Faculdade de Tecnologia de Jundiaí – "Deputado Ary Fossen"

Curso Superior de Tecnologia em Ciência de Dados - 2024

REGRESSÃO LINEAR SIMPLES EM AMBIENTES DE COMPUTAÇÃO EM 'NUVEM'

André Santos de Oliveira

Resumo

Neste artigo apresentamos uma análise comparativa da implementação de modelos de regressão linear simples em três das principais plataformas de computação em nuvem: Google Cloud Platform (Vertex AI), Amazon Web Services (SageMaker) e Microsoft Azure (Azure ML). Através de um experimento prático, utilizando um conjunto de dados sintético sobre a temperatura corporal de cães em relação à idade, são exploradas as características e funcionalidades de cada plataforma. As análises se concentram em aspectos como a configuração do ambiente, a facilidade de uso das ferramentas de desenvolvimento, a escalabilidade dos recursos computacionais, a integração com outras ferramentas e serviços. Os resultados deste estudo fornecem insights valiosos para pesquisadores e profissionais de dados que buscam a plataforma de nuvem ideal para desenvolver e implantar modelos de regressão linear, considerando as demandas específicas de seus projetos.

Palavras-chave: Regressão Linear; Computação em Nuvem; Vertex AI; SageMaker; Azure ML; JupyterLab; Python; Comparação de Plataformas

Abstract

This paper presents a comparative analysis of the implementation of simple linear regression models on three major cloud computing platforms: Google Cloud Platform (Vertex AI), Amazon Web Services (SageMaker), and Microsoft Azure (Azure ML). Through a practical experiment using a synthetic dataset on the relationship between dog body temperature and age, the characteristics and functionalities of each platform are explored. The analyses focus on aspects such as environment configuration, ease of use of development tools, scalability of computational resources, integration with other tools and services. The results of this study provide valuable insights for researchers and data professionals seeking the ideal cloud platform for developing and deploying linear regression models, considering the specific demands of their projects.

Keywords: Linear Regression; Cloud Computing; Vertex AI; SageMaker; Azure ML; JupyterLab; Python; Platform Comparison

Figuras

Introdução

O aprendizado de máquina ('*machine learning*') se consolidou como uma ferramenta poderosa para solucionar problemas complexos em diversas áreas, impulsionando a tomada de decisão estratégica a partir de dados.

> "Deep learning has transformed the use of machine learning technologies for the analysis of large experimental datasets. In science, such datasets are typically generated by large- scale experimental facilities, and machine learning focuses on the identification of patterns, trends and anomalies to extract meaningful scientific insights from the data. (Thiyagalingam, Shankar, Fox, & Hey, 2022)"

A crescente disponibilidade de plataformas de computação em nuvem, como Google Cloud Platform (GCP), Amazon Web Services (AWS) e Microsoft Azure, democratizou o acesso a recursos de processamento e armazenamento de dados, tornando viável a aplicação de soluções de aprendizado de máquina em escalas jamais vistas. Estas plataformas oferecem uma ampla gama de serviços, incluindo armazenamento de dados, processamento em larga escala, algoritmos de aprendizado de máquina predefinidos e ferramentas para treinamento, ajuste e implantação de modelos (Madan & Reich, 2021).

Compreender as semelhanças e diferenças entre as soluções torna-se crucial para a melhor escolha de qual ou quais ferramentas usar. Entretanto, isto demandaria o trabalho de centenas de usuários, pesquisadores, estudiosos. Seria necessário abordar os mais variados tipos de problemas para, então, traçar com assertividade qual ferramenta é mais adequada para cada caso. Nosso objetivo é mais modesto. Partimos de um problema simples, que pode ser abordado por um modelo de regressão linear simples e examinamos como é a sua implementação.

