Simulando projetos de Machine Learning com dados e clientes reais em cursos de graduação: qual é a opinião dos estudantes sobre isto?

Fabrício J. Barth[1], **Fabio Roberto de Miranda**[1,2], **Marcelo Graglia**[2]

[1]INSPER Instituto de Ensino e Pesquisa
São Paulo – SP – Brasil

[2]PUC-SP
São Paulo – SP – Brasil

{fabriciojb,fabiomiranda}@insper.edu.br, mgraglia@pucsp.br

Resumo. *Este trabalho avalia a Sprint Session do curso de Ciência da Computação do* INSPER*, iniciativa que aproxima estudantes da realidade de Machine Learning (ML) por meio de desafios baseados em demandas reais. A análise de três edições (2024-2 a 2025-2), envolvendo 74 alunos organizados em 18 equipes que atuaram em 3 projetos, indica, via análise qualitativa, que os estudantes valorizam a aplicação prática com dados reais, o contato corporativo e o desenvolvimento de competências em modelagem, MLOps e deploy. Embora persistam desafios como tempo limitado e complexidade dos dados, a iniciativa demonstra eficácia no desenvolvimento integrado de competências técnicas e profissionais essenciais a projetos de ML.*

1. Introdução

Projetos de Machine Learning (ML) envolvem processos complexos que vão desde a coleta e preparação de dados até a implantação e o monitoramento de modelos em produção. Para que esses projetos sejam conduzidos com eficiência, é necessário um conjunto abrangente de habilidades que perpassa diversas áreas, incluindo engenharia de software, gerenciamento de dados, conhecimento estatístico e competências interpessoais. A formação de profissionais capazes de lidar com essa complexidade representa um desafio significativo para cursos de graduação. Uma das principais dificuldades enfrentadas pelas instituições de ensino superior é oferecer aos alunos experiências que simulem, de forma realista, os desafios encontrados no desenvolvimento de projetos de ML no mercado de trabalho.

A natureza interdisciplinar, prática e iterativa desses projetos nem sempre encontra espaço nas disciplinas tradicionais, o que pode limitar a preparação dos estudantes para atuar profissionalmente na área. Com o objetivo de reduzir essa lacuna, vários cursos de engenharia ou ciência da computação já desenvolveram iniciativas que propõem o desenvolvimento intensivo de soluções para problemas reais em parceria com organizações externas [Sasajima et al. 2024, Lanubile et al. 2023, Dogan 2023, Soares et al. 2024, Barth et al. 2012]. Nessas iniciativas, os alunos trabalham em equipe, em um curto espaço de tempo, para entregar soluções que integram aspectos técnicos e colaborativos, de maneira próxima ao que se espera em ambientes profissionais.

Este trabalho descreve e avalia uma iniciativa denominada *Sprint Session*, desenvolvida no contexto de um curso de Ciência da Computação. Essas sprints ocorrem ao

final de cada semestre, do primeiro ao quarto, e consistem em projetos desenvolvidos por estudantes a partir de demandas concretas provenientes de organizações externas. Cada *sprint* possui duração de três semanas e tem como propósito principal proporcionar aos alunos a oportunidade de aplicar os conhecimentos adquiridos ao longo do semestre em um contexto prático, colaborativo e conectado com o mundo real.

Este artigo investiga em que medida a participação em iniciativas como estas contribui para o desenvolvimento das habilidades necessárias à condução de projetos de ML. Especificamente, buscamos compreender a percepção dos alunos sobre esta experiência durante o processo. Para isso, realizamos uma análise qualitativa dos relatos fornecidos pelos participantes. Espera-se que ao fazer esta análise, tenhamos uma compreensão mais clara de quão relevante é esta iniciativa do ponto de vista dos alunos. Os dados analisados foram dados de *sprints* executados entre o segundo semestre de 2024 e o segundo semestre de 2025. Durante este período foram executadas três (3) sprints diferentes. Ao longo destes sprints foram entregues 18 soluções referentes a 3 projetos distintos. Participaram desta atividade 74 estudantes de graduação do quarto semestre do curso de Ciência da Computação do INSPER.

2. Iniciativas correlatas

Diversas instituições utilizam o *Project-based Learning* (PBL) para ensinar MLOps. Essas iniciativas são projetadas para que os alunos construam e mantenham componentes de ML com qualidade de produção, superando a abordagem acadêmica tradicional focada apenas na otimização de modelos em laboratório [Lanubile et al. 2023, Dogan 2023, Sasajima et al. 2024].

