UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

BRUNO FAUSTINO AMORIM

UMA INVESTIGAÇÃO DOS DESAFIOS NO CICLO DE VIDA DO APRENDIZADO DE MÁQUINA E A IMPORTÂNCIA DO MLOPS: UM SURVEY

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

BRUNO FAUSTINO AMORIM

UMA INVESTIGAÇÃO DOS DESAFIOS NO CICLO DE VIDA DO APRENDIZADO DE MÁQUINA E A IMPORTÂNCIA DO MLOPS: UM SURVEY

AN INVESTIGATION OF CHALLENGES IN THE MACHINE LEARNING LIFECYCLE AND THE IMPORTANCE OF MLOPS: A SURVEY

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Francisco Carlos Monteiro Souza
Coorientador: Prof. Dr. Alinne Souza

DOIS VIZINHOS
2022

BRUNO FAUSTINO AMORIM

UMA INVESTIGAÇÃO DOS DESAFIOS NO CICLO DE VIDA DO APRENDIZADO DE MÁQUINA E A IMPORTÂNCIA DO MLOPS: UM SURVEY

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 11/novembro/2022

Francisco Carlos Monteiro Souza
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Rodolfo Adamshuk Silva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

André Roberto Ortoncelli
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

DOIS VIZINHOS
2022

AGRADECIMENTOS

Agradeço por todo o apoio dado por minha família, amigos, professores, orientadores e a todos os profissionais participantes dessa pesquisa. Além disso, agradeço especialmente aos professores orientados Prof. Dr. Francisco Carlos Monteiro Souza e a Prof. Dr. Alinne C. Corrêa Souza, que foram fundamentais para a conclusão desse trabalho, por toda a ajuda e dedicação durante o seu desenvolvimento.

RESUMO

Muito recentemente, foi dada uma atenção considerável ao desenvolvimento de áreas de inteligência artificial e ciência de dados. Isso foi impulsionado pelos avanços científicos e pelo número crescente de softwares e serviços que estão popularizando técnicas e algoritmos de aprendizado de máquina e levando pessoas com menos conhecimento em áreas como estatística e matemática a criar seus modelos preditivos. Como resultado, o campo de aprendizado de máquina deixou de ser apenas acadêmico e passou a despertar o interesse de empresas de diferentes domínios. Esses eventos levaram ao surgimento de várias ferramentas, como Scikit-Learn, Tensorflow, Keras, Pycaret e um vasto número de serviços de Aprendizado de Máquina baseados em nuvem, acelerando o tempo de desenvolvimento de modelos preditivos em velocidades jamais vistas. No entanto, muitos desafios permanecem na operacionalização e manutenção de produtos centrados no Aprendizado de Máquina, fazendo com que muitas iniciativas empresariais com o seu uso sejam frustradas. Nesse cenário, a experiência prática mostra que o Aprendizado de Máquina é apenas uma fatia de um conjunto mais extenso de práticas e tecnologias necessárias para construir soluções nessa área. Neste artigo, o objetivo principal é identificar os desafios enfrentados atualmente pelos cientistas de dados no desenvolvimento de produtos centrados em Aprendizado de Máquina e como as Operações de Aprendizado de Máquina pode ajudar a vencê-los. Para tanto, foi realizada uma pesquisa que coletou respostas de 66 profissionais brasileiros em ciência de dados. A partir dos desafios identificados, foi explorada a importância das práticas de Operações de Machine Learning como parte integrante do ciclo de vida do Aprendizado de Máquina. Por fim, este trabalho contribui para preencher a lacuna em Operações de Aprendizado de Máquina nas atividades diárias que envolvem ciência de dados e o avanço desse campo de pesquisa no Brasil.

Palavras-chave: palavra; Aprendizado de Máquina; MLOps.

ABSTRACT

Quite recently, considerable attention has been paid to developing artificial intelligence and data science areas. This has been driven by scientific advances and the growing number of software and services that are popularizing machine learning techniques and algorithms and driving people with less knowledge in areas such as statistics and mathematics to create their predictive models. As a result, the machine learning field is no longer only scientific and has aroused the interest of companies from different domains. These events led to the emergence of multiple tools such as Scikit-Learn, Tensorflow, Keras, Pycaret, and a vast number of cloud-based machine learning services that provide an acceleration in the development of predictive models at speeds never seen. However, many challenges remain in operationalizing and maintaining machine learning-centered products, making many business initiatives frustrated. In this scenario, practical experience shows that machine learning is only a slice of a more extensive set of practices and technologies necessary to build solutions in this area. In this paper, the main goal is to identify the challenges currently faced by data scientists in developing Machine Learning-centric products and how Machine Learning Operations can support overcoming them. For this purpose, a survey was conducted that collected answers from 66 Brazilian professionals in data science. From the challenges identified, the importance of Machine Learning Operations practices as an integrated part of the Machine Learning lifecycle was explored. Finally, this work contributes to filling the gap in Machine Learning Operations in daily activities involving data science and advancing this research field in Brazil.

Keywords: Machile Learning; MLOps.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
MLOps	Machine Learning Operations

SUMÁRIO

1 INTRODUÇÃO

As atividades do ciclo de vida de Aprendizado de Máquina são de natureza exploratória, iterativa e requerem muita experimentação. Durante esse ciclo, existe um árduo trabalho entre a primeira fase de especificação dos requisitos e a disponibilização dos modelos preditivos para os usuários ou sistemas finais. Grande parte desse processo é centrado na manipulação de dados, já que esses dados podem sofrer diversos tipos de variações, seja por alterações nas aplicações que os geram, problemas de qualidade, ou mudança no comportamento de determinados eventos (AMERSHI et al., 2019).

Fundamentado nisso, esse trabalho objetiva explorar os desafios enfrentados pelos times de ciência de dados durante todo o ciclo de vida de AM, contemplando desde o desenvolvimento até a implantação de modelos em produção. A partir desse estudo, abordou-se como o MLOps[1] pode otimizar o ciclo de vida de AM, por meio de práticas, automação de processos e estrutura. Práticas de engenharia de software podem ser aplicadas para o desenvolvimento, implantação, manutenção e monitoramento dos modelos como parte das aplicações empresarias, isso objetiva acelerar as entregas de modelos, otimizando o trabalho das equipes de ciência de dados.

Este tema de pesquisa também é analisado em trabalhos como (MÄKINEN et al., 2021), (PALEYES; URMA; LAWRENCE, 2020) e (SOUZA, 2021). A literatura existente estuda a aplicação de MLOps nas atividades de Aprendizado de Máquina com o intuito de ajudar a vencer os desafios diários enfrentados pelos times de ciência de dados (MÄKINEN et al., 2021). Porém, há uma lacuna na abrangência dos aspectos práticos observados, já que na indústria a importância do MLOps ainda é vagamente explorada, fazendo com que muitas empresas enfrentem problemas em diversas fases do ciclo de vida de Aprendizado de Máquina (KREUZBERGER; KÜHL; HIRSCHL, 2022).

As principais contribuições deste estudo podem ser sumarizadas como: *i*) um *survey* para identificar possíveis desafios no ciclo de vida de Aprendizado de Máquina; *ii*) síntese dos desafios identificados por meio de agrupamentos de atividades do ciclo de vida; e *iii*) a percepção sobre a área de MLOps na visão de profissionais da ciência de dados. Como principais observações, foi identificado que 90% dos participantes já realizam a implantação de modelos de Aprendizado de Máquina em ambiente produtivo. Contudo, nota-se que os times de ciência de dados estão empenhando tempo e esforço considerável com atividades além do desenvolvimento de modelos preditivos. As operações de Aprendizado de Máquina também carecem de profissionais de Engenharia.