> "In practice, the selection of an ML algorithm for a given scientific problem is more complex than just selecting one of the ML technologies and any particular algorithm. The selection of the most effective ML algorithm is based on many factors, including the type, quantity and quality of the training data, the

> availability of labelled data, the type of problem being addressed (prediction, classification and so on), the overall accuracy and performance required, and the hardware systems available for training and inferencing. With such a multidimensional problem consisting of a choice of ML algorithms, hardware architectures and a range of scientific problems, selecting an optimal. (NG, 2019)"

A Regressão Linear

Regressão Linear é uma ferramenta estatística poderosa que permite modelar e examinar a relação entre variáveis. Dentro da análise de regressão, as regressões lineares simples e múltiplas são amplamente utilizadas em diversas áreas do conhecimento, desde a pesquisa científica, tomada de decisão em negócios, economia, saúde, engenharia, entre outras (kfollis, 2023). A regressão linear fornece *insights* valiosos sobre a relação entre variáveis e permite a construção de modelos preditivos.

> "Regression – Regression establishes a relationship between the input variables (also known as independent variables or features) and the target variable (also known as the dependent variable). This relationship is captured through a mathematical function or model that maps the input variables to a continuous output. It is commonly used for tasks such as predicting house prices based on features like square footage and the number of bathrooms, stock market trends, or estimating sales figures. (AMAZON WEB SERVICES, 2024)"

O termo "regressão" tem origem nos estudos de Sir Francis Galton sobre hereditariedade no final do século XIX. Ao analisar a relação entre a altura dos pais e a altura de seus filhos, Galton observou que, embora filhos de pais altos tendessem a ser altos e filhos de pais baixos tendessem a ser baixos, a altura dos filhos tendia a ser mais próxima da média da população do que a altura de seus pais. Esse fenômeno, que Galton chamou de "regressão à média", é a base para o termo "regressão" em estatística. (Morettin, 2010)

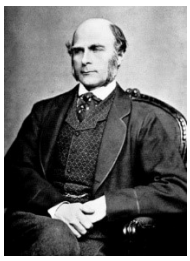

Figura 1: Sir Francis Galton, 1850s (Wikipedia)

Regressão Linear Simples

A regressão linear simples é utilizada para modelar a relação linear entre duas variáveis quantitativas: uma variável dependente (Y) e uma variável independente (X). O objetivo é encontrar uma equação matemática que represente a reta que melhor se ajusta aos dados, permitindo prever o valor de Y a partir de um determinado valor de X (Developers, 2024). Por exemplo, podemos utilizar a regressão linear simples para modelar a relação entre a idade de um cão (variável independente - X) e a temperatura corporal observada (variável dependente - Y). A equação da reta, nesse caso, seria:

Y = β0 + β1*X, onde:

- Y é a temperatura corporal (variável dependente)
- X é a idade do cão (variável independente)
- β0 é o coeficiente linear (valor de Y quando X = 0)
- β1 é o coeficiente angular da reta (inclinação da reta, que indica o quanto Y varia para cada unidade de variação em X)

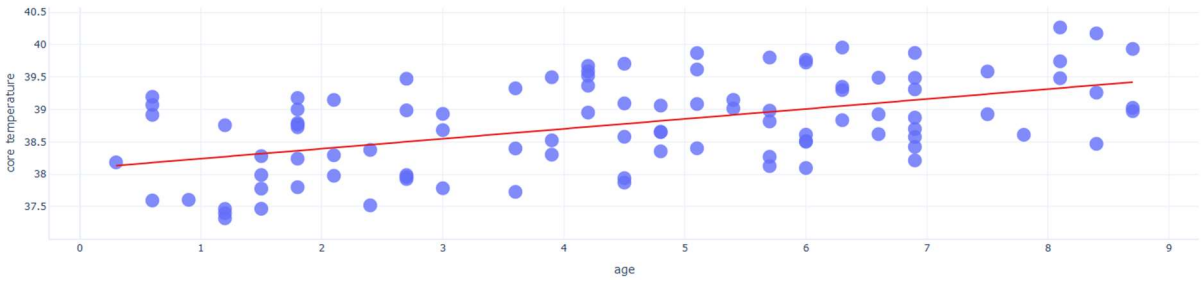

Figura 2: Regressão linear simples, relação entre duas variáveis: idade e temperatura