Os cursos de MLOps na Universidade de Bari (Itália) e na Universitat Politècnica de Catalunya (Espanha) organizam o aprendizado em torno de seis marcos de um projeto prático que cobre o ciclo de vida completo do componente ML [Lanubile et al. 2023]: (i) iniciação do projeto: os alunos definem um problema do mundo real e especificam o componente ML; (ii) reprodutibilidade da pipeline: foca em controle de versão de código (Git) e de dados (DVC), além do rastreamento de experimentos usando ferramentas como MLflow; (iii) garantia de qualidade: inclui análise estática (Pylint, flake8), e teste de código e modelos (Pytest). Também abordam a medição da eficiência energética (usando Code Carbon); (iv) implantação de API: os alunos projetam a arquitetura física e desenvolvem APIs RESTful para integrar o modelo em um sistema maior; (v) entrega do componente: utiliza containerização (Docker) para portabilidade e automação de processos com fluxos de CI/CD (GitHub Actions), e; (vi) monitoramento: foca no monitoramento de desempenho do modelo e monitoramento de recursos. Este curso de MLOps na Universidade de Bari demonstrou que as equipes conseguiram implantar seus componentes ML em um ambiente de produção baseado em nuvem, e os alunos consideraram as tecnologias propostas úteis [Lanubile et al. 2023].

Outras iniciativas, como a descrita em [Dogan 2023], em que os alunos trabalham com perguntas de pesquisa e conjuntos de dados reais, brutos e mal estruturados fornecidos por esses mentores de outras áreas. O projeto culmina na redação de um artigo de pesquisa que é submetido a uma conferência internacional. Os alunos desenvolvem habilidades de gerenciamento de projetos, comunicação com outras disciplinas e enfrentam desafios práticos como a limpeza e preparação de dados reais.

O objetivo de [Sasajima et al. 2024] foi desenvolver e avaliar um currículo introdutório de ciência de dados baseado em PBL com uso de dados reais, voltado a alunos do primeiro e segundo ano da graduação. A iniciativa, realizada entre 2019 e 2023 na Universidade de Hyogo, buscou tornar a ciência de dados mais concreta desde o início da formação, aproximando os estudantes de problemas reais de negócio antes do ensino aprofundado de técnicas matemáticas e computacionais. O método envolveu parcerias com empresas, uso de dados reais, atividades práticas em grupo e avaliações anuais por questionários. Os resultados mostram que o PBL foi eficaz para aumentar a motivação dos alunos, fortalecer habilidades de trabalho em equipe e tornar mais clara a atuação profissional do cientista de dados. O estudo também indica que o uso de dados reais é seguro quando bem contextualizado, embora destaque desafios como a variação na qualidade da orientação docente e a importância do interesse dos alunos pelo contexto do negócio para o sucesso dos projetos.

Em resumo, as fontes convergem para a ideia de que o ensino de *Machine Learning* deve ser feito através da resolução de problemas reais do mercado ou de pesquisa, utilizando dados não curados e integrando práticas de engenharia de software e gestão ágil para preparar o estudante para os desafios industriais modernos.

3. Competências para a execução de projetos de ML

Nesta seção serão descritas algumas habilidades que são essenciais para a execução de projetos de ML. Estas habilidades podem ser agrupadas em: habilidades fundamentais de computação; gerenciamento de ciclo de vida de dados; conhecimentos específicos de machine learning; engenharia de software para sistemas de ML, e; competências interdisciplinares e profissionais [Force 2021, Amershi et al. 2019, Schröer et al. 2021, Stodden 2020]. A seguir, os principais grupos de competências são detalhados.

3.1. Conhecimentos específicos de ML

O conhecimento específico em ML envolve desde a seleção e design de modelos até sua avaliação, implantação e manutenção. A escolha adequada de algoritmos, considerando o tipo de modelo (supervisionado, não supervisionado, por reforço), deve estar alinhada aos atributos disponíveis e ao domínio do problema. O treinamento de modelos requer preparação adequada dos dados, ajuste de hiperparâmetros e atenção à desatualização de modelos.

A avaliação dos modelos deve ser criteriosa, utilizando métricas apropriadas e testes em diferentes fatias de dados, incluindo avaliação humana quando necessário. Desafios comuns, como overfitting, instabilidade numérica e generalização inadequada, exigem estratégias específicas de mitigação. A interpretabilidade também é um aspecto essencial, com a aplicação de técnicas que tornem os modelos mais compreensíveis, especialmente em contextos sensíveis ou regulados.