O restante desse artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. Na Seção 3, os detalhes do *survey* são apresentados. A Seção 4 discute os resultados obtidos e a importância do MLOps, e a Seção 5 encerra este trabalho apresentando as conclusões.

[1] https://blog.nvidia.com.br/2020/09/08/o-que-e-mlops/

2 ASPECTOS CONCEITUAIS

Nesta Seção são apresentados os conceitos relacionados ao Ciclo de vida de Aprendizado de Máquina, bem como MLOps.

2.1 Ciclo de vida de Aprendizado de Máquina

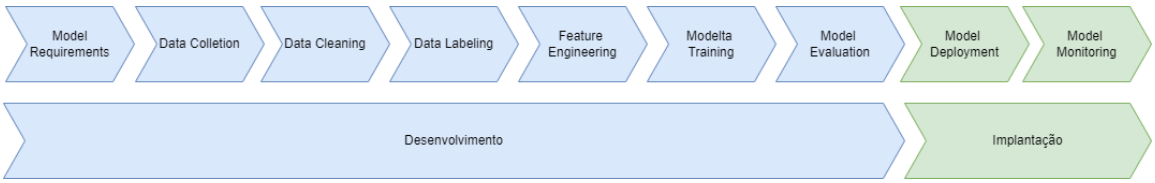
Com diferentes formas de representação, encontrou-se na literatura várias propostas de fluxos para o ciclo de vida de Aprendizado de Máquina [1] [2] (AMERSHI et al., 2019) (MORABITO, 2015) (ASHMORE; CALINESCU; PATERSON, 2021). A construção de modelos de AM é um processo constituído por diversas etapas, onde durante a busca do modelo generalizável com melhor desempenho, várias dessas etapas são reiteradas, tornando o processo de desenvolvimento altamente iterativo e exploratório.

Como embasamento teórico foi utilizado um estudo de caso da Microsoft (AMERSHI et al., 2019), que estruturou o ciclo de vida de Aprendizado de Máquina em nove atividades, conforme pode ser visto na Figura 1. Para o nosso atual contexto, essas atividades foram agrupadas em dois conjuntos: 1) Desenvolvimento: contendo todas as etapas referentes ao desenvolvimento do modelo, desde o mapeamento de requisitos até a avaliação do modelo. 2) Implantação: contendo as etapas de implantação e monitoramento do modelo.

A primeira etapa é a análise dos requisitos do modelo, onde são definidos quais recursos podem ser implementados com AM, recursos úteis para um determinado produto existente ou novo e tipos de modelos mais apropriados para o problema (AMERSHI et al., 2019). Na coleta de dados os cientistas obtêm os dados necessários para o trabalho em questão. Nessa fase, surgem os desafios, principalmente relacionados as dimensões de variedade e volumetria dos dados (MORABITO, 2015). Durante a preparação e transformação, é realizada a limpeza e transformação dos dados brutos para análise [3]. A fase de engenharia de recursos é utilizada para selecionar e transformar variáveis. Isso envolve a criação, transformação, extração e seleção de recursos [4]. Na fase de treinamento do modelo, os modelos escolhidos são treinados e ajustados sobre os dados transformados (AMERSHI et al., 2019). A avaliação do modelo é responsável por determinar se o desempenho do modelo é satisfatório para atingir as metas definidas pelo negócio [5]. Em seguida, o modelo poder ser implantado e disponibilizado para os sistemas ou usuários finais (ASHMORE; CALINESCU; PATERSON, 2021).

[1] https://www.oracle.com/a/ocom/docs/data-science-lifecycle-ebook.pdf
[2] https://docs.aws.amazon.com/pt_br/wellarchitected/latest/machine-learning-lens/feature-engineering.html
[3] https://www.oracle.com/a/ocom/docs/data-science-lifecycle-ebook.pdf
[4] https://docs.aws.amazon.com/pt_br/wellarchitected/latest/machine-learning-lens/feature-engineering.html
[5] https://docs.aws.amazon.com/pt_br/wellarchitected/latest/machine-learning-lens/feature-engineering.html

Figura 1 – Ciclo de vida de Aprendizado de Máquina.



Fonte: Adaptado de AMERSHI et al. (2019).

2.2 Operações de Aprendizado de Máquina (MLOps)

MLOps tem por objetivo unificar o desenvolvimento de sistemas de AM (*ML*) e a operação de sistemas de AM (*Ops*) [6]. Apesar dos cientistas de dados possuírem ferramentas e habilidades para implementação e treino de modelos preditivos, a implantação e sustentação desses modelos em ambiente produtivo requer outros conhecimentos técnicos fora do domínio de ciência de dados. O maior desafio está na criação de sistemas de AM operantes continuamente em produção (SCULLEY et al., 2015) [7].

Apesar do rápido crescimento no desenvolvimento dos modelos, as empresas possuem muitas dificuldades para operacionaliza-los, fazendo com que grande parte dos modelos criados nunca cheguem em produção (HECHT, 2019) (WIGGERS, 2019). Como a criação de produtos ou serviços baseados em AM envolve outros componentes além dos modelos preditivos, cientistas de dados não são capazes de sozinhos realizar todo o trabalho contido em todo o ciclo de vida e AM, sendo necessária uma equipe multidisciplinar e conhecimentos especializados em outras áreas de conhecimento como Engenharia de Dados e Engenharia de Software.

Além disso, a construção de sistemas de AM apresenta um conjunto de particularidades que não são tradicionalmente vistas em engenharia de software. Contrariamente ao desenvolvimento de software tradicional, as aplicações de AM acrescentam bases de dados que mudam constantemente e afetam a maneira de desenvolver e manter esses tipos de sistemas.

[6] https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[7] https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

3 TRABALHOS RELACIONADOS

Embora a literatura sobre MLOps seja ampla, alguns *surveys* fornecem uma visão das dificuldades durante os estágios do ciclo de vida de AM. O trabalho de makinen2021needs aborda uma pesquisa com 331 cientistas de dados, em que foi identificada a importância do MLOps no contexto das atividades diárias desses profissionais. Neste trabalho também foi levantado aspectos profissionais e empresarias dos participantes, além de abranger um número maior de elementos focados em todo o ciclo de vida de AM. Por fim, os resultados apontam que os cientistas de dados estão expandindo suas atuações para tarefas relacionadas a infraestrutura e implantação de modelos de AM.

Em (PALEYES; URMA; LAWRENCE, 2020) foi conduzida uma pesquisa que descreve obstáculos que surgem durante as fases do ciclo de vida de implantação de modelo de AM, considerando os estágios de gerenciamento de dados, aprendizado de modelo, verificação de modelo e implantação de modelo. Os principais obstáculos identificados estão relacionados à reutilização de código e anti-padrões de Engenharia de Software. É importante destacar que foram explorados aspectos como ética e segurança da informação, que devem ser considerados durante o desenvolvimento dos modelos.

No estudo de (SOUZA, 2021) foi investigado a complexidade na adoção de ferramentas e práticas de MLOps. Como principal contribuição, por meio de experimentos, foram apresentadas como algumas ferramentas de código aberto podem ser aplicadas na implantação do MLOps.