Nosso Exemplo de Regressão Linear Simples

Para esta exploração dos recursos disponíveis nas plataformas Cloud, escolhemos um problema simples. Analisamos a variação da temperatura corporal de cães. O objetivo é identificar se estão ou não doentes com base na temperatura observada e outras variáveis, como por exemplo a idade do cão. (MICROSOFT, 2024)

Para desenvolvimento deste modelo, seguimos o tutorial disponível na plataforma 'Microsoft Learn' (MICROSOFT, Treinar e entender os modelos de regressão no

aprendizado de máquina, 2024). A implementação em linguagem Python para execução em notebook JupyterLab está disponível em: https://github.com/aso1976/artigo-regressao-cloud

O problema explorado no conjunto de dados sobre a temperatura corporal dos cães é a relação entre a idade do cão e sua temperatura corporal registrada. O objetivo é construir um modelo de regressão linear simples para prever a temperatura corporal com base na idade e entender essa relação.

Os dados indicam que os cães, em sua maioria, apresentam temperatura corporal levemente elevada, sugerindo que estão doentes. Um pequeno número de cães tem temperatura acima de 40 graus, o que indica um estado de saúde mais grave.

Observando a distribuição dos dados, percebe-se que a idade dos cães não parece ter uma forte influência na temperatura. Apesar disso, a análise visual dos dados sugere que cães mais velhos tendem a ter temperaturas mais altas, embora essa relação seja "ruidosa", com variações significativas de temperatura entre cães da mesma idade.

Um modelo de regressão linear simples foi ajustado aos dados para avaliar formalmente a relação entre idade e temperatura. O modelo resultante confirma a hipótese de uma correlação positiva, indicando que a temperatura corporal aumenta com a idade. Os parâmetros do modelo indicam que a temperatura média de um cão com idade 0 é de 38 graus Celsius. A cada ano de vida, a temperatura corporal aumenta em média 0.15 graus Celsius.

É importante destacar que, apesar do modelo indicar uma correlação positiva, a relação entre idade e temperatura corporal não é perfeita. Outros fatores, além da idade, podem influenciar a temperatura corporal de um cão. Entretanto, este modelo de regressão linear simples, de fácil entendimento, nos é útil para explorarmos como seria sua implementação nas soluções 'cloud' oferecidas pelas plataformas.

Soluções de Computação em Nuvem

Para este ensaio, selecionamos as soluções de machine learning mais populares do mercado, das principais plataformas em nuvem: Vertex AI (Google Cloud Platform), SageMaker (Amazon Web Services) e Azure Machine Learning (Microsoft Azure). Essas plataformas foram escolhidas por sua ampla adoção e recursos robustos.

As grandes plataformas em nuvem não esgotam as opções para ciência de dados. Plataformas como RapidMiner, DataRobot, H2O.ai e Alteryx complementam esse cenário, oferecendo alternativas e especializações em diferentes áreas.

Nosso objetivo é configurar um ambiente de processamento em cada uma das soluções selecionadas, instanciar uma sessão do JupyterLab e executar nosso modelo de regressão linear simples.

Google Cloud Plataform (GCP), Vertex AI.

O Vertex AI é uma plataforma abrangente do Google Cloud Platform (GCP) que oferece uma variedade de ferramentas e serviços para desenvolver, treinar e implantar modelos de machine learning. Ele simplifica o processo de criação de soluções de inteligência artificial, desde a preparação dos dados até a produção.

O Workbench é um componente essencial do Vertex AI. Ele fornece um ambiente de desenvolvimento integrado (IDE) baseado em notebook Jupyter, onde cientistas de dados podem escrever e executar código, experimentar diferentes algoritmos e visualizar resultados. O Workbench oferece uma interface intuitiva e personalizável, além de integração com outras ferramentas do Google Cloud, como BigQuery e Cloud Storage.