A implantação dos modelos deve considerar automação dos pipelines, integração com sistemas legados, versionamento de código e dados, além de estratégias de rollback seguro. Após a implantação, o monitoramento contínuo do modelo é fundamental para detectar alterações em dependências, desvios entre treinamento e inferência e outras anomalias.

Os conhecimentos específicos de ML podem ser sumarizados pelas nove etapas do workflow de ML [Amershi et al. 2019]. Essas etapas abrangem desde a compreensão do problema e dos dados até a modelagem, avaliação e manutenção do sistema em produção, oferecendo uma estrutura abrangente que serve tanto como guia técnico quanto como base para o desenvolvimento de habilidades práticas (Figura 1). Essa estrutura se alinha com e expande os processos clássicos propostos em KDD [Fayyad et al. 1996] e CRISP-DM [Wirth and Hipp 2000], que também organizam o trabalho em fases encadeadas, enfatizando a importância de entender o domínio do problema, preparar adequadamente os dados, aplicar técnicas de modelagem e interpretar os resultados de maneira iterativa. O workflow em nove etapas aprofunda essas ideias ao incorporar práticas modernas de engenharia de ML, como automação de pipelines, versionamento e monitoramento contínuo, competências essenciais no cenário atual de desenvolvimento e operação de sistemas inteligentes.

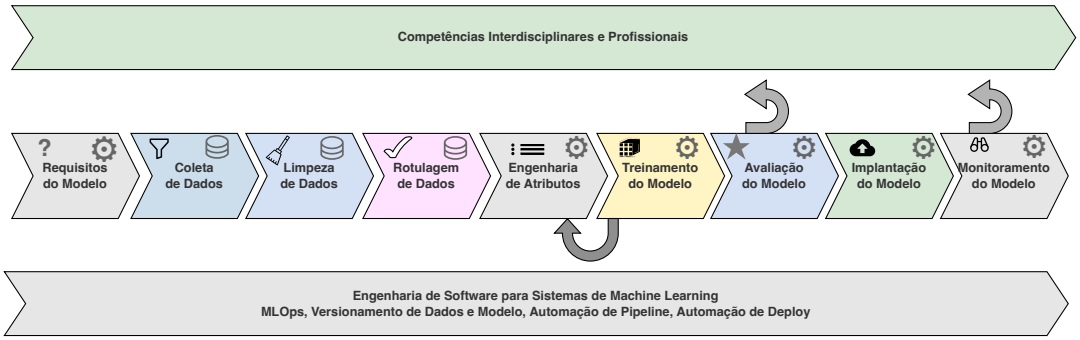

Figura 1. As nove etapas do workflow de Machine Learning, destacando habilidades interdisciplinares e práticas de engenharia de software. Imagem adaptada a partir de [Amershi et al. 2019]

3.2. Gerenciamento do ciclo de vida dos dados

Projetos de ML são inerentemente centrados em dados, tornando o domínio de todo o ciclo de vida dos dados uma competência central. A aquisição de dados envolve a escolha de fontes adequadas e a integração com sensores e APIs. A preparação de dados inclui atividades como limpeza, fusão, transformação, rotulagem e engenharia de features, sempre com foco na qualidade e na adequação ao modelo.

Outro aspecto crítico é o gerenciamento e versionamento de dados, o que exige controle rigoroso de acessos, rastreabilidade e padronização dos dados. A segurança e privacidade dos dados devem ser garantidas em todas as fases, com remoção de informações pessoais e controle de acesso. Compreender os aspectos legais e implementar mecanismos técnicos adequados é parte indispensável dessa responsabilidade.

A integração de ML em ambientes de produção exige práticas de engenharia de software, denominadas *Machine Learning Operations* (MLOps). Isso inclui o desenvolvimento de pipelines automatizados de ponta a ponta, desde a coleta de dados até o monitoramento do modelo. Testes rigorosos, tanto de unidade quanto de integração, garantem a confiabilidade de cada etapa do sistema.

3.3. Competências interdisciplinares e profissionais

Além das habilidades técnicas, o sucesso em projetos de ML depende de competências interpessoais e éticas. A comunicação eficaz, oral e escrita, permite que os resultados sejam compreendidos por públicos diversos, técnicos ou não. O trabalho em equipe e a colaboração interdisciplinar são imprescindíveis em sistemas complexos, especialmente quando envolvem múltiplos modelos ou áreas de conhecimento distintas.

