Bibliotecas Python Para Machine Learning

Pedro Inácio de Oliveira Filho, Eduardo Camilo Inacio

Faculdade Senai - SC

Tecnólogo Em Análise e Desenvolvimento De Sistemas

pedro_oliveira-fi1@estudante.sesisenai.org.br,
eduardo.inacio@edu.sc.senai.br

Resumo. *Este artigo traz um breve estudo sobre a aplicabilidade de bibliotecas Python para o desenvolvimento de Machine Learning (aprendizado da máquina) no âmbito da inteligência artificial. Partindo da hipótese de que as bibliotecas fornecem conteúdo necessário para que o programador otimize o seu trabalho, subsiste a necessidade em distinguir as bibliotecas adequadas para implementação de um sistema, programa ou análise de dados. Posto isso, o artigo almeja melhorar a compreensão a respeito da aplicabilidade de bibliotecas Python. A pesquisa feita para este artigo, concebe fundamentação e resultados experimentais para o desenvolvimento de trabalhos futuros nesta área. O desenvolvimento e referencial teórico ora apresentado, elucida sobre a importância das bibliotecas dentro das análises e manipulação de dados. A pesquisa pretende averiguar o grau de importância de cada uma das bibliotecas estabelecidas dentro do processo de desenvolvimento machine learning. Por se tratar de um apanhado de informações sobre um projeto ainda em andamento, não haverá pretensão em obter respostas conclusivas definitivas sobre o tema proposto.*

1. Introdução

A área de desenvolvimento Machine Learning é importante para o avanço tecnológico, visto que, o mercado está cada vez mais desenvolvido e a velocidade com que a tecnologia avança, requer rapidez e agilidade no desenvolvimento de novos produtos. Há décadas, existiam problemas no processo de inserção de dados, que era feito de forma lenta e ineficiente em consequência da falta de ferramentas que hoje nos auxiliam na programação, no desenvolvimento de sistemas e inteligência artificial. Tecnologias como reconhecimento de imagem e voz, dentre várias outras, só foram possíveis a partir da criação de métodos eficientes que otimizam o trabalho do programador. A evolução desses sistemas fez-se plausível, devido à criação de bibliotecas que hoje auxiliam no processo de programação. O Python e o desenvolvimento Machine Learning serão objeto de estudo desta pesquisa, que foi estruturada, a princípio, em análises curtas sobre o assunto abordado.. Neste contexto há um esforço para que o artigo faça um breve esboço sobre algumas das bibliotecas utilizadas em Python para Machine Learning. Sendo Machine Learning um dos mais

importantes campos de pesquisa e desenvolvimento para a inteligência artificial, concernente ao aprendizado da máquina. o principal objetivo do cientista é receber os dados, organizá-los, analisar, desenvolver e proporcionar assim, o aprendizado da máquina em torno da resolução de problemas cotidianos.

2. Bibliotecas e Machine learning

O Machine Learning (Aprendizado de máquina), termo originado do inglês, é o ramo da ciência da informação que desenvolve inteligência artificial. O desenvolvimento da inteligência artificial é gerenciado e acompanhado por cientistas de dados que inserem informações necessárias para que haja o aprendizado na máquina sem que ela seja programada.

A análise de dados é um processo que requer foco principalmente para que os dados sejam estruturados, uma vez que, há tipos de formatos diferentes e comuns. Nesta perspectiva, trabalha-se com dados tabulares ou do tipo planilha, em que se pode ter string, data, numérico ou outro tipo. Há também as matrizes, tabelas, colunas de chaves, séries temporais, etc. O conjunto de dados extraídos das diversas formas citadas anteriormente, pode ser estruturado para que se crie novas análises delimitadas num determinado assunto, por exemplo, machine learning para manutenção preditiva.