> "O Vertex AI Workbench oferece uma experiência do JupyterLab e recursos avançados de personalização.
> Os notebooks da Vertex AI fornecem uma infraestrutura de computação totalmente gerenciada, escalonável e pronta para empresas com controles de segurança e recursos de gerenciamento de usuários." (CLOUD, Vertex AI, 2024)

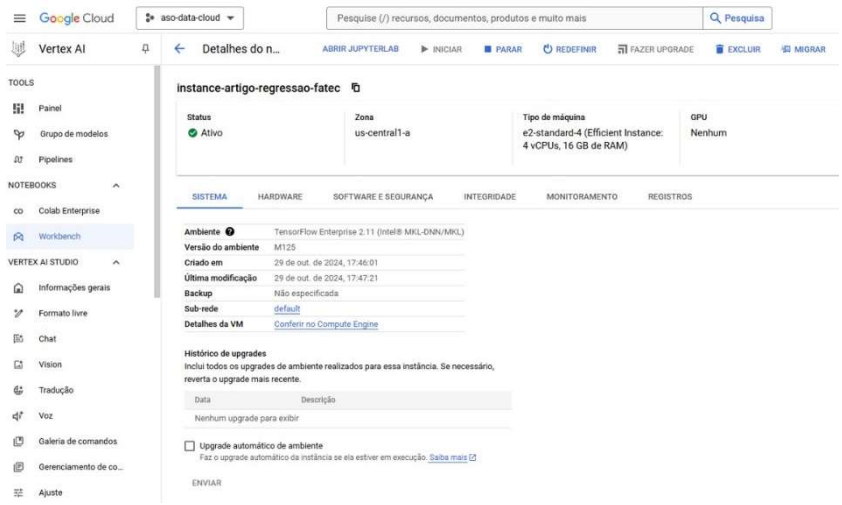

Figura 3: Ambiente Workbench, na Vertex AI, configurado para execução do JupyterLab.

Para implementação de nosso modelo de regressão linear no GCP, ativamos o acesso ao componente 'Notebook API'. E então, utilizando o console do Vertex AI, criamos

uma instância do WorkBench que levantou um ambiente calçado no TensorFlow. Nele abrimos um notebook JupyterLab e executamos nosso modelo.

Amazon Web Services (AWS), SageMaker.

O SageMaker é um serviço de aprendizado de máquina totalmente gerenciável, disponível na plataforma AWS, que possibilita aos cientistas de dados e desenvolvedores: construir, treinar e implantar modelos de forma rápida e eficiente.

Ele oferece interface de usuário amigável para executar fluxos de trabalho, disponibilizando ferramentas de aprendizado de máquina em vários ambientes de desenvolvimento integrados (IDEs). Eliminando a necessidade de construir e gerenciar servidores próprios para armazenamento e compartilhamento de dados. Isto libera tempo para a construção e desenvolvimento de fluxos de trabalho de forma colaborativa.

> "Amazon SageMaker is a fully managed machine learning (ML) service. With SageMaker, data scientists and developers can quickly and confidently build, train, and deploy ML models into a production-ready hosted environment. It provides a UI experience for running ML workflows that makes SageMaker ML tools available across multiple integrated development environments (IDEs)." (AMAZON WEB SERVICES, 2024)

SageMaker Studio oferece um conjunto de IDEs que incluem, além do JupyterLab, o Code Editor, RStudio e SageMaker Studio Classic. O JupyterLab no Studio fornece *kernels* que iniciam em segundos, com tempo de execução pré-configurados, *frameworks* populares para ciência de dados, aprendizado de máquina e armazenamento em bloco de alto desempenho.

> "Amazon SageMaker Notebook Instances: Lets you prepare and process data, and train and deploy machine learning models from a compute instance running the Jupyter Notebook application" (AMAZON WEB SERVICES, 2024).