O estudo de kim2017data apresenta uma pesquisa com 793 cientistas de dados da Microsoft investigando os problemas enfrentados, práticas recomendadas durante a construção de modelos de AM e o atual contexto de trabalho dos cientistas de dados. Os resultados apontam uma tendência desses profissionais atuarem em funções diretamente inseridas nas equipes de software. Além disso, o trabalho também mostra que funcionários contratados para outras posições na empresa assumiram tarefas de ciência de dados como parte do seu trabalho. O estudo é limitado ao contexto corporativo da Microsoft, logo algumas conclusões podem não ser representativas para empresas de outros setores e de diferentes outros portes.

Por fim, o trabalho de (JAIN et al., 2020) destaca a importância do controle da qualidade dos dados em tarefas de AM, por meio de métricas de qualidade de dados como desequilíbrio de classes e homogeneidade dos dados que podem ser utilizadas e monitoradas como parte do ciclo de vida de AM.

Apesar da abrangência dos trabalhos apresentados em contemplar as dificuldades durante todos os estágios do ciclo de vida de AM, ainda existe uma lacuna na cobertura dos aspectos não técnicos. Entre elas, a percepção de valor vista pelas áreas de negócio, bem como a falta de informações sobre as composições e estruturas dos times de dados das empresas estudadas. Neste contexto, o presente artigo tem como objetivo preencher essa lacuna

e contribuir para a literatura existente, trazendo uma visão holística que investiga os atuais desafios do ciclo de vida de AM. Portanto, foi conduzido um estudo exploratório referente às adversidades práticas enfrentadas pelos profissionais atuantes na área de ciência de dados envolvendo a manipulação de dados.

4 ESTUDO EXPERIMENTAL

Esta seção detalha o *survey* conduzido para identificar os desafios enfrentados atualmente por profissionais que atuam na área de ciência de dados no desenvolvimento de produtos centrados em AM e como o MLOps pode ajudar a minimizá-los. O *survey* foi planejado seguindo o processo proposto por Kasunic (KASUNIC, 2005) para design efetivo de *surveys*. Além disso, foram utilizados direcionamentos descritos por Kitchenham e Pfleeger (A.; PFLEEGER, 2008).

4.1 Identificação dos objetivos de investigação e Questões de Pesquisa

O objetivo do *survey* consiste na identificação dos desafios no desenvolvimento de produtos utilizando AM e como a aplicação de MLOps pode miniminizar esses desafios. Neste contexto, os objetivos do *survey* foram especificados segundo o modelo *Goal-Question-Metric (GQM)* proposto por Basili e Weiss (BASILI; WEISS, 1984), conforme pode ser visto a seguir:

"Analisar **desenvolvimento de produtos centrados em AM** *com o propósito de* **identificar os desafios enfrentados e como o MLOps pode minimizar esses desafios** *do ponto de vista de* **profissionais que atuam na área de ciência de dados** *no contexto de* **ciclo de vida de AM**.*"*

A partir do objetivo, as seguintes Questões de Pesquisas (QPs) foram identificadas:

QP_1**: Qual perfil dos profissionais atuantes na área de ciência de dados e da empresa na qual trabalham?** Nesta QP, buscou-se identificar informações relacionadas à formação, cargo, experiência no cargo ocupado, nível (júnior, pleno ou sênior) do cargo ocupado e o nível de conhecimento sobre MLOps. Além disso, foram coletadas informações relacionadas à empresa como o tipo (nacional ou multinacional), número de funcionários, ramo, como a empresa se encontra atualmente em relação ao desenvolvimento e execução de modelos e quais profissionais compõem os times de dados.

QP_2**: Quais são as principais atividades desempenhadas pelos profissionais atuantes na área de ciência de dados nas empresas?** Esta QP busca identificar as atividades realizadas com mais frequência durante o desenvolvimento de modelos de AM, bem como a frequência dos tipos de dados manipulados. Além disso, investigou-se qual ambiente, linguagens e ferramentas são utilizadas para auxiliar o desenvolvimento de modelos de AM.

QP_3**: Quais desafios são enfrentados durante o ciclo de vida de AM?** Esta QP visa identificar os principais desafios relacionados à manipulação dos dados, ao desenvolvimento dos modelos, a não implantação de modelos em produção, bem como a implantação e o monitoramento. Além disso, nesta QP espera-se apresentar como as práticas de MLOps podem minimizar os desafios identificados no ciclo de vida.

4.2 Identificação do público-alvo e planejamento de amostragem

O público-alvo definido para participar do *survey* é composto por profissionais que atuam na área de ciência de dados que estejam trabalhando em empresas de diferentes ramos distribuídos geograficamente pelo Brasil. Após a definição do público-alvo, foi planejado o processo de obtenção de amostras. Nesta etapa considerou-se localizar profissionais que trabalham em qualquer nicho de negócio. Assim, a identificação desses profissionais foi realizada por meio de contato via e-mail e rede social.

4.3 Planejamento e escrita do questionário

O detalhamento do protocolo desenvolvido para a condução do *survey* pode ser visualizado no link *https://shre.ink/mNKt*. Além do protocolo, o questionário aplicado junto aos participantes pode ser acessado em: *https://bityli.com/fBqFghav*.

As 25 questões foram estruturadas em cinco seções: (i) contém uma apresentação do *survey* em que é descrito o objetivo do mesmo e o público-alvo; (ii) coleta informações que caracterizam os participantes; (iii) identifica informações relacionadas as atividades conduzidas pelos participantes; (iv) coleta informações a respeito dos desafios enfrentados pelos cientistas de dados durante o ciclo de vida de AM; e (v) obtém formas de como MLOps podem otimizar os processos de AM. O *survey* pode ser acessado em: *https://drive.google.com/file/d/1q8Ob2LIPKCY9Cf0lrbf2Nzm37iiOd7wF*.

4.4 Execução do survey piloto

Segundo Kasunic (KASUNIC, 2005), é fundamental a condução de um *survey* piloto, pois é possível detectar possíveis problemas existentes no mesmo, verificar se as perguntas são compreensíveis, se as perguntas certas foram feitas para atingir o objetivo, e quanto tempo os participantes levam para completarem o questionário. Para avaliar o *survey* foram aplicadas quatro questões abertas propostas por Hauck et. al (HAUCK; WANGENHEIM; WANGENHEIM, 2011), sendo: (1) O questionário contém tudo que é esperado para contemplar o seu objetivo?; (2) O questionário contém quaisquer informações não desejáveis ou desnecessárias ao contexto e objetivo da pesquisa?; (3) Você conseguiu compreender adequadamente as perguntas?; e (4) Existe algum erro ou inconsistência no questionário?. Além destas questões também foram consideradas as opiniões de especialistas (LITWIN, 1995).

Nesse contexto, um grupo de quatro participantes distribuídos entre formadas e não formadas, estagiando e trabalhando, foram convidadas por e-mail para participar do teste piloto. Essas participantes, foram selecionadas pelos critérios de disponibilidade e proximidade, participaram do estudo piloto respondendo às questões propostas por Hauck et. al (HAUCK; WANGENHEIM; WANGENHEIM, 2011) e enviaram *feedbacks* sobre o *survey*. A avaliação das participantes foi positiva, com sugestões para: reduzir o número de perguntas, deixar algumas

não obrigatórias; e incluir uma pergunta aberta.

