Proposta de uma arquitetura geral de gerência de dados para tarefas de Machine Learning e análise de dados com aplicações na agricultura

Jennifer do Prado da Silva[1], **Daniel dos Santos Kaster**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

jennifer.prado@uel.br, dskaster@uel.br

Abstract. *It is possible to observe that Artificial Intelligences (AIs) are currently gaining strength and quickly becoming the pillar of innovation, with the ability to identify patterns, solve problems, perform solutions to automate tasks and provide a more comprehensive understanding of the abundance of data available. In the construction of Machine Learning (ML) systems, it is necessary to collect, transform and organize the data before inducing the AI models, these data being of different types, sources and scales. Aiming at this stage, the objective of this project is to present a general architecture of tools for data collection, cleaning and pre-processing, filtering constructed and representative data for ML tasks and problems with applications outside the area of agriculture. It is expected that the results presented are an architecture with integrated open source tools, as well as use cases, capable of implementing pipelines of real situations related to tasks supporting Machine Learning in agriculture.*

Resumo. *É possível observar que atualmente a Inteligência Artificial está ganhando força e rapidamente se tornando o pilar da inovação, com capacidade de identificar padrões, resolver problemas, realizar previsões para automatizar tarefas e proporcionar uma compreensão mais abrangente em relação a abundância de dados disponíveis. Na construção de sistemas de Aprendizado de Máquina, é preciso realizar coleta, transformação e organização dos dados antes de induzir os modelos de IA, sendo estes dados de diferentes tipos, fontes e escalas. Visando esta etapa, o objetivo deste projeto é apresentar uma arquitetura geral de ferramentas para coleta, limpeza e pré-processamento dos dados, filtrando dados qualificados e representativos para tarefas e problemas de Aprendizado de Máquina com aplicações voltadas a área da agricultura. Espera-se que os resultados apresentem uma arquitetura com ferramentas de código aberto integradas, assim como casos de uso, capazes de implementar pipelines de situações reais relacionadas a tarefas de suporte ao Aprendizado de Máquina na agricultura.*

1. Introdução

Em 2017, o site de publicidade inglês de notícias e assuntos internacionais, *The Economist*, afirmou que o recurso mais valioso do mundo não é mais o petróleo, mas os dados [7]. Diante desta afirmação, é comum observar que empresas de todo porte, instituições e organizações vem lidando com uma quantidade de dados significativamente

grande, de todo tipo e de uma variedade de fontes, tornando cada vez mais evidente que a transformação de dados é frequentemente necessária [25].

É comum observar inúmeras aplicações do aprendizado de máquina em diversas áreas do conhecimento, como por exemplo na área da saúde, tecnológica, biológica e agrária. No caso desta última, inúmeras mudanças e inovações tecnológicas ocorreram nos anos atuais, com isso é possível observar que a inteligência artificial (IA) e o aprendizado de máquina são cada vez mais usados para previsão na agricultura [4] [16], uma das principais áreas em que os grandes volumes de dados e tipos de dados complexos são predominantes.

Em uma de suas matérias, no ano de 2019, a empresa ucraniana de TI, *Sciforce*, afirmou que o aprendizado de máquina está em toda parte durante todo o ciclo de cultivo e colheita [27]. Pode-se destacar os estudos realizados por pesquisadores que recorrem aos métodos de aprendizado de máquina e tecnologia de visão computacional para a identificação de doenças e pragas agrícolas [32]. Essas ideias e estudos enfocam que a agricultura pode sim se beneficiar do aprendizado de máquina para um melhor desempenho e resultados em suas atividades.

Em sua maioria, as tarefas de aprendizado de máquina direcionadas a área agrícola envolvem uma grande quantidade de dados. Com isso, a agricultura passa a ser cada vez mais dependente de soluções baseadas em dados e informações, para geração de diferenciais e elaboração de estratégias de impulsionamento do setor [14]. Isso acontece devido ao crescente volume e variedades de dados no campo. As técnicas convencionais de processamento de dados são incapazes de atender às demandas cada vez maiores na nova era da agricultura inteligente, o que é um importante obstáculo para extrair informações valiosas dos dados de campo [4].