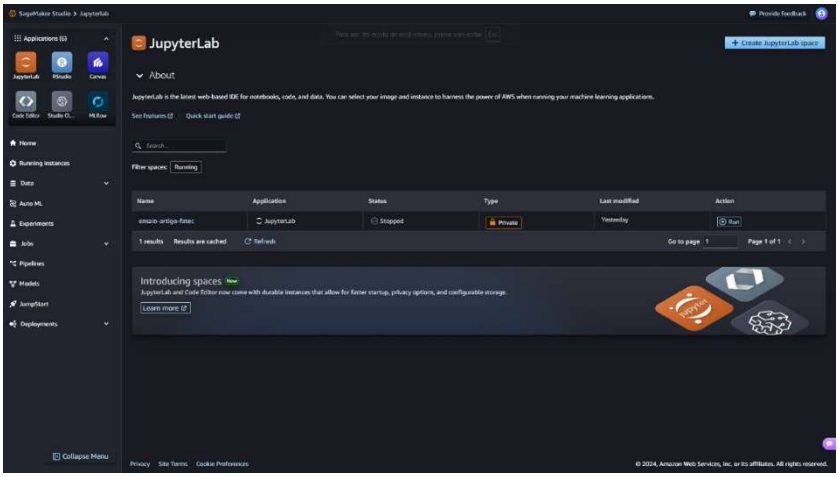

Figura 4: Ambiente SageMaker Studio, no SageMaker, configurado para execução do JupyterLab

Para implementação de nosso modelo de regressão linear na AWS criamos um domínio do Amazon SageMaker e iniciamos a partir dele uma instância do JupyterLab que nos foi entregue operacional e executamos nosso modelo.

Microsoft Azure (AZURE), Azure ML

O Azure Machine Learning (Azure ML) é uma plataforma abrangente da Microsoft que oferece um ambiente completo para cientistas de dados e engenheiros de aprendizado de máquina criarem, treinarem e implantarem modelos de machine learning. Essa plataforma na nuvem proporciona ferramentas e serviços para acelerar todo o ciclo de vida de um projeto de ML, desde a experimentação inicial até a produção em larga escala.

> "O Azure Machine Learning é um serviço de nuvem para acelerar e gerenciar o ciclo de vida dos projetos de ML (machine learning). Profissionais, cientistas de dados e engenheiros de ML podem usá-lo em seus fluxos de trabalho cotidianos para: treinar e implantar modelos e gerenciar MLOps (operações de aprendizado de máquina).
> Você pode criar um modelo em Machine Learning ou usar um modelo criado em uma plataforma de código aberto, como PyTorch, TensorFlow ou scikit-learn. As ferramentas do MLOps ajudam você a monitorar, treinar e reimplantar modelos." (Microsoft, 2024)

Utilizamos o 'Jupyter Notebook' personalizado disponível no 'Azure Machine Learning'. Este notebook possui funcionalidades adicionais e integrações específicas para a plataforma. Essa integração oferece uma experiência aprimorada para cientistas de dados e engenheiros de machine learning, permitindo a execução de experimentos, a construção de modelos e a colaboração em projetos diretamente na nuvem. (Microsoft, 2024)

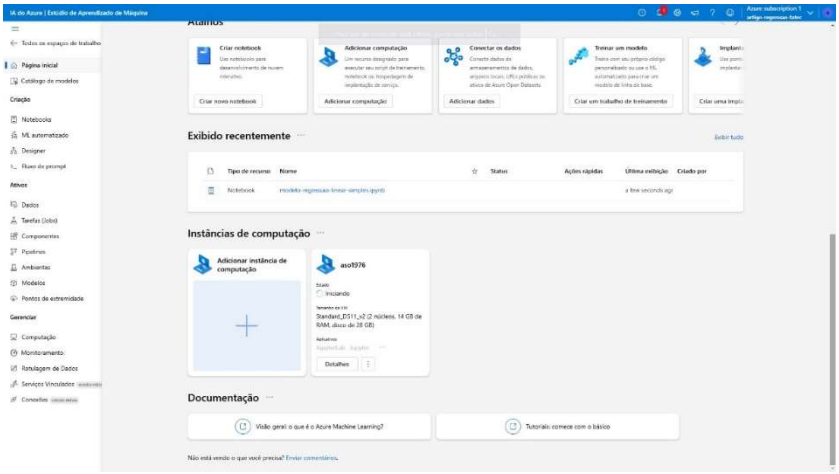

Figura 5: Ambiente Azure ML, na Azure, configurado para execução do notebook "Jupyter"

Para implementação de nosso modelo de regressão linear na AZURE iniciamos a instância de computação, acessamos o notebook Jupyter personalizado da plataforma, selecionamos o kernel apropriado, e executamos nosso modelo.