Um grupo de profissionais foi convidado por e-mail para participar do estudo piloto. Esses profissionais foram escolhidos pelos critérios de disponibilidade e proximidade como o grupo onde esta pesquisa foi realizada. A avaliação dos profissionais foi positiva, com sugestões para: *(i)* reduzir o número de questões; *(ii)* deixar algumas perguntas como não obrigatórias; e *(iii)* incluir pelo menos uma pergunta aberta.

4.5 Coleta e Análise dos Dados

O questionário foi divulgado para os profissionais da área de ciência de dados via e-mail e a rede social LinkedIn[1] e ficou disponível no período de 29 de Junho a 5 de Agosto de 2022. Após a coleta dos dados, os mesmos foram analisados com o objetivo de verificar a consistência e completude das respostas(A.; PFLEEGER, 2008). Em seguida, foram realizadas análises quantitativas e qualitativas conforme o tipo de perguntas utilizadas na pesquisa. Para a interpretação dos dados quantitativos foi utilizada estatística descritiva; e análise de discurso e visualização de dados para a análise qualitativa.

[1] https://br.linkedin.com/

5 RESULTADOS

O *survey* obteve a contribuição de 66 profissionais atuantes na área de ciência de dados. Os resultados serão apresentados de acordo com as QPs, conforme apresentado na Seção 4.

5.1 Perfil dos profissionais atuantes na área de ciência de dados e das empresas (QP_1)

A pesquisa coletou respostas de profissionais de diferentes organizações, sendo 68% (45/66) empresas nacionais brasileiras e 31% (21/66) multinacionais. Com base na quantidade de funcionários, de acordo com o Sebrae [1], a maioria dessas empresas são de grande porte, sendo que 27% (18/66) possuem de 100 a 500 funcionários, e 54% (36/66) possuem mais de 500 funcionários. Dessas empresas, 19% (13/66) foram classificadas como empresas de Software/Internet e 24% (16/66) de Serviços Financeiros.

Sobre os profissionais participantes, 45% (30/66) possuem ensino superior completo, 27% (18/66) são mestres, 16% (11/66) são especialistas, 7% (5/66) são doutores e 3% (2/66) possuem ensino superior incompleto.

Uma classificação de quatro níveis foi criada para identificar o nível de maturidade das empresas na utilização de Aprendizado de Máquina nas operações. Para isso, realizou-se a seguinte pergunta: "Qual das frases abaixo mais se aproxima do momento atual da sua empresa?".

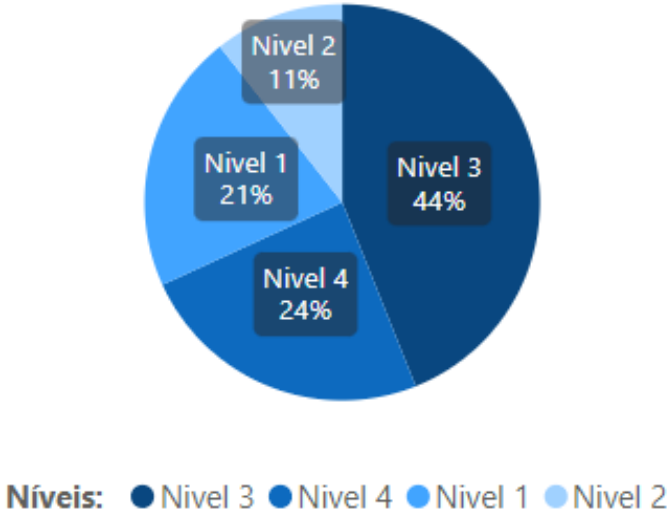
- **Nível 1**: a empresa está desenvolvendo seus primeiros modelos, porém ainda sem gerar valor para as frentes de negócio.
- **Nível 2**: as áreas de negócio já consomem os modelos, porém são executados localmente e não há monitoramento ou automações.
- **Nível 3**: a empresa executa seus modelos em ambiente produtivo em nuvem, com consumos via API ou *Batch*, porém ainda sem processos automatizados.
- **Nível 4**: a empresa executa seus modelos em ambiente produtivo em nuvem, e com processos como monitoramento, implantação e retreino automatizados.

A criação dessa classificação, em parte, foi inspirada em um trabalho realizado pelo Google [2].

A maioria das empresas está no nível 3, sendo empresas que já possuem modelos de AM em produção, porém ainda sem processos automatizados. Ademais, um percentual considerável de empresas está no nível 1, desenvolvendo seus primeiros modelos de AM, porém ainda sem a geração de valor para o negócio (Figura 2).

[1] https://www.sebrae.com.br/Sebrae/Portal%20Sebrae/UFs/SP/Pesquisas/MPE_conceito_empregados.pdf
[2] https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

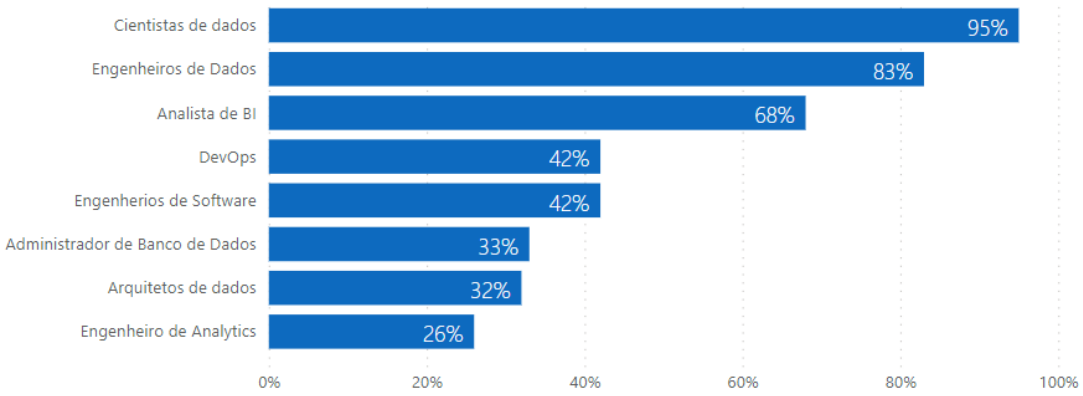
Figura 2 - Níveis de maturidade.



Fonte: Autoria própria (2022).

Os times de dados das empresas são compostos por diferentes perfis, observando que em sua maioria são 95% (63/66) formados por cientistas de dados, 83% (55/66) engenheiros de dados e 68% (45/66) analistas de BI (Figura 3). Apesar da maior parte das empresas serem de médio e grande porte, 37% (25/66) não possuem equipes dedicadas para a implantação de modelos de AM em produção.

Figura 3 - Composição dos times de dados.