A linguagem python vem como forte proposta no que diz respeito à linguagens de programação. surgida em 1991, a linguagem phyton se tornou uma linguagem muito utilizada em todo o planeta nas últimas décadas. Python além de se tornar uma linguagem essencial para a análise de dados, também se tornou uma ferramenta importante em Machine Learning e desenvolvimento do mercado de softwares, além de contribuir significativamente com a área acadêmica.

Python é uma linguagem que possui facilidade em aglutinar-se e integrar-se com outros códigos, por isso, torna seu uso mais prático e viável. A necessidade de uma linguagem que seja mais completa, aglutinadora, e que permita o usuário utilizar um conjunto de ferramentas mais completo, faz com que a linguagem Python esteja em evidência na preferência dos profissionais, pesquisadores, engenheiros de software e outros profissionais da área de tecnologia.

Diante da necessidade de se inserir grande quantidade de dados, o programador que utiliza uma linguagem, seja ela Python, Java ou qualquer outra, dependerá de uma série de conhecimentos que podem tornar o seu trabalho demorado, cansativo e inviável. Sendo assim, nasce a necessidade de utilização das bibliotecas. Neste artigo, serão vistas as principais bibliotecas para Python, mas antes, deve-se definir o conceito básico sobre bibliotecas.

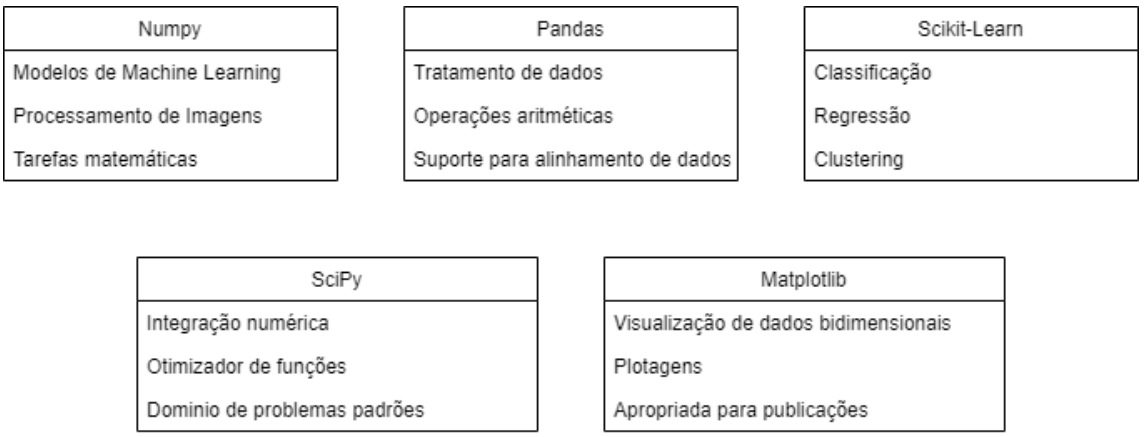
As bibliotecas foram criadas a partir da necessidade de otimizar o processo de inserção de dados na programação, facilitando assim, a criação de códigos e diminuindo eventuais repetições desnecessárias dentro da programação. O uso das bibliotecas

garante mais eficiência, agilidade, qualidade e menos erros ao programar, portanto, é sabido que ao iniciarmos o aprendizado em linguagens de programação, é de extrema importância que saibamos antes a importância das bibliotecas. O termo “biblioteca”, em inglês, denomina-se como “library”, refere-se à coleção de livros, ou melhor, conjunto de subprogramas (ou códigos) que são desenvolvidos para auxiliar no desenvolvimento de softwares. Os programas, códigos, aplicações, são predefinidos e aglutinados de forma organizada para que o usuário (ou programador) tenha acesso, por meio da pesquisa nas bibliotecas, ao melhor conteúdo ou pelo menos ao que o levará ao melhor resultado na busca de incrementos e soluções para o produto a ser criado. São inúmeras as bibliotecas que poderiam ser citadas a seguir, porém, limitamo-nos às mais conhecidas e utilizadas pelos programadores Python.