Semelhanças

As soluções Vertex AI, SageMaker e Azure ML compartilham diversas semelhanças na usabilidade e execução de notebooks Jupyter. Todas elas oferecem ambientes de desenvolvimento integrados (IDEs) baseados na web, o que facilita a criação, edição e execução dos notebooks. O SageMaker Studio, da Amazon, é um IDE abrangente projetado para fluxos de trabalho de ML. Ele fornece uma interface baseada na web para todas as etapas de desenvolvimento de ML, incluindo notebooks Jupyter gerenciados que se integram a outros serviços da AWS. Já a Microsoft com o Azure ML Studio entrega um portal web para desenvolvimento de ML sem código ou com pouco necessidade de código, além de instâncias de computação, que são maquinas virtuais (VMs) gerenciadas para executar notebooks Jupyter. O Google, por sua vez, oferece o Vertex AI Workbench, um ambiente unificado para desenvolvimento e execução de notebooks Jupyter. (CloudOptimo, 2024)

Os notebooks Jupyter são entregues pré-configurados com os kernels e pacotes Python necessários para ciência de dados e aprendizado de máquina. Vale destacar que, em nosso teste, o kernel escolhido por padrão no Azure ML, "Python 3", apresentou diversas incompatibilidade de bibliotecas. Problema resolvido com a seleção de outro kernel mais recente, "Python 3.8". Já as soluções do Google e AWS entregaram instâncias do JupyterLab totalmente funcionais.

As plataformas incluem o suporte a frameworks de ML populares, como TensorFlow, PyTorch, Scikit-learn e MXNet. É possível escalar os recursos de processamento, permitindo que os usuários executem notebooks Jupyter em instâncias de computação com diferentes capacidades de CPU, memória e GPU. Por fim, as plataformas oferecem recursos para colaboração em notebooks Jupyter, permitindo que os usuários compartilhem seus notebooks com outras pessoas, trabalhem em notebooks simultaneamente e controlem as versões de seus notebooks. (CloudOptimo, 2024)

Diferenças

O SageMaker, profundamente integrado ao ecossistema AWS, pode oferecer uma experiência mais intuitiva para usuários familiarizados com outros serviços AWS (CloudOptimo, 2024). Em nosso teste a velocidade para disponibilizar o ambiente JupyterLab foi destaque. A instância de processamento escolhida iniciou em segundos.

Similarmente, o Azure ML, com sua forte integração com os outros produtos Microsoft, pode ser mais fácil de usar para aqueles familiarizados com as ferramentas da Microsoft (CloudOptimo, 2024). Entretanto, a configuração da instância de computação necessária e a versão do notebook Jupyter personalizado, dentro do 'Workspace do Azure Machine Learning', se mostraram menos intuitivos que a solução da AWS.

Já o Vertex AI, aproveita a expertise do Google em IA e a integração com o Google Cloud. Pode ser mais adequado para usuários que já utilizam outros serviços do Google (CLOUD, 2024). Nos entregou o ambiente JupyterLab com a mesma velocidade observada na AWS com as bibliotecas Python para ciência de dados instaladas e compatíveis.

Custos

Este artigo tem como objetivo apresentar uma visão geral das plataformas Vertex AI Workbench, SageMaker Studio e Azure ML, destacando suas funcionalidades e benefícios. Devido à complexidade e variabilidade dos custos de processamento em plataformas de *cloud computing*, não realizamos uma comparação de preços neste material. A escolha da plataforma mais adequada depende de diversos fatores específicos a cada projeto, incluindo os requisitos como recursos computacionais, o volume de dados e as necessidades de negócios. Recomendamos que os leitores

consultem a documentação oficial de cada plataforma e realizem testes práticos para obter estimativas precisas de custos.