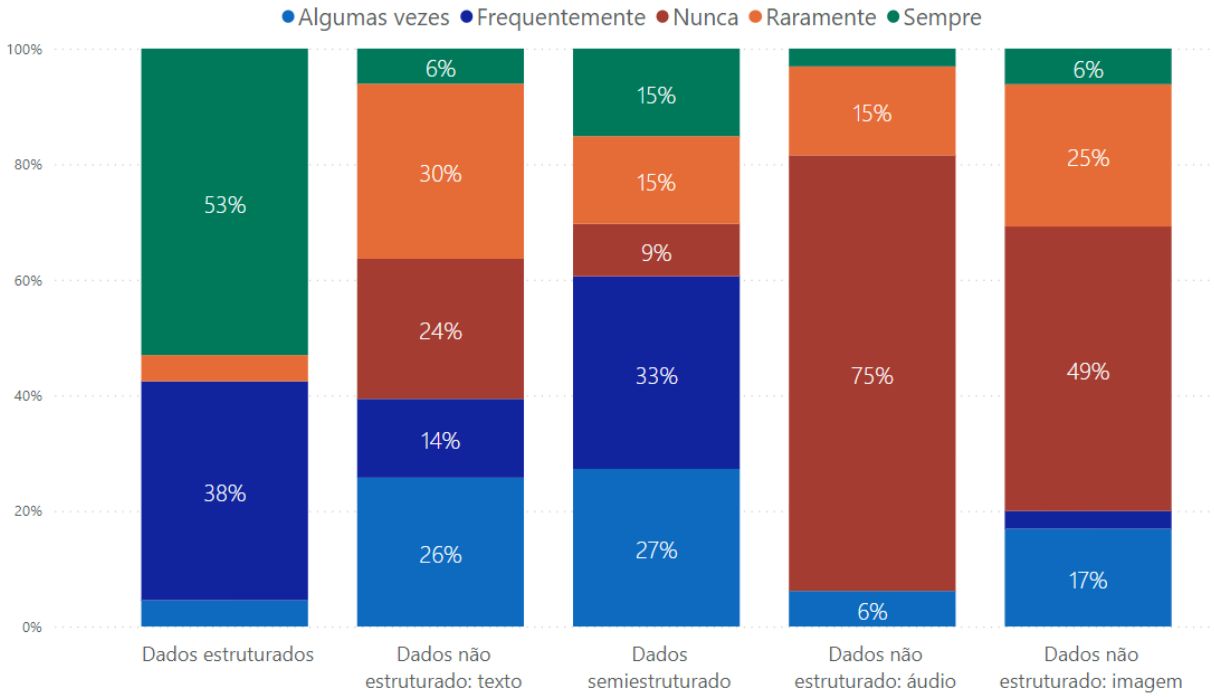
Fonte: Autoria própria (2022).

5.2 Principais atividades desempenhadas pelos profissionais atuantes na área de ciência de dados nas empresas (QP_2)

Para atingir o objetivo dessa seção, foram elencadas perguntas para identificar os vários aspectos que contemplam todas as atividades de desenvolvimento no ciclo de vida de AM, realizando perguntas relacionadas a diferentes domínios, desde o ambiente utilizado para o desenvolvimento dos modelos, incluindo ferramentas e tecnologias, até as atividades mais frequentes e maiores problemas e dificuldades enfrentados durante essa fase.

O primeiro aspecto a ser considerado é que apesar do aumento na variedade das informações, causado com o advento do *Big Data*, nota-se que os times de ciência de dados raramente trabalham com dados não estruturados de imagem ou áudio. Suas atuações concentram-se praticamente sempre com dados estruturados e semi-estruturados (Figura 4).

Figura 4 - Tipos de dados frequentemente manipulados.



Fonte: Autoria própria (2022).

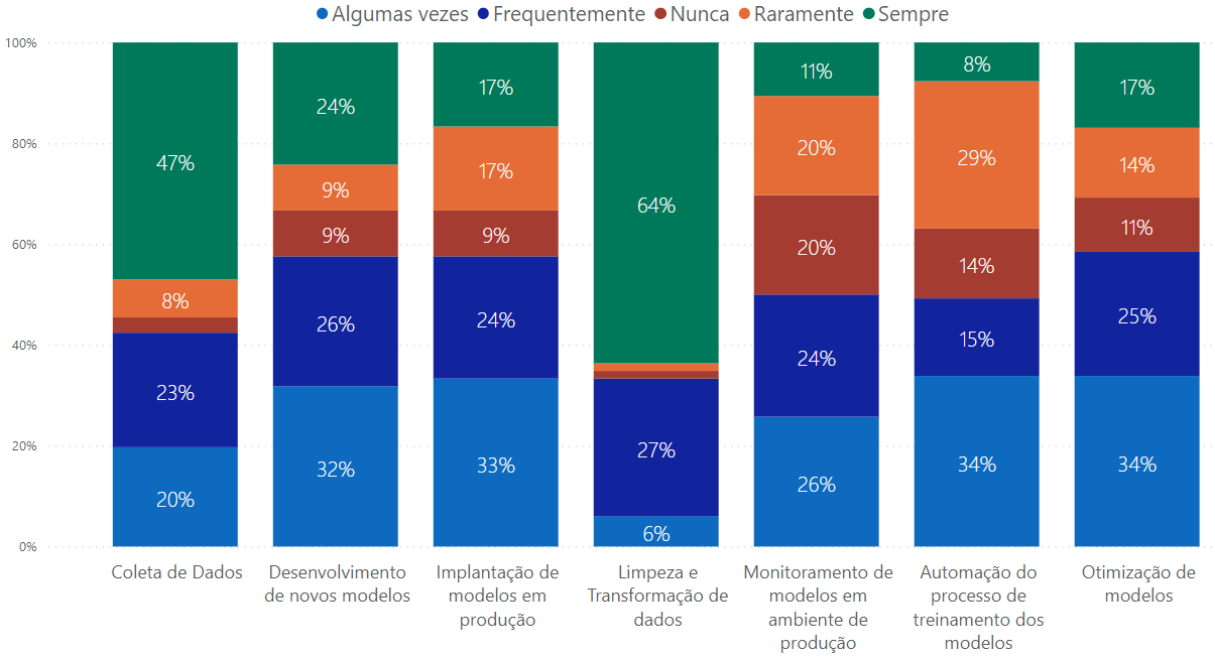
Sobre as atividades mais frequentes, a "limpeza e transformação dos dados"é a atividade apontada como mais recorrente pelos participantes. Isso não é surpresa, já que essa atividade é conhecidamente uma das mais comuns e trabalhosas nessa área. Houve também um alto índice de respostas apontando para atividades de coleta de dados.

A implantação de novos modelos em produção mostra-se como atividade de alta frequência. Apesar disso, dos 57 participantes que afirmam que suas empresas possuem modelos em produção, 45% (26/57) deles responderam que suas empresas tem até no máximo 5 modelos em produção e 19% (11/57) deles tem de 6 a 10 modelos em produção. Dada a quantidade relativamente baixa de modelos em produção, suspeita-se que os poucos modelos em operação trazem uma sobrecarga operacional considerável para os times envolvidos. Nesse cenário, o MLOps pode trazer ganhos significativos automatizando tarefas como implantação, retreino, e monitoramento dos modelos.

Outro ponto importante a se notar é que o desenvolvimento de novos modelos foi marcado por 57% (38/66) dos participantes como "Frequentemente"ou "Sempre". Relacionando essa informação a outros estudos como (WIGGERS, 2019), suspeita-se que boa parte dos modelos desenvolvidos não chegam até a produção (Figura 5). Isso também pode estar relacionado diretamente a aspectos de dificuldades encontradas no contexto de dados, como

será apresentado no decorrer do artigo.

Figura 5 - Atividades mais frequentes durante o desenvolvimento.



Fonte: Autoria própria (2022).