Trabalhar com dados agrícolas requer levar em consideração que alcançar alta precisão e interpretabilidade é um desafio [5]. Esse grande volume de dados exigirá abordagens inteligentes a fim de evitar que se tornem apenas um enorme acúmulo [9]. Tal afirmação se direciona aos inúmeros tipos de dados existentes no meio agrícola que, se não preparados, podem prejudicar na previsão de uma dada variável dependente, na acurácia e até mesmo na otimização do processo, ao invés de oferecer *insights* que contribuam no acompanhamento desde a produção até o consumidor final.

Visando este panorama no aprendizado de máquina, muitas empresas se adaptaram e se especializaram para oferecer ferramentas úteis, no quesito de preparação dos dados, capazes de solucionar os problemas existentes no conjunto de dados, como por exemplo o *Azure Databricks* da empresa *Microsoft*, uma plataforma unificada que disponibiliza recursos para ingestão, preparação, análise e monitoramento de dados [20], dividindo o cenário com outras companhias especialistas, assim como o serviço *Oracle Cloud Infrastructure Data Integration*, da empresa *Oracle*, uma interface gerenciada, sem servidor e nativa da nuvem, responsável por extrair, carregar, transformar, limpar e remodelar dados [22], a interface *Amazon SageMaker Data Wrangler*, da empresa *Amazon*, a qual proporciona recursos para limpeza, exploração, visualização e processamento de dados tabulares e de imagem em grande escala [1] e outros sistemas disponíveis das diversas empresas no mercado da tecnologia. Estes serviços oferecem uma gama de recursos favoráveis para a preparação de dados, mas são plataformas limitadas, ja que cobram pe-

los recursos consumidos ou entregam pacotes de recursos por um valor preestabelecido. Além deste impasse relacionado a monetização, pode ser destacado também a limitação que o cliente tem em relação aos recursos próprios da empresa fornecedora do serviço, privando seus usuários e fazendo com que estes sejam incapazes de aplicar ao seu conjunto de dados, processos de preparação de dados provenientes de companhias concorrentes.

Neste contexto, este projeto propõe apresentar uma arquitetura geral para o processamento de grandes volumes de dados, com ferramentas integradas de código aberto que sejam compatíveis entre as demais, próprias para a coleta e limpeza de dados, bem como extrair informações relevantes destes, de preparar e gerenciar os diversos tipos de dados agro meteorológicos existentes e integrar as ferramentas selecionadas de maneira estratégica, de modo que seja possível desfrutar o máximo de suas funcionalidades e manter informações válidas, relevantes e precisas para aplicações em tarefas e soluções agrícolas baseadas em Aprendizado de Máquina.

2. Fundamentação Teórico-Metodológica e Estado da Arte

2.1. Tarefas de preparação de dados para aprendizado de máquina ponta a ponta

O aprendizado de máquina é uma ramificação da inteligência artificial e da ciência da computação, e tem como foco, o uso de dados, algoritmos e modelos estatísticos para simular o modo como os humanos aprendem. Para isso, essa ramificação aprimora gradualmente sua precisão de forma autônoma, utilizando redes neurais e aprendizado profundo sem a necessidade de ser programado explicitamente, confiando em padrões e inferências. Com isso, quanto maior for a alimentação de dados no sistema, mais precisos serão os resultados [13] [11].