Resultados da Execução do Modelo de Regressão Linear

Todas as execuções do modelo entregaram os mesmos resultados esperados: Intercept: 38.087867548892085 Slope: 0.15333957754731858

Através das interfaces dos próprios notebooks, fizemos o carregamento (*upload*) do nosso arquivo Jupyter "*modelo-regressao-linear-simples.ipynb*" e o executamos.

A arquivo está disponível em: https://github.com/aso1976/artigo-regressao-cloud.

Considerações Finais

As três plataformas oferecem ambientes de desenvolvimento integrados baseados na web, com JupyterLab pré-configurado e suporte a frameworks populares de machine learning. Além disso, permitem o escalonamento dos recursos de processamento e oferecem recursos para colaboração. A integração com outros serviços das plataformas, a velocidade de provisionamento do ambiente e a intuitividade da interface do usuário variam entre as soluções. O SageMaker se destaca pela rápida inicialização da instância de processamento e integração com o ecossistema AWS. O Azure ML se integra bem com outros produtos Microsoft, mas a configuração da instância de computação e a versão do Jupyter Notebook personalizado podem ser menos intuitivas. O Vertex AI aproveita a expertise do Google em IA e integração com o Google Cloud, entregando o ambiente JupyterLab funcional com bibliotecas Python compatíveis.

REFERÊNCIAS:

AMAZON WEB SERVICES, I. (28 de 10 de 2024). *Amazon SageMaker - Developer Guide.* Acesso em 28 de 10 de 2024, disponível em Amazon SageMaker: https://docs.aws.amazon.com/sagemaker/latest/dg/

CLOUD, G. (29 de 10 de 2024). *Vertex AI.* Fonte: Google Cloud: https://www.google.com/url?sa=E&source=gmail&q=https://cloud.google.com/vertex-ai/docs?hl=pt-br

CloudOptimo. (31 de 10 de 2024). *SageMaker vs Azure ML vs Google AI Platform: A Comprehensive Comparison.* Fonte: CloudOptimo: https://www.cloudoptimo.com/blog/sagemaker-vs-azure-ml-vs-google-ai-platform-a-comprehensive-comparison/

Developers, G. f. (13 de 08 de 2024). *Regressão linear*. Fonte: Machine Learning: https://developers.google.com/machine-learning/crash-course/linear-regression?hl=pt-br

kfollis, p. T. (23 de 12 de 2023). *Algoritmo Regressão Linear da Microsof*. Fonte: learn.microsoft.com: https://learn.microsoft.com/pt-br/analysis-services/data-mining/microsoft-linear-regression-algorithm?view=asallproducts-allversions

Madan, M., & Reich, C. (2021). Comparison of Benchmarks for Machine Learning Cloud Infrastructures. *CLOUD COMPUTING 2021 : The Twelfth International Conference on Cloud Computing, GRIDs, and Virtualization*.

MICROSOFT. (06 de 10 de 2024). *Exercise - Train a Linear Regression Model.* Acesso em 06 de 10 de 2024, disponível em https://learn.microsoft.com/pt-br/training/modules/understand-regression-machine-learning/3-exercise-train-linear-regression

Microsoft. (31 de 10 de 2024). *Learn - Azure - Machine Learning.* Fonte: Learn: https://learn.microsoft.com/pt-br/azure/machine-learning/how-to-run-jupyter-notebooks?view=azureml-api-2

MICROSOFT. (28 de 10 de 2024). *Treinar e entender os modelos de regressão no aprendizado de máquina.* Fonte: Learn: https://learn.microsoft.com/pt-br/training/modules/understand-regression-machine-learning/3-exercise-train-linear-regression

Morettin, B. e. (2010). *Estatística Básica, Capitulo 16.*

NG, A. (2019). *Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning. [S.l.]:.* DeepLearning.AI.

Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (6 de April de 2022). Scientific machine learning. *Nature Reviews Physics*, pp. https://www.nature.com/articles/s42254-022-00441-7.