A respeito das ferramentas mais utilizadas durante o desenvolvimento, 87% (58/66) dos participantes utilizam Scikit-learn [3], 60% (40/66) XGBoost [4]. Apesar da baixa quantidade de manipulação de dados não estruturados, como áudio ou textos, nota-se um alto uso de ferramentas de Aprendizado Profundo como TensorFlow [5] e Keras [6], onde 53% (35/66) utilizam TensorFlow e 48% (32/66) Keras. Com isso, questiona-se se não seria possível utilizar algoritmos e ferramentas menos complexas para atingir os mesmos resultados, talvez tornando o processo de modelagem mais simples. Sobre o Aprendizado de Máquina Automatizado (*AutoML*), vemos que 22% (15/66) dos participantes utilizam ferramentas como Pycaret [7]. Por fim, nota-se que a minoria, isto é, 30% (20/66) dos profissionais participantes, utiliza ferramentas e pacotes R [8].

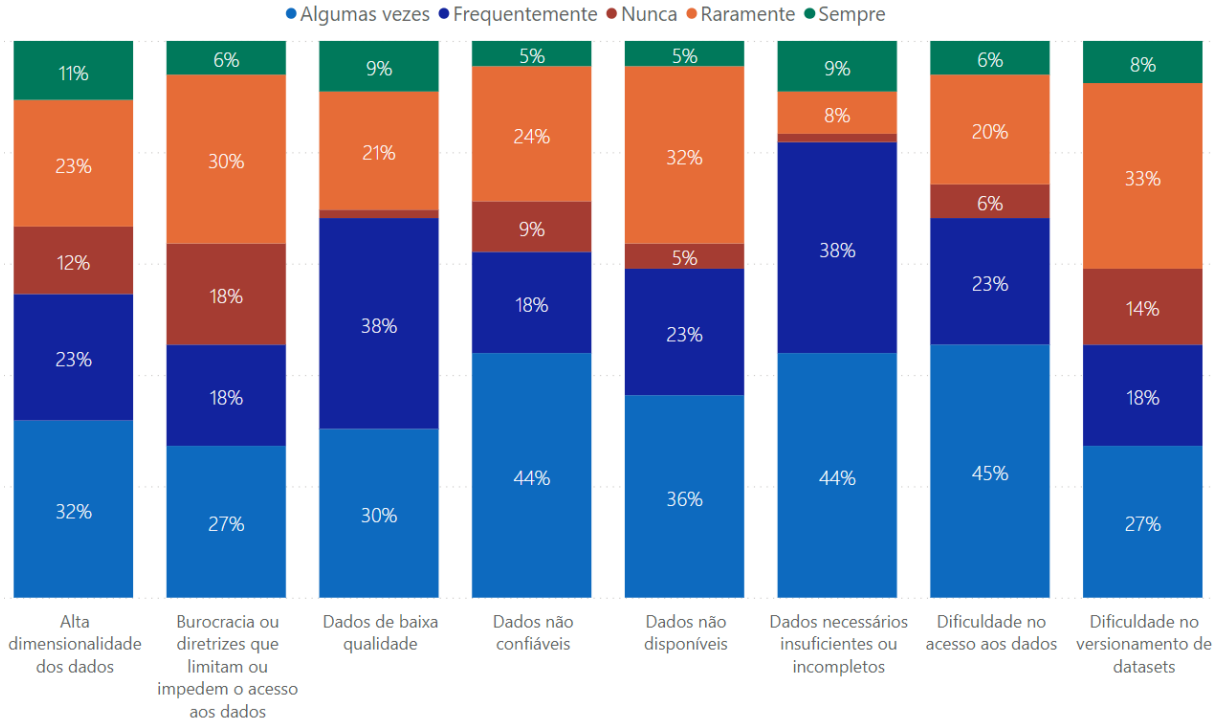
Quanto ao ambiente de trabalho, cerca de 63% (42/66) dos participantes desenvolvem seus modelos de AM em nuvem e 27% (18/66) em ambiente local ou *on-premisse*, ou seja, diretamente na infraestrutura de servidores da empresa. Para o desenvolvimento dos modelos várias linguagens de programação podem ser aplicadas, sendo Python utilizada em 98% (65/66) e R 28% (19/66). Além disso, linguagens como Java, Scala ou *SaaS* raramente são utilizadas. As respostas indicam que boa parte dos times de ciência de dados dificilmente sofrem com a falta de ferramentas ou conhecimento técnico durante o desenvolvimento dos modelos.

[3] https://scikit-learn.org/stable/
[4] https://xgboost.readthedocs.io/en/stable/
[5] https://www.tensorflow.org/
[6] https://keras.io/
[7] https://pycaret.org/
[8] https://www.r-project.org/

5.3 Desafios enfrentados durante o ciclo de vida de AM (QP_3)

Os resultados demonstram que há uma grande dificuldade com aspectos de qualidade dos dados. Afirmações de "dados necessários insuficiente ou incompletos", mostram-se em 81% (54/66) das vezes como uma das mais altas ocorrências quando somadas as 29 respostas marcadas como "Algumas vezes"e 25 respostas "Frequentemente". Nota-se também um alto índice de respostas afirmando dificuldades no acesso aos dados. O problema na qualidade dos dados se agrava quando é analisado o total marcado como "dados não confiáveis". Além disso, cerca de 53% (35/66) dos participantes relataram sofrer em algum nível com problemas no versionamento dos conjuntos de dados (*datasets*) (Figura 6).

Figura 6 - Problemas identificados na manipulação de dados.



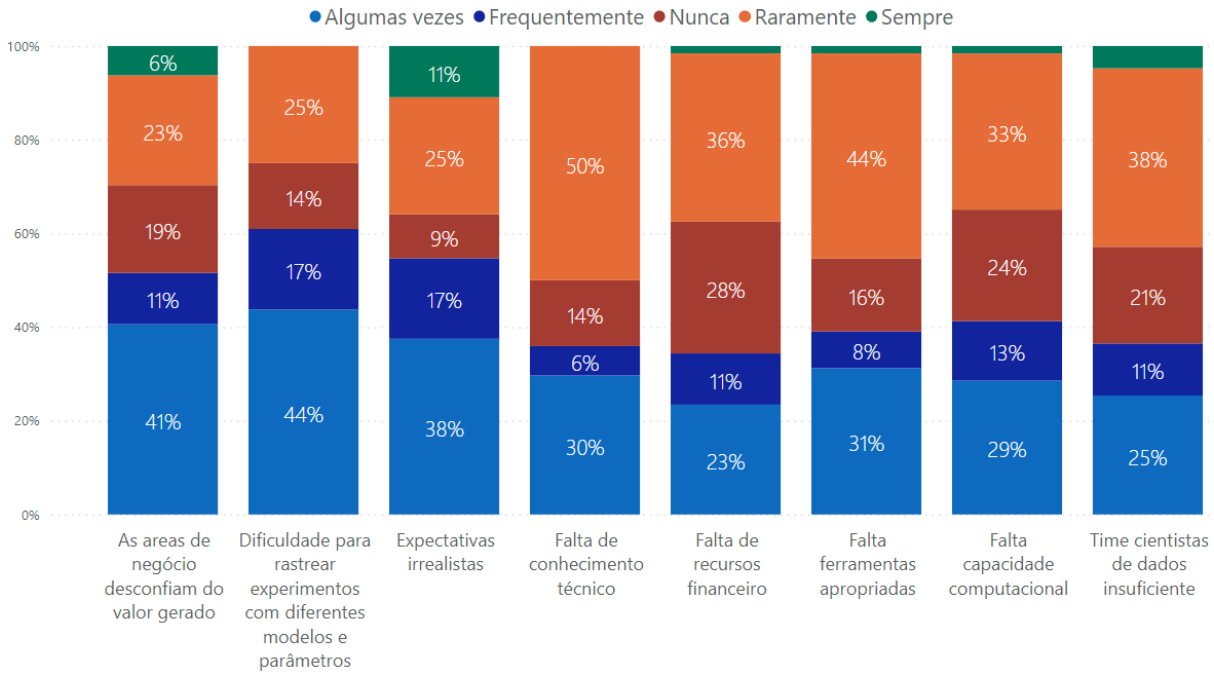
Fonte: Autoria própria (2022).