3. Contribuição/proposta

A linguagem Python possui, conforme levantamento realizado pelo Google, um acervo disponibilizado de aproximadamente 137 mil bibliotecas. Seria genérico poder dizer que cada uma delas tem uma finalidade, pois, dentro de cada biblioteca, encontram-se os mais variados tipos de programas prontos e códigos predefinidos para diversas finalidades. Ao pesquisar uma biblioteca, o programador realiza uma busca sistemática por determinado programa que atenda às necessidades e se encaixe no que está sendo desenvolvido. É interessante ressaltar que a busca por códigos nem sempre vai alcançar o resultado esperado, pois, cada item disponibilizado dentro de uma biblioteca é desenvolvido para determinados objetivos, contexto e finalidade. Assim, a pesquisa sistemática nas bibliotecas deve ser criteriosa quanto à necessidade real do que está sendo desenvolvido e a finalidade do que foi encontrado como complemento para o projeto. Antes de citar as bibliotecas, é importante destacarmos a plataforma de distribuição das linguagens python Anaconda (https://www.anaconda.com/products/distribution) que oferece mais de 8.000 pacotes de ciência de dados e aprendizado de máquina de código aberto, construídos e compilados pelo Anaconda para todos os principais sistemas operacionais e arquiteturas. A seguir, este artigo abordará alguns conceitos sobre bibliotecas utilizadas em Python. A figura 1, resume algumas das funcionalidades oferecidas em cada biblioteca:

Figura 1



2.1 Numpy

NumPy (Numerical Python) é a biblioteca da linguagem de programação Python que possui objetos denominados arrays (matrizes) multidimensionais e também está equipada com uma série de rotinas de processamento para esses arrays. Programadores utilizam essa biblioteca para executar cálculos numéricos em modelos de Machine Learning, processamento de imagem e computação gráfica e outras tarefas matemáticas.

De acordo com Wes McKinney, 2018,pag.22, A Numpy [...] tem sido a pedra angular do processo numérico em Python há muito tempo. Ela oferece o código aglutinador para as estruturas de dados, os algoritmos e a biblioteca necessários à maioria das aplicações científicas que envolvam dados numéricos.

Dentro dos modelos de Machine learning, Numpy fornece a possibilidade de se realizar diversos cálculos matemáticos como multiplicação, divisão entre outros. Além de ter uma ótima utilidade para cálculos, também armazena dados de treinamento em arrays, que facilitam no aprendizado e desenvolvimento em Machine Learning.

Wes McKinney, 2018, pag 22, ainda cita que, além dos recursos de processamento rápido de arrays que arrays que a Numpy acrescenta no Python, um de seus principais usos em análise de dados é como um contêiner para que dados sejam passados entre algoritmos e bibliotecas [...] Assim, muitas ferramentas de processamento numérico para Python supõem os arrays Numpy com estrutura de dados principal ou têm como meta uma interoperabilidade suave com a Numpy.

2.2 Pandas

Pandas (http://pandas.pydata.org/) é uma biblioteca python que oferece ferramentas eficientes, desenvolvidas para análise e manipulação de dados estruturados ou tabulares. Surgiu entre 2008 e 2010 e tem auxiliado produtivamente o Python na programação e análise de dados, tornando-o assim, mais fácil e rápido.

O pandas, segundo McKinney, 2018, pag 22, combina as ideias de processamento de alto desempenho de arrays da Numpy com os recursos flexíveis de manipulação de dados das planilhas e dos bancos de dados relacionais (como SQL). Ele disponibiliza uma funcionalidade sofisticada de indexação para facilitar a reformatação, a manipulação, as agregações e a seleção de subconjuntos de dados.

Em resumo, o Pandas se tornou ideal para a manipulação, leitura, agregação e visualização de dados, além de poder obter dados de diversos tipos de arquivos como: excel, Csv entre outros.

