



Universidade Estadual de Campinas
Instituto de Computação



Thales Eduardo Nazatto

Construção de um *Pipeline* de *Machine Learning*
autônomo com métricas de *Fairness*

CAMPINAS
2022

Thales Eduardo Nazatto

**Construção de um *Pipeline de Machine Learning* autônomo com
métricas de *Fairness***

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Cecília Mary Fischer Rubira
Coorientador: Prof. Dr. Leonardo Montecchi

Este exemplar corresponde à versão da
Dissertação entregue à banca antes da
defesa.

CAMPINAS
2022

Na versão final, esta página será substituída pela ficha catalográfica.

De acordo com o padrão da CCPG: “Quando se tratar de Teses e Dissertações financiadas por agências de fomento, os beneficiados deverão fazer referência ao apoio recebido e inserir esta informação na ficha catalográfica, além do nome da agência, o número do processo pelo qual recebeu o auxílio.”

e

“caso a tese de doutorado seja feita em Cotutela, será necessário informar na ficha catalográfica o fato, a Universidade convenente, o país e o nome do orientador.”

Na versão final, esta página será substituída por outra informando a composição da banca e que a ata de defesa está arquivada pela Unicamp.

Você não consegue ligar os pontos olhando pra frente; você só consegue ligá-los olhando pra trás. Então você tem que confiar que os pontos se ligarão algum dia no futuro. Você tem que confiar em algo – seu instinto, destino, vida, carma, o que for. Esta abordagem nunca me desapontou, e fez toda diferença na minha vida.

(Steve Jobs)

Agradecimentos

Primeiramente, à minha família, pela dedicação que tiveram em me criar, pela liberdade de escolha para fazer o que gosto e pela compreensão atual de que hoje seguimos caminhos completamente distintos, apesar de mantermos contato constantemente.

A meus orientadores, Profa. Dra. Cecília Mary Fischer Rubira e Prof. Dr. Leonardo Montecchi, pelo desafio de orientar uma dissertação com temas fora de seus domínios de estudo, e também ao Prof. Dr. Gerberth Adín Ramírez Rivera, que me aceitou como orientador anteriormente, foi compreensivo no momento de minha desistência e que pude fazer reflexões com tal experiência que levei para esta dissertação final de alguma forma.

A todas as pessoas com quem morei em Campinas nesses anos desde que comecei a frequentar aulas, presentes nas Repúblicas Borritos, KioBio e Galinheiro, pelos momentos de lazer que me fizeram relaxar das tensões encaradas em um curso de Pós-Graduação.

A todas as pessoas que conheci na época em que eu estudei em Rio Claro e reencontrei em Campinas, e também a todas as pessoas que mantive contato de Rio Claro desde que comecei a frequentar aulas, principalmente por entender que as conexões e comunicações se mantém mesmo quando um ciclo se fecha e um ciclo diferente é iniciado.

A todas as pessoas que trabalhei junto na CI&T, Dextra e Zup, pela compreensão de que o estudo também é essencial para a evolução profissional de uma pessoa.

E finalmente, a todas as pessoas que pude conhecer na Unicamp, pela troca de experiências e pelo contato com pessoas de altíssimo nível e acima de todas as minhas expectativas.

Resumo

Esta dissertação de Mestrado possui o objetivo de contribuir com o tema de Inteligência Artificial (IA) do ponto de vista da Engenharia de Software, visto que o uso de IA envolvendo grandes volumes de dados vem crescendo conforme nossa sociedade migra processos manuais de trabalho para soluções digitais e necessita de tomadas de decisão mais rápidas e assertivas. Entretanto, as métricas usadas inicialmente para definir a eficácia de um algoritmo se mostraram limitadas devido a barreiras éticas e legais, resultando em vieses que refletem a sociedade de maneira que não era esperada pelos desenvolvedores da solução. Desta forma, o uso de novas métricas, como métricas de Fairness, e proveniência de dados pode ajudar neste problema. Desenvolvendo um pipeline autônomo utilizando a arquitetura MAPE-K, documentando métodos e ferramentas, e realizando estudos de caso, é esperado que este problema seja controlado e facilite o desenvolvimento de softwares com o uso responsável de dados.

Abstract

This Master's Degree dissertation aims to contribute to the theme of Artificial Intelligence (AI) from the Software Engineering point of view, since the use of AI involving large volumes of data has been growing as our society migrates manual work processes to digital solutions and needs faster and more assertive decision making. However, the metrics used initially to define the effectiveness of an algorithm proved to be limited due to ethical and legal barriers, resulting in biases that reflect society in a way that was not expected by the solution developers. Thus, the use of new metrics, like Fairness metrics, and data provenance can help with this problem. By developing an autonomous pipeline using MAPE-K architecture, documenting methods and tools, and making case studies, it is expected that this problem will be controlled and facilitate software development with the responsible use of data.