Um alto número de participantes afirmou ter problemas com o rastreamento de experimentos e parâmetros durante a fase de desenvolvimento (Figura 7). Além disso, a fim da ampliar a abrangência deste trabalho, cobrimos também aspectos não-técnicos. Grande parte das respostas demonstra que as áreas de negócio desconfiam do valor gerado com os modelos de AM (Figura 7). Com isso, a hipótese criada é que a falta de reprodutibilidade e falta de qualidade dos dados sejam um dos principais fatores que alimentam esse cenário de desconfiança. Outro importante aspecto é o alto nível de afirmações sobre expectativas irreais (Figura 7). A suspeita é que as áreas de negócio podem não compreender que a sua participação é importante para o processo de modelagem e definição de métricas. As expectativas também podem ser frustradas devido aos problemas de ausência de dados necessários e dados de baixa qualidade. As organizações parecem não ter consciência de que modelos matemáticos são

incapazes de gerar resultados positivos sem dados minimamente confiáveis. Muitas vezes as áreas de negócio criam a expectativa de que os modelos de AM são a saída para quase todos os problemas. Como resultado, isso pode gerar alta desconfiança.

Por fim, percebem-se poucas ocorrências apontando para a falta de recursos financeiros durante essa fase.

Figura 7 - Dificuldade no desenvolvimento dos modelos de AM.



Fonte: Autoria própria (2022).

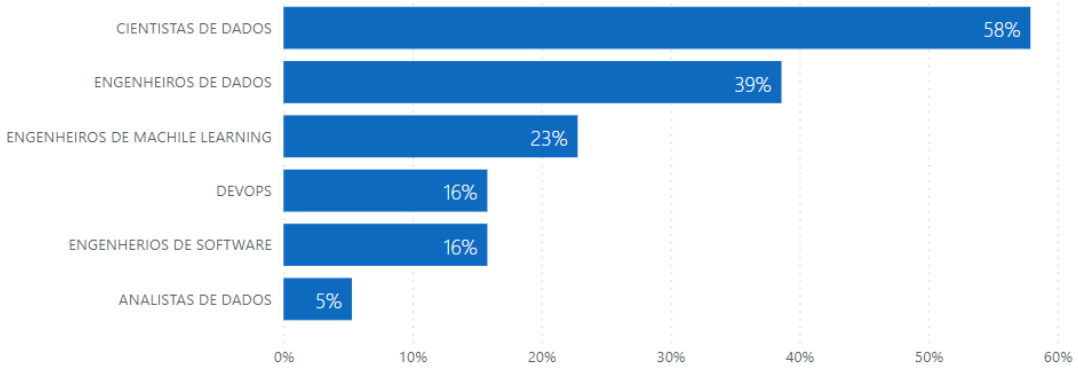
5.4 Implantação de modelos de AM

Neste ponto, analisou-se os 57 participantes que afirmaram realizar a implantação de modelos de AM em ambiente produtivo. Desses participantes, 43% (25/57) apontou não ter uma equipe dedicada para essa frente de trabalho.

Em geral, dois perfis profissionais ou mais atuam juntos durante a implantação, sendo que na maioria dos casos os cientistas e/ou engenheiros de dados são os responsáveis pelas implantações. Na minoria das implantações, observa-se a atuação em 23% (13/57) dos casos os Engenheiros de Aprendizado de Máquina, 16% (9/57) Engenheiros de Software e 16% (9/57) equipes de DevOps (Figura 8). Em poucos casos observam-se profissionais mais orientados a práticas de MLOps, como Engenheiros de Aprendizado de Máquina, atuando nas implantações. A estrutura de MLOps, pode ajudar os cientistas de dados a dedicar mais tempo em trabalhos de modelagem, do que implantando e orquestrando modelos de AM em produção.

A respeito das dificuldades enfrentadas durante a implantação e monitoramento dos modelos, 59% (34/57) dos participantes queixam-se de ter times insuficientes de engenharia (Engenharia de Dados, Engenharia de Aprendizado de Máquina ou DevOps) e 52% (30/57)

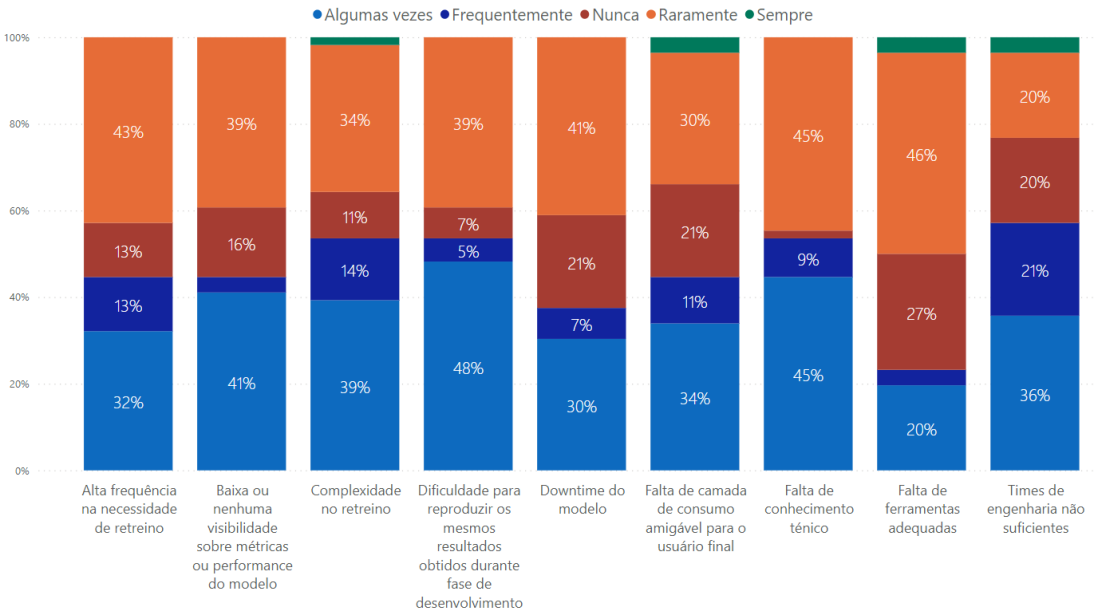
Figura 8 - Profissionais que realizam a implanta dos modelos de AM.



Fonte: Autoria própria (2022).

dizem não ter conhecimento técnico suficiente para esse tipo de tarefa. Todavia, 43% (25/57) dos participantes dizem ter baixa ou nenhuma visibilidade sobre as métricas de modelos em produção. Outro fator importante é que boa parte, isto é 52% (30/57), se queixa da dificuldade em reproduzir na produção, os mesmos resultados obtidos durante a fase de desenvolvimento. Problemas com a visibilidade das métricas dos modelos em produção e a complexidade de retreino, mostram-se também como dificuldades enfrentadas pelos times de ciência de dados (Figura 9).

Figura 9 - Desafios durante a fase de implementação.



Fonte: Autoria própria (2022).