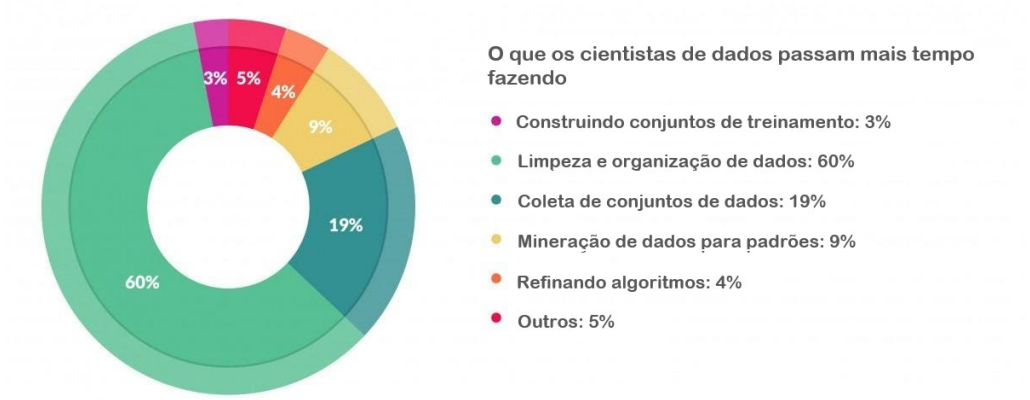
O conceito de aprendizado de máquina é muito importante para se ter ideia de como funciona este subconjunto da inteligência artificial, mas para compreender todo o processo e funcionamento de um projeto que abrange a concepção deste subconjunto é essencial estar inteirado sobre as etapas do aprendizado de máquina ponta a ponta. Este é um processo, em que uma máquina aprende um mapeamento entre uma série de entradas (X) para algumas saídas conhecidas (y), sem ser explicitamente programada [31]. Pode ser definido em três etapas. A primeira, foca em possuir ampla compreensão dos dados, meio de coleta e métodos para limpeza de dados. Em seguida, a segunda etapa, destina-se em selecionar e implementar o modelo. Por fim, a terceira etapa, é a definição dos parâmetros do modelo e ajuste de dados [19].

No aprendizado de máquina ponta a ponta, como ilustrado na figura 1, os cientistas de dados gastam 19% de seu tempo na coleta de conjuntos de dados [23], uma das tarefas mais longas na construção de um modelo. Isso acontece pois na execução deste processo há dois fatores fundamentais a serem considerados, a qualidade e quantidade. Contando com um conjunto de dados grande o suficiente é possível ter mais análises significativas e dispor de dados de alta qualidade garante que as informações sejam relevantes o bastante para o caso de uso em questão e proporcionam modelos resultantes com altas taxas de precisão [24].

Já a limpeza e organização dos dados é responsável por cerca de 60% do trabalho dos cientistas de dados [23] e com isso a etapa mais demorada e importante na construção de um modelo. Neste processo, são realizadas diversas verificações, como tratamento

de valores ausentes com dados inferidos a partir dos dados existentes, modificação de variáveis categóricas, identificação de valores discrepantes, expandir *features* existentes, limpeza e várias outras transformações dos dados de treinamento para que estejam prontos para análise de dados e otimização de modelo [24] [2].

Figura 1. Gráfico percentual do tempo gasto por cientistas de dados



Fonte: Adaptado de Forbes, 2016.

Após o longo processo de limpeza dos dados, é realizado a etapa de engenharia de *features*. *Features* são as entradas que os modelos de aprendizado de máquina utilizam durante o treinamento e a inferência para fazer previsões. Esta etapa, envolve combinações de diversas análises de dados e conhecimento e domínio da área relacionada aos dados selecionados. É essencial considerar, quais dados são necessários e relevantes, se informando com especialistas, realizando *brainstorming* ou por meio de pesquisas aprofundadas sobre o assunto. Concluindo esses exercícios, é possível evitar a perda de variáveis preditoras importantes [3].

É possível utilizar a engenharia de *features* em diversas áreas de aplicação, como exemplo, tem-se o projeto realizado por estudantes da Universidade Federal da Bahia(UFBA), que aplicou engenharia de *features* linguísticas para classificação de triplas relacionais no Galego, Português Brasileiro e Espanhol Europeu [17]. Em outro projeto, foi aplicada em conjunto com a avaliação de *features*, para realizar extração de informações em Notas Fiscais. Neste último, a engenharia de *features* foi utilizada para determinar conjuntos de *features* que possuíam maior relevância para identificação de elementos de notas fiscais eletrônicas [6].