Listas de Figuras

2.1	Ciclo de vida e ecossistema do dado [21].	18
2.2	Exemplo de uma rede neural utilizada em <i>Deep Learning</i>	21
2.3	Processo padrão para aprendizado de máquina	31
2.4	IBM Analytics and AI Reference Architecture.	32
2.5	Ciclo de vida da proveniência. [59]	34
2.6	Diagrama de funcionamento da arquitetura MAPE-K [16].	37
2.7	Diagrama de uma arquitetura <i>Pipe-and-Filter</i>	38
3.1	Diagrama de atividades com subdivisão de cada papel em uma aplicação de IA utilizando métricas de Fairness, com base na IBM AI Reference Architecture [4]	41
3.2	<i>Assurance Cases</i> feitos para detalhar os objetivos necessários para executar o Pipeline mais adequado	43
3.3	Diagrama de classes do <i>Framework</i> baseado na arquitetura <i>Pipe-and-Filter</i>	45
3.4	Fases e etapas do Pipeline implementado	48
3.5	Adequação do Pipeline ao gerenciador MAPE-K	50
3.6	Comportamento das opções de menu.	55
3.7	Configuração da etapa de análise para o pipeline autônomo.	56
3.8	Configuração das métricas para etapa de análise do pipeline autônomo.	57
3.9	Cenários possíveis na configuração das métricas.	58
3.10	Configuração da etapa de planejamento para o pipeline autônomo.	59
3.11	Execução simples e manual do Pipeline.	60
3.12	Informações do resultado do pipeline.	61
3.13	Métricas do resultado do pipeline.	62
3.14	Execução autônoma do Pipeline.	63
3.15	Seleção do pipeline após análise.	64
3.16	Execução do pipeline após seleção.	65
A.1	Fases de um processo baseado em IA explicável.	86

Listas de Tabelas

2.1	Matriz de confusão	21
4.1	Melhores opções escolhidas pelo modelo MAPE-K Todos os métodos - 50% Performance/50% Fairness	67
4.2	Melhores opções escolhidas pelo modelo MAPE-K Todos os métodos - 75% Performance/25% Fairness	68
4.3	Melhores opções escolhidas pelo modelo MAPE-K Todos os métodos - 25% Performance/75% Fairness	68
4.4	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no dado - 50% Performance/50% Fairness . .	69
4.5	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no dado - 75% Performance/25% Fairness . .	69
4.6	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no dado - 25% Performance/75% Fairness . .	69
4.7	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no treinamento - 50% Performance/50% Fairness	70
4.8	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no treinamento - 75% Performance/25% Fairness	70
4.9	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no treinamento - 25% Performance/75% Fairness	70
4.10	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no resultado - 50% Performance/50% Fairness	71
4.11	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no resultado - 75% Performance/25% Fairness	71
4.12	Melhores opções escolhidas pelo modelo MAPE-K Apenas com redução de viés no resultado - 25% Performance/75% Fairness	71
4.13	Melhores opções escolhidas pelo modelo MAPE-K Apenas sem redução de viés - 50% Performance/50% Fairness	72
4.14	Melhores opções escolhidas pelo modelo MAPE-K Apenas sem redução de viés - 75% Performance/25% Fairness	72
4.15	Melhores opções escolhidas pelo modelo MAPE-K Apenas sem redução de viés - 25% Performance/75% Fairness	72
4.16	Métricas de performance das execuções de Pipeline	73
4.17	Métricas de Fairness das execuções de Pipeline	74

Sumário

Agradecimentos	6
Resumo	7
Abstract	8
1 Introdução	13
1.1 Motivação	13
1.2 Objetivo	14
1.3 Organização	14
2 Conceitos Relacionados	15
2.1 Ciência de Dados e Engenharia de Dados	15
2.1.1 Engenharia de Dados	15
2.1.2 Ciência de Dados	16
2.1.3 O ciclo do dado: Semelhanças e diferenças entre Engenharia de Dados	17
2.2 Machine Learning	19
2.2.1 Métricas de Avaliação	21
2.2.2 Algoritmos	22
2.3 Fairness em Machine Learning	27
2.3.1 Métricas de Fairness	27
2.3.2 Algoritmos para redução de vieses	29
2.4 Engenharia de Software	31
2.4.1 Engenharia de Software para Aplicações de IA	31
2.5 Proveniência de Dados	32
2.6 Arquitetura de Software	35
2.6.1 Arquitetura MAPE-K	36
2.6.2 Arquitetura <i>Pipe-and-Filter</i>	37
3 Metodologia	40
3.1 Detalhamento do processo	40
3.2 Arquitetura do código	43
3.2.1 Transformação dos conjuntos de dados	43
3.2.2 Pipeline	45
3.2.3 Componente MAPE-K	50
3.2.4 Interface Humano-Computador	54
3.3 Experimentos e limitações	65
4 Resultados e Discussões	67