Analisou-se também o grupo de 13% (9/66) dos participantes que afirmaram que suas empresas não realizam implantação de modelo em produção. Nesse caso, as principais dificuldades relatadas foram a falta de conhecimento técnico e times de engenharia (Engenharia de Dados, Engenharia de Aprendizado de Máquina ou DevOps) não suficientes, seguido da falta de uma camada de consumo amigável como *dashboards* ou aplicações *web* e a falta de ferramentas adequadas.

Sobre o formato de consumo dos modelos, os resultados indicam que 47% (27/57) dos modelos em produção são entregues em formato de APIs, 14% (8/57) como aplicações para processamento em lote (*batch*), enquanto 10% (6/57) são entregues como *scripts* em formato *jupyter notebook* [9] para serem executados pelos usuários finais.

Observa-se que apesar da maioria das organizações já realizarem a implantação de modelos em produção, muitos desafios ainda são enfrentados. Com isso, há uma ampla gama de oportunidades para a aplicação de práticas de MLOps. Dificuldades envolvendo o rastreamento de experimentos, reprodutibilidade de resultados, versionamento de conjuntos de dados (*datasets*) e automações são questões intrínsecas ao MLOps. A aplicação de práticas de MLOps pode reduzir a sobrecarga operacional dos times de ciência de dados, automatizar tarefas e beneficiar a implantação e sustentação de modelos de AM em diferentes escalas.

Por fim, notou-se que as empresas normalmente possuem poucos modelos preditivos em produção. Levantou-se como hipótese que isso se deve, em parte, por times de engenharia insuficientes e falta de conhecimento técnico. Além disso, a medida que novos modelos são colocados em produção, torna-se cada vez mais difícil mantê-los sem processos automatizados.

[9] https://jupyter.org/

6 CONCLUSÃO

Neste artigo, os desafios enfrentados durante todo o ciclo de vida de AM foram explorados. Notou-se que os times de ciência de dados têm gasto tempo e esforço consideráveis com atividades além do desenvolvimento de modelos preditivos. Em contraste com outros trabalhos, os resultados deste artigo exibem os problemas enfrentados com questões não técnicas, falta de profissionais de engenharia e uma visão geral sobre o valor percebido pelos negócios. Identificou-se que 86% (57/66) dos participantes já realizam a implantação de modelos em produção, porém a maioria ainda sem processos automatizados. Grande parte dessas implantações são realizados pelos cientistas de dados em conjunto com times de engenharia. Como resultado, foi observado que há falta de recursos como times de engenharia e desafios técnicos e não-técnicos são percebidos durante a implantação.

Os cientistas de dados dedicam boa parte do seu tempo trabalhando em tarefas como limpeza e transformação de dados. Nesse sentido, o MLOps tem um papel importante, podendo ajudar diretamente em todas as fases do ciclo de vida de AM. Em problemas apontados como os do caso da alta frequência de retreino, abordagens como Treinamento Contínuo (*Continuous Training*) podem ajudar na automação desse tipo de tarefa. A construção de *pipelines* de Integração Continua (*Continuous Integration*) e Entrega Contínua (*Continuous Delivery*) possibilitam a implantação de modelos de AM em produção de forma ágil e confiável.

Manter uma operação de modelos de AM requer uma estrutura de times de dados multifuncionais. Modelos de Aprendizado de Máquina apenas entregam seu real valor quando são mantidos de forma saudável em ambiente produtivo e disponível constantemente para consumo pelos usuários. Para que isso ocorra, apenas o desenvolvimento de modelos de AM não é o suficiente, sendo necessário uma estrutura capaz de manter tais modelos continuamente operantes.

Essa pesquisa contribui para o campo de AM, porque em primeiro lugar, aborda de forma ampla o ciclo de vida de Aprendizado de Máquina, discorrendo sobre questões como problemas relacionados a gestão de dados, ferramentas e tecnologias de desenvolvimento, profissionais responsáveis pela implantação, expectativas não realistas, falta de profissionais de engenharia e desconfiança das áreas de negócio na geração de valor, e em segundo lugar, explora como MLOps é importante para a otimização das operações diárias de Aprendizado de Máquina. Como limitações desse trabalho, a maior parte dos participantes são de empresas nacionais.

Referências

A., K. B.; PFLEEGER, S. L. Guide to advanced empirical software engineering. In: SHULL, F.; SINGER, J.; SJøBERG, D. I. K. (Ed.). London: Springer London, 2008. cap. Personal opinion surveys., p. 63–92. Citado 2 vezes nas páginas 14 e 16.

AMERSHI, S. et al. Software engineering for machine learning: A case study. In: IEEE. **2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)**. [S.l.], 2019. p. 291–300. Citado 2 vezes nas páginas 9 e 10.

ASHMORE, R.; CALINESCU, R.; PATERSON, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 5, may 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3453444>. Citado na página 10.

BASILI, V.; WEISS, D. A methodology for collecting valid software engineering data. **IEEE Transactions on Software Engineering**, v. 10, n. 6, p. 728–738, 1984. Citado na página 14.

HAUCK, J. C. R.; WANGENHEIM, C. G. V.; WANGENHEIM, A. V. **Método de Aquisição de Conhecimento para Customização de Modelos de Capacidade/Maturidade de Processos de Software**. Florianópolis, SC, 2011. Citado na página 15.

HECHT, L. E. Add it up: How long does a machine learning deployment take. **The New Stack. TheNewStack**, v. 12, 2019. Citado na página 11.

JAIN, A. et al. Overview and importance of data quality for machine learning tasks. In: **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining**. New York, NY, USA: Association for Computing Machinery, 2020. (KDD '20), p. 3561–3562. ISBN 9781450379984. Disponível em: <https://doi.org/10.1145/3394486.3406477>. Citado na página 12.

KASUNIC, M. **Designing an effective survey**. [S.l.]: Pittsburgh, PA.: Carnegie Mellon University, 2005. Citado 2 vezes nas páginas 14 e 15.

KREUZBERGER, D.; KÜHL, N.; HIRSCHL, S. Machine learning operations (mlops): Overview, definition, and architecture. **arXiv preprint arXiv:2205.02302**, 2022. Citado na página 9.

LITWIN, M. S. **How to Measure Survey Reliability and Validity**. [S.l.]: SAGE Publication, 1995. Citado na página 15.

MÄKINEN, S. et al. Who needs mlops: What data scientists seek to accomplish and how can mlops help? In: IEEE. **2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)**. [S.l.], 2021. p. 109–112. Citado na página 9.

MORABITO, V. Big data governance. **Big data and analytics**, Springer, p. 83–104, 2015. Citado na página 10.

PALEYES, A.; URMA, R.-G.; LAWRENCE, N. D. Challenges in deploying machine learning: a survey of case studies. **ACM Computing Surveys (CSUR)**, ACM New York, NY, 2020. Citado 2 vezes nas páginas 9 e 12.

SCULLEY, D. et al. Hidden technical debt in machine learning systems. In: **Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2015. (NIPS'15), p. 2503–2511. Citado na página 11.

SOUZA, J. V. R. d. Adoção de mlops: desafios de gerenciar código, modelo e dados automaticamente. 2021. Citado 2 vezes nas páginas 9 e 12.

WIGGERS, K. **IDC: For 1 in 4 companies, half of all AI projects fail, 2019**. 2019. Disponível em: <https://venturebeat.com/ai/idc-for-1-in-4-companies-half-of-all-ai-projects-fail/>. Citado 2 vezes nas páginas 11 e 19.