos.

4. CONCLUSÕES

Análise de dados é alvo de estudos de diversas áreas acadêmicas e comerciais. Suas vantagens são diversas e sempre quando bem feita pode apresentar insights de grande importância, e até mesmo mostrar e precaver fraudes presentes no conjunto de dados. Além disso pode mostrar caminhos para a resolução de problemas, diminuição de custos, apresentar padrões e auxilia em diversas áreas como: agricultura, educação e saúde.

Com isso, esse trabalho propôs um estudo sobre análise e visualização de dados com Data Science e Open Data, com a utilização de tecnologias e ferramentas open Source.

5. REFERÊNCIAS

DEVMEDIA. Aprendendo a programar em Python – Introdução. 2016. Disponível em< https://www.devmedia.com.br/aprendendo-a-programar-em-pythonintroducao/17093>, Acesso em 27/11/2019.

MATPLOTLIB. MATPLOTLIB: Python Plotting. 2012. Disponível em< https://www.devmedia.com.br/aprendendo-a-programar-em-pythonintroducao/17093>, Acesso em 27/11/2019.

TABLEAU. Guia prático da visualização de dados: definição, exemplos e recursos de aprendizado. Disponível em< https://www.tableau.com/pt-br/learn/articles/data-visualization>, Acesso em 27/11/2019.

tableau

CIELNE, D., MEYSMAN, A. D., & ALI, M. (2016). Introducing Data Science. Manning Publications.

EMC[1] Gantz, John and Reinsel. (2012). David. The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. EMC Corporation. Acesso em: jul. 2015. Disponível em: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

MATOS, D. (12 de MARÇO de 2017). R ou Python para Análise de Dados? Acesso em 12 de Novembro de 2018, disponível em Ciência e Dados: http://www.cienciaedados.com/r-ou-python-para-analise-de-dados/

PAIXÃO, A. d., SILVA, V. A., & TANAKA, A. (2015). De business intelligence a Data Science: um estudo comparativo entre áreas de conhecimento relacionadas.

ISOTANI, S. (2015). DADOS ABERTOS CONECTADOS. Novatec.