5 Conclusões	78
A Conceitos complementares	86
A.1 AI Explainability	86
A.2 <i>Fluent API</i>	88
A.3 <i>Feature Toggles</i>	89
A.4 <i>MLOps</i>	90

Capítulo 1

Introdução

1.1 Motivação

Técnicas de Inteligência Artificial e Aprendizado de Máquina já são utilizadas há bastante tempo no ramo da Computação. Ramos como robótica e jogos são grandes exemplos, dada a necessidade nos mesmos de automatizar comportamentos que seriam tidos como triviais para um ser humano. Entretanto, nos últimos anos ocorreu um crescimento no uso dessas tecnologias em aplicações tradicionais, com previsão de US\$ 57 bilhões em investimentos em 2021, 480% maior em relação a 2017 [13]. No Brasil, o número de empresas de IA aumentou de 120 em 2018 para 206 em 2020 [14].

Isso se mostra possível devido a grande quantidade de dados processada diariamente pelas empresas, que coletam estatísticas toda vez que um usuário acessa suas aplicações. Com esses dados, podem traçar diferentes perfis e usar soluções de IA para ter tomadas de decisão mais assertivas com o objetivo de melhorar a experiência de usuário e corrigir problemas. Porém, muitas dessas soluções foram projetadas sem pensar em governança de dados como requisito de projeto, e se mostram ineficientes quando ela é tratada em consideração.

Governança de dados é um tema que entrou em evidência recentemente em países como o Brasil: Iniciativas como a LGPD - Lei Geral de Proteção de Dados [15], de 14 de agosto de 2018, mostram como as aplicações e seus dados possuem cada vez mais influência na sociedade moderna. E nesse ponto muitas aplicações de IA falham: muitas implementações foram implementadas como *black boxes*, onde o determinante para estabelecer a confiança no modelo implementado é sua entrada e sua saída.

Um outro efeito colateral dessa estratégia é a exposição de vieses que, embora sejam vistos como não-intencionais pelos desenvolvedores por ter a possibilidade de ser um *outlier* no modelo treinado, refletem preconceitos escancarados da sociedade atual. Uma entrada de dados enviesada resulta em um algoritmo que realiza discriminações em sua classificação [27], e uma vez que as métricas utilizadas para medir a qualidade de um modelo *black box* são geralmente baseadas em acurácia, precisão e recall, discriminações não são facilmente percebidas por tais métricas. Ao mesmo tempo, um modelo *black box* pode ter uma alta dependência de poucos dados, determinando problemas de acoplamento. Para resolver este problema, é possível que a criação de um novo modelo seja uma melhor opção que realizar a correção em apenas parte dele.

Devido a esses tipos de problemas, o termo *Explainable AI* (XAI) ganha força para envolver o desenvolvimento de uma IA que seja acurada e simultaneamente transparente. Como IA possui diversos tipos de métodos diferentes para enquadrar diversos tipos de dados, o mesmo acaba se aplicando em Explainable AI, podendo enquadrar em diversos tipos de dados [72], ou dados específicos como imagens [47] e tabelas [55]. Como o objetivo em XAI é fazer com que os resultados alcançados pela solução de IA sejam compreendidos por humanos, é possível considerar este fato como requisito no design de uma solução de IA, fazendo com que a mesma seja reusável e testável.

No mesmo tema, é possível estabelecer métricas para determinar o quanto o modelo está preparado para dados sensíveis [20], termo que é conhecido como *Fairness*. Com a evolução das pesquisas na comunidade acadêmica, foram descobertos algoritmos para redução dos vieses presentes nos conjuntos de dados, como *Reweighting* [45], *Adversarial Debiasing* [78] e *Reject Option Classification* [46]. Por consequência, ocorre melhora nas métricas em questão, mas pode desfavorecer métricas que já são amplamente utilizadas como garantia de um bom modelo desenvolvido com técnicas de Aprendizado de Máquina.

1.2 Objetivo

O objetivo desta dissertação de mestrado é desenvolver uma estrutura de *Pipeline* para *Machine Learning* que seja completamente autônoma, por três fatores principais:

- Facilitar a criação de modelos justos e confiáveis com a automatização da escolha dos algoritmos, cuja complexidade aumenta com a escolha dos algoritmos a serem utilizados e suas execuções nas etapas corretas do processo, onde eles foram escolhidos para atuar.
- Estabelecer um balanceamento entre métricas para avaliar bons modelos com métricas para avaliar modelos justos.
- Considerar proveniência de dados como requisito no design de uma solução de IA, e como uma alternativa a XAI através da utilização de metadados.