2.2. Integração entre sistemas

A definição de integração de sistemas pode ser descrita como a capacidade de realizar conexões de dados, aplicativos, *APIs* e dispositivos [12], ou seja, diferentes ferramentas conversarem entre si e trocarem dados, reunindo ou incorporando partes em um todo, de maneira que a integração desses elementos contribuam para melhorar a eficiência, intensificar a produtividade, criar rotinas mais inteligentes com comunicação eficiente, e a garantir agilidade dos processos [30].

Para formar uma integração entre sistemas, alguns fatores devem ser considerados neste processo, como definir objetivos, mapear os processos internos, e definir quais

integrações são necessárias. Para atuar na formação deste processo, é essencial contar com o suporte de plataformas tecnológicas [28].

A integração entre sistemas mantém a sincronização entre os sistemas sempre que ocorrer modificações de dados ou eventos, vinculando sistemas a nível funcional, permite criar uma integração orientada a eventos e mensagens de forma dinâmica e extremamente adaptáveis com transferências de dados em alta velocidade. As integrações apresentam ações controladas por eventos, que ocorrem quando um acionador, ou evento, dispara um procedimento ou uma solicitação. Também conta com integrações de *APIs*, mapeamento de dados entre os sistemas para definir como estes dados serão transferidos, facilitando a exportação, agrupamento e análise destas informações [26] [8].

Há diversas pesquisas realizadas relacionadas a integração de sistemas em diversas áreas, como um exemplo tem-se o livro *'Handbook of Medical Image Computing and Computer Assisted Intervention'*, que apresenta processos, ferramentas e melhores práticas para desenvolvimento, teste e manutenção de software baseado em componentes, na área de integração de sistemas de intervenção assistida por computador (CAI, pela sigla em inglês), discute sobre as diversas considerações de design, opções para estruturas de aplicativos e *middleware* para lidar com a comunicação entre os componentes, como o *Robot Operating System(ROS)*, para sistemas CAI que integram robôs e o *OpenIGTLink*, para lidar com integração de imagens médicas [15]

Da mesma forma, existem diversos softwares e plataformas especializadas que realizam integrações entre sistemas, como um exemplo, é válido mencionar a plataforma *TensorFlow Extended(TFX)*, baseada em *TensorFlow* e implementada no *Google*, própria para implantação e gerenciamento de *pipelines* de produção de aprendizado de máquina [29]. Esta realiza a integração de componentes, como por exemplo para gerar modelos baseados em dados de treinamento, módulos para analisar e validar dados e modelos e infraestrutura para servir modelos em produção, tudo em uma unica plataforma, padroniza os componentes, simplifica a configuração da plataforma e reduz o tempo de produção da ordem de meses para semanas, ao mesmo tempo que proporciona estabilidade da plataforma, minimizando interrupções [21]. Também é importante citar a plataforma de integração *IBM Cloud Pak® for Integration*, que apresenta recursos para aumentar a velocidade e a qualidade do sistema, com gerenciamento de *API*, integração de aplicativos, *streaming* de eventos e outras funcionalidades, além de possuir conectores inteligentes pré-construídos e recursos de automação, como integração de dados com tecnologia de IA, processamento de linguagem natural(PNL, pela sigla em inglês) e ferramentas de baixo código [8].

3. Objetivos

Este projeto tem como objetivo propor uma arquitetura completa para preparação de dados agro meteorológicos, com um conjunto reduzido de ferramentas de código aberto que seja completo o suficiente para realizar a coleta, limpeza e tratamento, de forma integrada, dos dados convencionais simples, como números, textos curtos e outros, e dados complexos, particularmente dados espaciais e mapas, aplicados em tarefas de aprendizado de máquina destinadas a agricultura.

4. Procedimentos metodológicos/Métodos e técnicas

Entende-se que para o desenvolvimento do projeto seja necessário o conhecimento de meios de coleta de dados e aprendizado profundo sobre preparação dos dados e um levantamento sobre as melhores técnicas e ferramentas de código aberto para serem aplicadas nas diferentes etapas do processo.