2.3 Scikit-Learn

Criada em 2010, (http://scikit-learn.org/stable/) se tornou uma biblioteca para codificação machine learning, de software livre, desenvolvida a partir de outras bibliotecas como Numpy, SciPy, Matplotlib, Pandas, etc. Segundo McKinney (2018, p. 27) O scikit-learn inclui submódulos para módulos como: Classificação,[...] regressão, [..] Clustering, [...] Redução de dimensionalidade, [...] Seleção de modelos, [...] pré-processamento. Scikit-learn, juntamente com pandas, IPython e o statsmodel, têm sido de extrema importância para que o Python seja hoje uma linguagem produtiva e necessária à ciência de dados, é nesse contexto que destacamos Scikit-Learning como uma das bibliotecas importantes para Python.

2.4 SciPy

O SciPy é uma junção de códigos para resolução de problemas no processo científico, nele encontramos ferramentas de processamento de sinais, solucionadores de matrizes e sistemas lineares, algoritmos para encontrar raízes, solucionador de equações entre outros. Utilizamos basicamente para soluções matemáticas e estatísticas.

2.4 IPython

O IPython é uma das mais importantes bibliotecas para Python, apesar de não oferecer nenhuma ferramenta de processamento ou análise de dados, possibilita o trabalho no ciclo execução-exploração, além de oferecer acesso fácil ao shell de seu sistema e também ao sistema de arquivos. Com a criação do projeto Jupyter, o IPython passou a ser mais produtivo e interativo, oferecendo novas possibilidades e suporte para outras linguagens de programação, portanto o IPython está integrado em grande parte dos trabalhos que envolvem execução, depuração e teste de código.

Após entender sobre a função básica das bibliotecas, foi feita a figura abaixo ilustrando a aplicabilidade das bibliotecas de acordo com as principais funções a que se destinam.