A execução eficaz de projetos de ML requer uma combinação equilibrada de conhecimentos técnicos profundos, práticas de engenharia robustas, responsabilidade ética e habilidades interpessoais. Esse conjunto multidisciplinar de competências permite não apenas construir modelos de alta qualidade, mas também garantir que eles operem de forma segura, justa e eficiente em ambientes de produção. Com a crescente complexidade e escala dos sistemas de ML, o desenvolvimento contínuo dessas habilidades torna-se essencial para profissionais e organizações que desejam extrair valor sustentável de soluções baseadas em dados.

Os conhecimentos específicos de ML podem ser compreendidos com base nas nove etapas do fluxo de trabalho em projetos de aprendizado de máquina, conforme documentado por [Amershi et al. 2019]. No entanto, a Figura 1 destaca não apenas as fases centrais de desenvolvimento mas também enfatiza aspectos fundamentais como automação de pipelines, versionamento de dados e modelos, e o papel das competências interpessoais ao longo de todo o processo. A figura fornece uma visão estruturada e integrada das diversas atividades envolvidas no ciclo de vida de sistemas de ML modernos. Atividades estas que precisam ser compreendidas e executadas por futuros profissionais da área.

4. Integração da Sprint Session no curso de Ciência da Computação

Esta seção tem como objetivo descrever a iniciativa denominada *Sprint Session* e evidenciar como essa abordagem pedagógica viabiliza a aplicação prática do conhecimento na resolução de problemas reais. As sprints ocorrem ao final de cada semestre, do primeiro ao quarto, e consistem em projetos desenvolvidos por estudantes a partir de demandas concretas, provenientes de organizações externas. Cada sprint possui duração de três semanas e tem como propósito principal proporcionar aos alunos a oportunidade de aplicar os conhecimentos adquiridos ao longo do semestre em um contexto prático, colaborativo e conectado com o mundo real.

A proposta pedagógica das *Sprint Session* busca promover o desenvolvimento de competências técnicas e interpessoais, incentivando a motivação intrínseca dos alunos por meio da resolução de problemas concretos. A interação com clientes externos contribui para a simulação de um ambiente profissional, aproximando o contexto acadêmico da realidade do mercado de trabalho.

No primeiro semestre, a disciplina central é Vida de Desenvolvedor, que possui uma carga horária extensa e concentra boa parte das atividades curriculares dos alunos nesse período. Ao longo dessa disciplina, os estudantes aprendem os fundamentos da programação utilizando a linguagem Python. Para isso, contam com um ambiente estruturado de exercícios com *feedback* automatizado, permitindo uma aprendizagem intensiva e contínua. Após essa fase inicial, os alunos passam a trabalhar com conceitos de desenvolvimento em equipe, usabilidade e interface do usuário. Ao final desse processo, os

estudantes estão aptos a participar da *Sprint Session* do primeiro semestre, na qual recebem, de parceiros externos, um conjunto de requisitos descritos na forma de *user stories*. A partir dessas histórias, os alunos devem desenvolver uma solução funcional e validá-la com os clientes ao longo das três semanas do Sprint.

No segundo semestre, a estrutura curricular se aproxima de um curso convencional de Ciência da Computação, com disciplinas como Ciência de Dados, Matemática do Contínuo, entre outras. Entretanto, o semestre é encerrado três semanas antes do período regular, a fim de reservar esse tempo final para a realização de uma nova *Sprint Session*. Nessa etapa, os estudantes são novamente desafiados a desenvolver um projeto proposto por um parceiro externo, geralmente envolvendo a criação de sistemas web do tipo CRUD (*Create, Read, Update, Delete*).

A mesma lógica se aplica ao terceiro semestre, no qual o período letivo também se encerra antecipadamente para dar lugar à sprint. Nessa fase, os projetos desenvolvidos têm como foco principal a solução de problemas de cunho social. Um exemplo recorrente de parceria nesse contexto é a colaboração com a UNAS[1], uma organização não governamental que atua em Heliópolis, São Paulo, cuja missão é promover a inclusão social.