Com isso, o primeiro passo será destinado a exploração e aplicação de alternativas de uso de ferramentas para coletas de dados, os quais incluem dados convencionais e dados espaciais. Para esta etapa, deve ser analisado de modo aprofundado, ferramentas de ingestão de dados que se ajustem a dados georreferenciados, como o *Apache Kafka*, que proporciona o *Kafka Connect* para realizar conexões com aplicativos externos em tarefas de importação e exportação de dados, *Apache Nifi*, uma ferramenta projetada especificamente para automatizar grandes fluxos de dados entre sistemas [18] e outras ferramentas concorrentes. Neste processo, espera-se que os dados ingeridos sejam armazenados em um banco de dados relacional, como o *PostgreSQL*, e um banco de dados espacial, como o *PostGIS*.

Para o desenvolvimento de um conjunto de *pipelines* de execução, esta etapa terá como foco a utilização de ferramentas apropriadas para gestão de *workflow*, como, por exemplo, o *Apache Airflow* e o *Prefect*, que se beneficiam dos Gráficos Acíclicos Dirigidos, mais conhecidos como *DAGs*, para gerenciar *workflows* e *pipelines* de dados e agendar trabalhos em vários servidores [10].

O próximo passo, na construção da arquitetura, será destinado a implementação das transformações, filtragem e limpeza dos dados ingeridos. Para isso, deve-se selecionar ferramentas de geoprocessamento, qualificadas para lidar com grandes quantidades de dados e de diferentes fontes. Tais ferramentas, deve realizar combinações de processos variados de preparação de dados, como a biblioteca *GeoPandas* que disponibiliza operações espaciais em tipos geométricos e a biblioteca de abstração de dados geoespaciais *GDAL*, que proporciona leitura, gravação e manipulação de formatos de dados geoespaciais raster e vetoriais. Também é possível citar as funcionalidades do *QGIS*, que permite a edição e análise de dados georreferenciados. Há a possibilidade, de outras bibliotecas e funcionalidades, serem adicionadas ao projeto no decorrer de seu desenvolvimento, para tratar de dados mais específicos.

No processo de engenharia de *features*, será aplicado os dados tratados nas etapas anteriores em ferramentas de *features stores* para serem armazenados e processados para reutilização, compartilhamento e desenvolvimento de modelos de aprendizado de máquina na arquitetura formulada. Para isso será utilizado a *feature store Hopsworks* para abranger as diferentes fontes de dados e ingerir features e a *feature store Feast* para que seja possível conectar em diferentes armazenamentos de dados *online/offline* e ser executado em qualquer plataforma desenvolvida no projeto.

Por fim, compreende-se que deve ser feita a integração das ferramentas selecionadas, e testes com *pipelines* e casos de uso variados, ajustar as entradas e saídas das ferramentas utilizadas, assim como armazenar e analisar os dados resultantes das diversas transformações aplicadas. Para finalizar a arquitetura, será desenvolvido um sistema com o uso de um servidor web como o *Apache2* ou *Node.JS* para a apresentação dos resultados obtidos após os tratamentos aplicados nos dados ingeridos. Esta etapa tem como

seguimento a criação de roteiros práticos relacionados ao uso da arquitetura proposta, contemplando as melhores práticas indicadas para cada uma das ferramentas.

5. Cronograma de Execução

Nesta seção, são listadas as atividades descritas na seção anterior. Também é apresentado na tabela 1 a seguir, o cronograma de execução das atividades.

Atividades:

1. Levantamento bibliográfico sobre ingestão de dados, gestão de *workflow*, engenharia de *features* e geoprocessamento;
2. Preparação do ambiente de trabalho e realização de testes de funcionalidades das ferramentas selecionadas;
3. Configuração e integração da ferramenta de coleta e ingestão de dados com a ferramenta para *workflows* e com os bancos de dados, relacional e espacial;
4. Definição e implementação de técnicas e tarefas de geoprocessamento a serem aplicadas ao conjunto de dados;
5. Determinação de *pipelines* de execução, realização de testes com os casos de uso selecionados e analise dos métodos e resultados obtidos;
6. Aplicação de engenharia de *features* no conjunto de dados para a geração de modelos de aprendizado de máquina e desenvolvimento da apresentação dos dados em um servidor web;
7. Escrita TCC (versão preliminar);
8. Escrita TCC (versão banca examinadora).