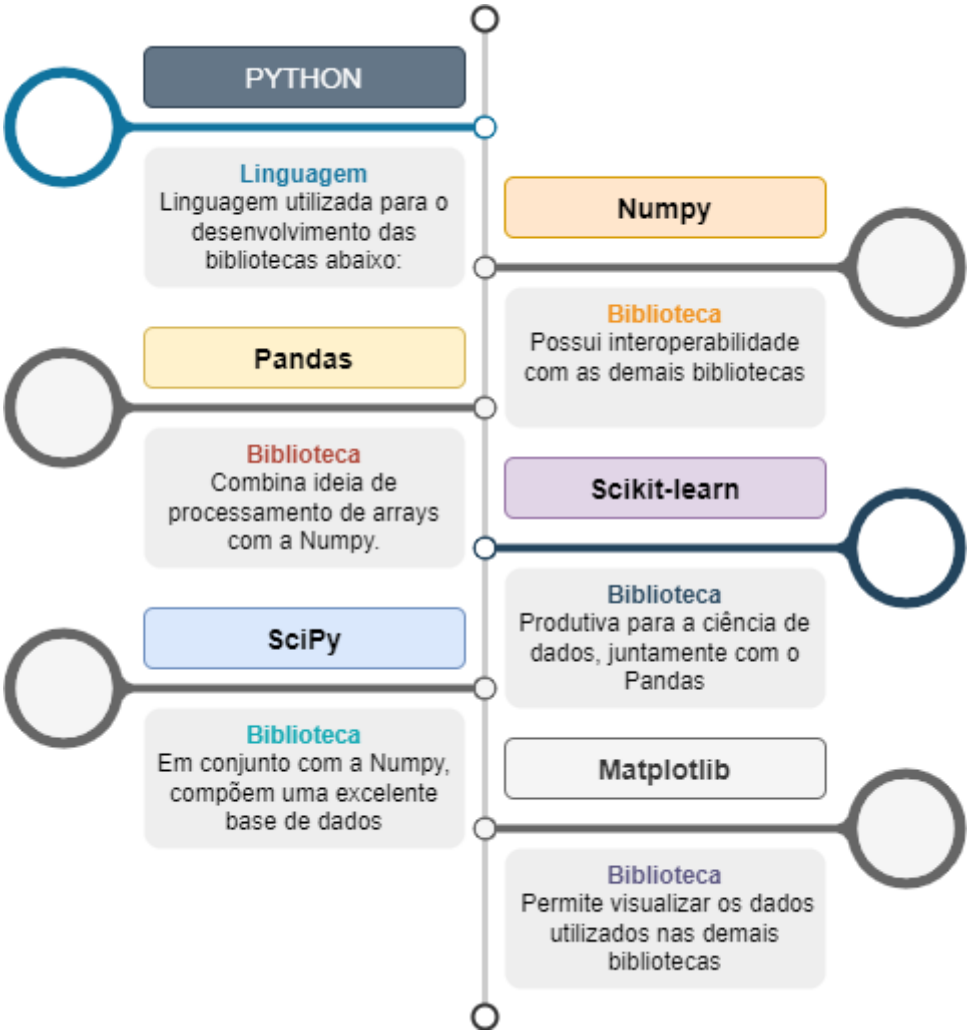
2.5 Matplotlib

A biblioteca Matplotlib (https://matplotlib.org/), criada por John D. Hunter, foi projetada para fazer plotagens, gráficos, visualização de dados, se tornou uma das mais utilizadas pelos programadores python, por ser criada da linguagem python, tem uma boa integração com essa linguagem.

4. Resultados Experimentais

As pesquisas realizadas demonstram que a função e aplicabilidade de bibliotecas para o uso de Python podem ser vistas com olhares distintos, em projetos diversos, mas sempre visando a otimização do código ou programa que está sendo desenvolvido. O resultado em todas as tentativas, demonstram que as bibliotecas desempenham específicas funções, dentro de uma cadeia hierárquica, criando uma intercooperação relacional entre elas, conforme expõe a figura:

FIGURA:



A atividade correlacional das bibliotecas Python, combina a ideia de cooperação entre a linguagem de programação e suas respectivas ferramentas de desenvolvimento, portanto, entende-se, nesta pesquisa, que a aplicabilidade das bibliotecas é de essencial importância dentro do objeto de pesquisa e desenvolvimento de códigos e programas.

As classificações foram definidas de forma bem sucinta, respeitando os parâmetros de uma pesquisa inicial, de cunho acadêmico, Faz-se assim necessário, continuidade para melhor compreensão do objeto de estudo a ser explorado posteriormente sobre as principais e mais utilizadas bibliotecas disponíveis para pesquisa e programação.

5. Conclusões e Trabalhos Futuros

O artigo evidenciou uma breve resposta sobre a aplicabilidade de bibliotecas Python para Machine Learning, no intuito de fomentar, inicialmente, o desenvolvimento de estudos mais aprofundados sobre o tema proposto, entretanto, neste primeiro momento, a pesquisa realizada elucidou um entendimento evidente à compreensão teórica. À progressão de novos parâmetros, em uma futura análise mais aprofundada sobre o desenvolvimento de pesquisas para Machine Learning, faz-se necessário novos estudos

e pesquisas a contextualizar de forma mais específica e com maior abrangência, o tema aqui elucidado.

7. Referências

SCIKIT-LEARN. Documentação da biblioteca. Disponível em: https://scikit-learn.org/stable/ Acesso: 07 mai. 2022.

SCIPY. Documentação da biblioteca. Disponível em: https://scipy.org/ Acesso: 07 mai. 2022.

NUMPY. Documentação da biblioteca. Disponível em: https://docs.scipy.org/doc/ Acesso: 03 mai. 2022.

PANDAS. Documentação operacional da biblioteca. Disponível em: https://pandas.pydata.org/docs/ Acesso: 08 mai. 2022.

PYTHON. Documentação da linguagem Python Disponível em: https://docs.python.org/ptbr/3/ Acesso: 03 mai. 2022.