O nível de complexidade dos projetos aumenta gradativamente ao longo dos semestres, refletindo o avanço dos alunos em termos de conhecimento e habilidades interpessoais. No terceiro semestre, os estudantes já possuem uma base sólida em programação e desenvolvimento web, o que lhes permite enfrentar desafios mais complexos e desenvolver soluções mais robustas. Neste momento, os requisitos da solução também são mais vagos. Por exemplo, no primeiro e segundo semestres os alunos recebem uma lista de requisitos, na forma de user stories, que devem ser implementados. Já no terceiro semestre, os alunos precisam compreender o dia a dia da comunidade, onde a UNAS atua, e devem propor soluções que atendam às necessidades reais da comunidade. Isso envolve não apenas o desenvolvimento técnico, mas também a capacidade de ouvir e entender as demandas dos usuários finais, promovendo uma abordagem mais centrada no usuário.

No quarto semestre, o escopo do projeto é relacionado com o tema *Machine Learning*. Neste momento, os alunos já tiveram disciplinas como Ciência de Dados, Álgebra Linear, Teoria da Informação, Inteligência Artificial e Robótica, e Aprendizagem de Máquina. Assim, os alunos são desafiados a desenvolver soluções que envolvam o uso de algoritmos de aprendizado de máquina, aplicando os conhecimentos adquiridos nas disciplinas anteriores.

Exemplos de projetos já executados incluem a análise de churn de clientes de uma startup, modelo para detecção de deficiência de recombinação homóloga para uma empresa de exames laboratoriais, modelo para identificação de fraudes em transações de compras com cartão de crédito em terminais físicos, e solução para extração de informações a partir de prontuários eletrônicos médicos.

O objetivo desta sprint é capacitar os alunos a desenvolver projetos de ciência de dados e machine learning de ponta a ponta, passando por todas as etapas do ciclo de vida de um projeto real, da definição do problema até a entrega de um sistema funcional e

[1]https://www.unas.org.br/

monitorado, com foco em: (i) formulação clara e realista de problemas de negócio; (ii) coleta, limpeza e análise de dados de forma eficiente; (iii) construção, avaliação e refinamento de modelos de ML; (iv) deploy de modelos de machine learning usando boas práticas de engenharia de software (i.e., teste unitário e de integração, deploy automatizado); e (v) documentação, comunicação de resultados e monitoramento contínuo do uso dos modelos.

A rubrica utilizada em todos os semestres/projetos é similar. Os grandes grupos avaliativos são: (i) setup do projeto; (ii) aquisição e pré-processamento dos dados; (iii) engenharia de atributos; (iv) desenvolvimento e avaliação dos modelos; (v) apresentação dos resultados; (vi) deploy dos sistemas e modelos; e (vii) organização da equipe. Os objetivos da ementa e dos grupos avaliativos da rubrica são coerentes com os nove estágios de um projeto de machine learning descritos em [Amershi et al. 2019, Idowu et al. 2022, Idowu et al. 2024, Biswas et al. 2022] e apresentados na figura 1. Além disso, a rubrica é muito similar ao que é exercitado e avaliado em [Lanubile et al. 2023], um dos trabalhos correlatos, descrito na seção 2.

Todas as *Sprint Sessions* têm duração de três semanas, todos os estudantes precisam trabalhar em grupo e como o curso é integral então a dedicação nestas semanas também é integral. A *Sprint Session* tem um docente coordenador, que é responsável por criar os documentos de escopo e de rubrica, além de avaliar os entregáveis ao longo do projeto. Além disso, todos os professores do semestre também participam da *Sprint* nos seus horários de aula exercendo o papel de consultores externos ao projeto. É evidente que alguns docentes são mais demandados do que outros, como, por exemplo, o professor responsável pela disciplina de Machine Learning. Em todos os semestres são utilizados problemas reais e dados reais fornecidos por organizações externas, da mesma forma como acontece em [Sasajima et al. 2024, Dogan 2023].

4.1. Problemas tratados na sprint de ML

No semestre 2024-2 o cliente da *Sprint* foi uma empresa que atua na área de medicina diagnóstica. Esta empresa propôs desenvolver uma *Convolutional Neural Network* (CNN) para a detecção e análise da Deficiência de Recombinação Homóloga (HRD[2]) em amostras de câncer a partir de dados de sequenciamento genético de nova geração. HRD é uma condição em que a via de reparo de DNA através da recombinação homóloga está comprometida, levando à instabilidade genômica e aumentando a suscetibilidade a determinadas terapias, como os inibidores de PARP. Esta CNN poderá fazer parte de uma pipeline para identificar o status de HRD em pacientes com câncer, permitindo estratégias de tratamento personalizadas e melhorando os resultados clínicos. De fato, os estudantes tiveram que replicar o estudo descrito em [Pozzorini et al. 2023]. Os dados manipulados eram representados na forma de matrizes, a quantidade de exemplos disponíveis era na ordem de centenas de milhares.