Tabela 1. Cronograma de Execução

	ago	set	out	nov	dez	jan	fev	mar	abr
Atividade 1	x								
Atividade 2	x	x							
Atividade 3			x						
Atividade 4			x	x	x				
Atividade 5					x	x			
Atividade 6						x	x		
Atividade 7		x	x	x	x				
Atividade 8						x	x	x	x

6. Contribuições e/ou Resultados esperados

Espera-se que os seguintes resultados sejam alcançados no projeto:

- Estabelecimento de uma arquitetura geral para coleta e preparação de dados, constituída somente de ferramentas de código aberto e com um alto nível de compatibilidade entre os meios selecionados;
- Apresentação de um conjunto de casos de uso para implementação de *pipelines* para viabilizar o desempenho e o potencial de utilização da arquitetura selecionada para aplicações em situações reais de tarefas relacionadas a suporte de aprendizado de máquina para aplicações na agricultura;
- Contribuição relevante para o conhecimento com relação a manipulação de dados, especificamente em coletas e preparações de dados direcionado a aplicações em ciência de dados e inteligências artificiais.

7. Espaço para assinaturas

Londrina, *18 de Setembro de 2023*.

_____	_____
Aluno	Orientador

Referências

[1] Amazon. Amazon sagemaker data wrangler. Disponível em: https://aws.amazon.com/pt/sagemaker/data-wrangler/. Acesso em: 09 ago 2023., 2023.

[2] Amazon. What is data preparation? Disponível em: https://aws.amazon.com/pt/what-is/data-preparation/. Acesso em: 12 set 2023., 2023.

[3] Amazon. What is feature engineering? Disponível em: https://aws.amazon.com/what-is/feature-engineering/. Acesso em: 15 set 2023., 2023.

[4] Lefteris Benos, Aristotelis C. Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11), 2021.

[5] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001.

[6] Eduardo Darrazão, Victor Amorim, Krerley Oliveira, and Luiz Gomes-Jr. Engenharia e avaliação de features para extração de informação em notas fiscais. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 80–89, Porto Alegre, RS, Brasil, 2023. SBC.

[7] The Economist. The world's most valuable resource is no longer oil, but data. Disponível em: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. Acesso em: 21 jun 2023., 2017.

[8] IBM Cloud Education. What is an integration platform? do i need one? Disponível em: https://www.ibm.com/blog/integration-platform/. Acesso em: 13 set 2023., 2021.

[9] Embrapa. Avanço da ciência de dados e big data, inteligência artificial, aprendizado de máquina e cooperativas de dados. Disponível em: https://www.embrapa.br/visao-de-futuro/agrodigital/sinal-e-tendencia/avanco-da-ciencia-de-dados-e-big-data-inteligencia-artificial-aprendizado-de-maquina-e-cooperativas-de-dados. Acesso em: 05 jul 2023.

[10] Shubhnoor Gill. 7 best airflow alternatives for 2023. Disponível em: https://hevodata.com/learn/airflow-alternatives/. Acesso em: 11 set 2023., 2023.

[11] Google. What is machine learning? Disponível em: https://cloud.google.com/learn/what-is-machine-learning. Acesso em: 11 set 2023., 2023.

[12] Red Hat. What is integration? Disponível em: https://www.redhat.com/en/topics/integration/what-is-integration. Acesso em: 12 set 2023., 2017.

[13] IBM. O que é machine learning? Disponível em: https://www.ibm.com/br-pt/topics/machine-learning. Acesso em: 11 set 2023., 2023.