No semestre 2025-1 o cliente foi uma plataforma de serviços financeiros e de tecnologia (fintech). O objetivo deste projeto foi desenvolver um modelo para identificação de fraudes em transações com cartão de crédito ou débito presenciais, sob o ponto de vista do adquirente. O dataset disponibilizado para o desenvolvimento deste modelo possuía aproximadamente 5 milhões de transações sintetizadas, simulando alguns padrões

[2]Do inglês, *Homologous Recombination Deficiency*

de fraude existentes na vida real. O dataset tinha poucos atributos prontos, ou seja, as equipes tiveram que se dedicar bastante na etapa de feature engineering. Como é um problema de fraude, é fácil perceber que se trata de um dataset altamente desbalanceado, aproximadamente apenas 1% das transações eram fraudes (classe positiva).

No semestre 2025-2 o cliente foi uma instituição de saúde de grande porte que propôs o desenvolvimento de uma solução para a construção automatizada de banco de dados clínicos a partir de prontuários eletrônicos desestruturados. Esta instituição disponibilizou 123 prontuários anonimizados de pacientes com adenocarcinoma de reto localmente avançado. Normalmente, este tipo de paciente tem um longo histórico de consultas e procedimentos na instituição, por isso, em média cada prontuário tinha aproximadamente 600 páginas de texto. Além dos prontuários, a instituição também forneceu um dataset com 80 variáveis por paciente que foram identificadas manualmente por um grupo de especialistas. Este dataset foi construído com o objetivo de realizar estudos retrospectivos sobre o tema e, neste projeto, serviu como *gold standard* para a validação das soluções criadas.

5. Análise qualitativa

Os dados utilizados nesta análise referem-se aos semestres 2024-2, 2025-1 e 2025-2. Na Tabela 1, apresenta-se o número de estudantes e equipes por semestre, bem como observações sobre a composição dessas equipes. No total, participaram 74 estudantes, distribuídos em 18 equipes, que trabalharam no desenvolvimento de soluções para 3 projetos distintos.

Semestres	**Estudantes**	**Equipes**	**Observações**
2024-2	29	7	Uma equipe foi composta por cinco integrantes e as demais por quatro.
2025-1	17	4	Uma equipe foi composta por cinco integrantes e as demais por quatro.
2025-2	28	7	Todas as equipes foram formadas por quatro integrantes.
Total	**74**	**18**	

Tabela 1. Número de alunos e equipes por semestre

Em todos os semestres, todas as equipes trabalharam com o mesmo objetivo e, ao longo de três semanas, precisaram entregar diversos artefatos correspondentes às diferentes etapas de um projeto de ML. No final do projeto, cada equipe apresentou os resultados para os professores e colaboradores da organização parceira. Além das entregas do projeto, os alunos também tiveram que responder um questionário para avaliar o trabalho em equipe e a seguinte pergunta: "*Descreva, em um parágrafo, como foi a sua experiência neste sprint, destacando os pontos positivos e negativos, bem como o que você exercitou e aprendeu durante o processo*". A partir das respostas fornecidas pelos alunos para esta questão aberta, foi realizada uma análise qualitativa por meio de codificação temática, na qual as respostas foram lidas, categorizadas e organizadas em temas representativos. Dois pesquisadores realizaram o processo de codificação de forma independente e, posteriormente, compararam os códigos atribuídos, discutindo divergências até alcançar consenso.

Esse procedimento visou aumentar a confiabilidade da análise, reduzindo vieses individuais e assegurando maior rigor metodológico na definição das categorias finais.

5.1. Semestre 2024-2: detecção de HRD

Neste semestre, 19 alunos responderam ao questionário. Entre as respostas, observou-se que os termos CNN, TensorFlow e desenvolvimento de modelos foram mencionados por sete (7) estudantes, sempre relacionados à etapa de modelagem do projeto. Além disso, referências a atividade prática (como aprender com a prática, entender como funciona na prática ou simplesmente colocar em prática) apareceram cinco (5) vezes. O versionamento de dados e o uso de DVC[3] foram citados quatro (4) vezes. Já os termos MLOps (acrônimo de *Machine Learning Operations*), MLFlow[4] e versionamento de modelos foram mencionados três (3) vezes; porém, ao considerar também os termos pipeline e deploy, é possível agrupá-los em uma categoria mais ampla relacionada a MLOps, totalizando nove (9) menções por estudantes diferentes. A documentação e o uso de README apareceram em duas (2) respostas, destacando a importância de documentar um projeto. Por fim, termos como comunicação, contato com o mercado de trabalho, boas práticas de projetos de ML, aprender biologia e bioinformática, além de interagir com colegas de classe como se fossem colegas de trabalho, foram citados uma única vez cada.