[14] D.D. Kühn. *Pesquisa e Análise de Dados: problematizando o rural e a agricultura numa perspectiva científica (DERAD604)*. Série Ensino, Aprendizagem e Tecnologias - UFRGS. PLAGEDER, 2017.

[15] Andras Lasso and Peter Kazanzides. Chapter 35 - system integration. In S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger, editors, *Handbook of Medical Image*

Computing and Computer Assisted Intervention, The Elsevier and MICCAI Society Book Series, pages 861–891. Academic Press, 2020.

[16] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8), 2018.

[17] Elian Luz, Camilla Silva, and Daniela Claro. Engenharia de features linguísticas para classificação de triplas relacionais. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 381–388, Porto Alegre, RS, Brasil, 2021. SBC.

[18] Ishwarya M. Top data ingestion tools in 2023. Disponível em: https://hevodata.com/learn/data-ingestion-tools/. Acesso em: 11 set 2023., 2023.

[19] Farhad Malik. End to end guide for machine learning project. Disponível em: https://medium.com/fintechexplained/end-to-end-guide-for-machine-learning-project-146c288186dc. Acesso em: 10 set 2023., 2018.

[20] Microsoft. What is azure databricks? Disponível em: https://learn.microsoft.com/pt-br/azure/databricks/introduction/. Acesso em: 09 ago 2023., 2023.

[21] Akshay Naresh Modi, Chiu Yuen Koo, Chuan Yu Foo, Clemens Mewald, Denis M. Baylor, Eric Breck, Heng-Tze Cheng, Jarek Wilkiewicz, Levent Koc, Lukasz Lew, Martin A. Zinkevich, Martin Wicke, Mustafa Ispir, Neoklis Polyzotis, Noah Fiedel, Salem Elie Haykal, Steven Whang, Sudip Roy, Sukriti Ramesh, Vihan Jain, Xin Zhang, and Zakaria Haque. Tfx: A tensorflow-based production-scale machine learning platform. In *KDD 2017*, 2017.

[22] Oracle. Build a secure oci data integration environment with pre-built tasks from templates. Disponível em: https://docs.oracle.com/en/solutions/oci-data-integration/. Acesso em: 09 ago 2023., 2023.

[23] Gil Press. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. Disponível em: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=28d139436f63. Acesso em: 12 set 2023., 2016.

[24] Jon Reilly. End to end machine learning workflow. Disponível em: https://www.akkio.com/post/end-to-end-machine-learning-workflow. Acesso em: 12 set 2023., 2022.

[25] Masahiro Ryo. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, 6:257–265, 2022.

[26] Sap. What is application integration? Disponível em: https://www.sap.com/brazil/products/technology-platform/what-is-enterprise-integration/application-integration.html. Acesso em: 13 set 2023., 2023.

[27] Sciforce. Machine learning in agriculture: Applications and techniques. Disponível em: https://medium.com/sciforce/machine-learning-in-agriculture-applications-and-techniques-6ab501f4d1b5. Acesso em: 05 jul 2023., 2019.

[28] Sydle. Systems integration: Learn the kinds, challenges, and importance. Disponível em: https://www.sydle.com/blog/systems-integration-6140d39a84679b13bf127a93. Acesso em: 12 set 2023., 2022.

[29] TensorFlow. Tfx is an end-to-end platform for deploying production ml pipelines. Disponível em: https://www.tensorflow.org/tfx. Acesso em: 13 set 2023., 2023.

[30] Iwan van Beurden. The meaning of tool integration. Disponível em: https://www.exida.com/blog/the-meaning-of-tool-integration. Acesso em: 12 set 2023., 2016.

[31] Rebeca Vickery. A beginner's guide to end to end machine learning. Disponível em: https://towardsdatascience.com/a-beginners-guide-to-end-to-end-machine-learning-a42949e15a47. Acesso em: 10 set 2023., 2021.

[32] Yuan Yuan, Lei Chen, Huarui Wu, and Lin Li. Advanced agricultural disease image recognition technologies: A review. *Information Processing in Agriculture*, 9(1):48–59, 2022.