5.2. Semestre 2025-1: identificação de fraude

Todos os 17 estudantes que participaram do projeto de desenvolvimento de um modelo para identificação de fraudes em transações físicas com cartões responderam ao questionário aplicado ao final da atividade. As respostas evidenciam um conjunto relativamente consistente de percepções sobre o projeto. Entre os aspectos positivos, destacam-se a relevância e a aplicabilidade prática do tema (4 menções), a oportunidade de interação direta com uma empresa de grande porte (5 menções) e os aprendizados técnicos associados às etapas do ciclo de ML, incluindo criação de modelos (2), engenharia de atributos (3), avaliação de desempenho (1), realização de deploy (2) e compreensão geral dos processos envolvidos (1). Ademais, os estudantes valorizaram o trabalho colaborativo (3 menções) e o contato frequente com o cliente (2 menções).

Como ocorre em todos os semestres, foi disponibilizado um canal assíncrono de comunicação com a equipe da organização parceira. No período analisado, os estudantes demonstraram utilização efetiva desse recurso. Embora empresas de grande porte participem regularmente dos sprints, neste semestre a atuação da empresa parceira recebeu atenção particular por parte dos alunos. Uma possível explicação é o fato de se tratar de uma organização inserida no mercado business to consumer (B2C), lidando com um tema sensível e próximo à realidade dos participantes, o que pode ter contribuído para um maior engajamento no desenvolvimento do projeto.

Por outro lado, alguns desafios foram recorrentes nas respostas. As principais dificuldades referem-se ao versionamento de dados e ao entendimento do funcionamento do DVC (4 menções), ao tratamento de conjuntos de dados extensos ou desbalanceados (4 menções) e ao surgimento de novas tarefas ao longo dos sprints (3 menções), o que gerou percepção de desorganização. Também foram mencionados o tempo reduzido para

[3]DVC é um software para versionamento de dados
[4]MLFlow é um software para gestão de pipelines e versionamento de modelos

execução das atividades (2 menções), a limitada disponibilidade dos professores para acompanhamento (1 menção), bem como desafios inerentes ao trabalho em equipe e à complexidade do problema proposto (2 menções).

5.3. Semestre 2025-2: extração de dados de prontuários eletrônicos

De forma geral, a sprint foi avaliada de maneira majoritariamente positiva (14 citações explícitas de "boa", "positiva", "ótima" ou "incrível" em um total de 28 respostas), principalmente pelo aprendizado de novos conceitos e ferramentas (cerca de 12 citações), como NLP[5], OCR[6], DVC e LLMs[7]. O contato com dados reais e com um projeto mais próximo do mercado apareceu em aproximadamente 8 relatos, sendo um dos pontos mais valorizados. Em contrapartida, o tempo insuficiente foi o ponto negativo mais recorrente (cerca de 10 menções), seguido pela complexidade e volume dos dados (aproximadamente 9 menções) e por limitações técnicas como hardware, consumo de tokens em LLMs e lentidão de processamento (cerca de 6 menções). Também surgiram relatos sobre dificuldades com métricas para avaliação de modelos, organização, comunicação em grupo e frustração com desempenho dos modelos (cerca de 7 citações). Assim, predominou a percepção de que foi uma experiência desafiadora e cansativa, mas altamente enriquecedora do ponto de vista técnico e profissional.

6. Conclusão

A análise conjunta das três turmas revela um conjunto consistente de elementos em comum, tanto nos aspectos positivos quanto nos desafios enfrentados. Em todos os semestres, observa-se forte valorização do caráter prático dos projetos e da aproximação com problemas reais, seja por meio do uso de dados do mundo real, da interação com empresas parceiras ou da aplicação de técnicas atuais de ML. Termos relacionados à modelagem e ao desenvolvimento técnico aparecem de forma recorrente, como no semestre 2024-2 (CNN, TensorFlow e desenvolvimento de modelos), em 2025-1 (menções a criação de modelos, engenharia de atributos e deploy) e em 2025-2 (NLP, OCR, DVC e LLMs). Além disso, tópicos associados a práticas mais amplas de MLOps (incluindo versionamento, pipelines e deploy) ganharam destaque, especialmente em 2024-2 e 2025-1.

Por outro lado, os desafios também apresentam padrões claros. A limitação de tempo é recorrente (2025-1 e 2025-2), assim como dificuldades relacionadas ao tratamento e à complexidade dos dados (2025-1 e 2025-2). Questões envolvendo versionamento de dados e uso do DVC aparecem tanto como aprendizado quanto como obstáculo. Também surgem, de forma transversal, percepções sobre desorganização, surgimento de tarefas não previstas, limitações técnicas e desafios no trabalho em equipe.

Em síntese, os resultados indicam que os sprints cumprem um papel relevante na formação dos estudantes ao proporcionar vivências próximas à realidade profissional, expondo-os não apenas às etapas técnicas do ciclo de vida de projetos de ML, mas também às incertezas, restrições de tempo, limitações de infraestrutura e desafios de comunicação típicos do contexto real. Ainda que as dificuldades sejam frequentemente mencionadas, elas coexistem com uma percepção amplamente positiva de aprendizado significativo e desenvolvimento técnico e profissional.

[5]*Natural Language Processing*
[6]*Optical Character Recognition*
[7]*Large Language Models*

Declaração sobre uso de Inteligência Artificial

Os autores deste artigo utilizaram ferramentas de Inteligência Artificial, especificamente o ChatGPT (OpenAI), com o objetivo de auxiliar na revisão gramatical e no aprimoramento da fluidez textual.

Referências

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300.

Barth, F., Burd, L., and Pimentel, M. (2012). Escritório de projetos: simulando o ambiente de projetos de software em cursos de tecnologia. In *Anais do XX Workshop sobre Educação em Computação*, pages 299–302, Porto Alegre, RS, Brasil. SBC.

Biswas, S., Wardat, M., and Rajan, H. (2022). The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, pages 2091–2103. ACM.

Dogan, G. (2023). Teaching machine learning with applied interdisciplinary real world projects. In Kinnaird, K. M., Steinbach, P., and Guhr, O., editors, *Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop*, volume 207 of *Proceedings of Machine Learning Research*, pages 12–15. PMLR.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34.

Force, A. D. S. T. (2021). *Computing competencies for undergraduate data science curricula*. Association for Computing Machinery, New York, NY, USA.

Idowu, S., Sens, Y., Berger, T., Krueger, J., and Vierhauser, M. (2024). A Large-Scale Study of ML-Related Python Projects. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, pages 1272–1281, New York, NY, USA. Association for Computing Machinery. event-place: Avila, Spain.

Idowu, S., Strüber, D., and Berger, T. (2022). Asset Management in Machine Learning: State-of-research and State-of-practice. *ACM Comput. Surv.*, 55(7).

Lanubile, F., Martínez-Fernández, S., and Quaranta, L. (2023). Teaching mlops in higher education through project-based learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, page 95–100. IEEE.

Pozzorini, C., Andre, G., Coletta, T., Buisson, A., Bieler, J., Ferrer, L., Kempfer, R., Saintigny, P., Harlé, A., Vacirca, D., Barberis, M., Gilson, P., Roma, C., Saitta, A., Smith, E., Consales Barras, F., Ripol, L., Fritzsche, M., Marques, A. C., Alkodsi, A., Marin, R., Normanno, N., Grimm, C., Müllauer, L., Harter, P., Pignata, S., Gonzalez-Martin, A., Denison, U., Fujiwara, K., Vergote, I., Colombo, N., Willig, A., Pujade-Lauraine, E., Just, P.-A., Ray-Coquard, I., and Xu, Z. (2023). Giinger predicts homologous recombination deficiency and patient response to parpi treatment from shallow genomic profiles. *Cell Reports Medicine*, 4(12):101344.

Sasajima, M., Ishibashi, K., Yamamoto, T., Yumoto, T., Ohshima, H., Fujie, T., and Katoh, N. (2024). Is problem-based learning exercises using real data effective on the education of lower grades in the faculty of data science? findings on pbl exercises for five years. In *2024 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 398–404.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

Soares, L. P., Lima, L. C., and Silva, R. I. G. (2024). Navigating engineering capstone strategies.

Stodden, V. (2020). The data science life cycle. *Communications of the ACM*, 63(7):58–66.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.