Para isso, além do desenvolvimento do *Pipeline* em questão, será utilizada a arquitetura MAPE-K [10] para analisar uma base de conhecimento e prover o melhor pipeline seguindo regras pré-determinadas.

1.3 Organização

O restante da dissertação se organizará da seguinte forma: o Capítulo 2 descreve os conceitos que irão ser abordados neste projeto; o Capítulo 3 mostra a metodologia e detalhamento do processo de desenvolvimento; o Capítulo 4 discute os resultados obtidos e, finalmente, o Capítulo 5 estabelece as conclusões, considerações finais e sugestões de trabalhos futuros e evoluções.

Capítulo 2

Conceitos Relacionados

2.1 Ciência de Dados e Engenharia de Dados

2.1.1 Engenharia de Dados

A engenharia de dados é o meio para entender um processo. Os dados podem ser gerados de várias maneiras, ou um subconjunto dos dados disponíveis pode usar técnicas de análise de dados de estatísticas, aprendizado de máquina, reconhecimento de padrões ou redes neurais, juntamente com outras tecnologias, como visualização, otimização, sistemas de banco de dados, dados, ferramentas de prototipagem e elicitação de conhecimento. O objetivo é usar os dados disponíveis ou gerar mais dados e assim entender o processo que está sendo investigado. O processo de analisar os dados, criar novas ferramentas de análise especificamente para a tarefa e trabalhar com especialistas do domínio é um aspecto fundamental dessa tarefa de engenharia. Atualmente é muito utilizado em conjunto com o termo *Big Data*, para a limpeza, tratamento e estabelecimento de processos para governança de grandes volumes de dados.

O termo *Big Data* apareceu uma vez como conceito em 1974 e novamente em editoriais em 2006 e 2007, e somente em 2008 seu uso como conceito começou a aparecer regularmente em artigos científicos, mas implementações desse conceito começaram a partir de 2010 [64]. Quanto a engenharia de dados, embora periódicos como o IEEE Transactions on Knowledge and Data Engineering, cuja primeira edição foi lançada em 1989, e conferências como a IEEE International Conference on Data Engineering (ICDE), cuja primeira edição foi realizada em 1984, possam ser consideradas pontapés iniciais para discussões e artigos acadêmicos, o embrião da engenharia de dados vem do paper "*A Business Intelligence System*", de 1958 [12]. Nele, é mencionada a ideia de sistemas rodados por máquinas para abstrair e padronizar informações de vários setores da sociedade, como industrial, científico e de organizações governamentais, e como estas podem disseminar informação de uma maneira mais eficiente. Embora os anos 50/60 foram o embrião para o conceito, os anos 70/80 o amadureceram e construíram a base para a estrutura de engenharia de dados [12].

Nos anos 70/80, os problemas em engenharia de dados eram classificados de acordo com 3 atributos [66]:

- **Completude do conhecimento e dos dados:** Os dados e o conhecimento disponíveis no ambiente poderiam ser considerados como completos ou incompletos. Se estiverem completos, não seria necessário conhecimento adicional para a resolução do problema. Se estiverem incompletos, como grandes volumes de dados, seria necessário determinar heurísticas para encontrar um conjunto de dados mais específico para a resolução do problema.
- **Exatidão do conhecimento e dos dados:** Os dados e o conhecimento disponíveis no ambiente poderiam ser considerados como exatos ou inexatos. Se estiverem exatos, poderiam ser representados em uma forma numérica ou lógica, como datas e coordenadas. Se estiverem inexatos, o número de casos possíveis seria infinito e seria impossível enumerar ou representar todos eles, sendo necessário determinar heurísticas para definir um número finito de possibilidades ou redefinir o significado de exatidão para que o que está disponível possa ser tratado como exato, como o reconhecimento de um objeto em uma imagem ou a definição do menor custo em uma determinada rota.
- **Conhecimento sobre o objetivo e especificações do problema:** Um objetivo de um problema pode ser bem ou mal definido. Um objetivo bem definido podia ser medido e representado em termos de parâmetros para que seja possível comparar a qualidade de uma solução com outra, enquanto um objetivo mal definido envolvia parâmetros que não podem ser mensurados ou não poder ser medido, sendo impossível comparar a qualidade de soluções alternativas e necessitando de tratamentos adicionais para que o objetivo deixe de ser mal definido e passe a ser bem definido.

Dado estas categorias, é possível pensar em soluções que conversam com os objetivos da Engenharia de Software: Abordar diversos tipos de conceitos; como teoria, projeto, desenvolvimento, avaliação e manutenção de novos dados e técnicas, metodologias e sistemas de gerenciamento de conhecimento; em prol do desenvolvimento e manutenção de sistemas e soluções com qualidade e com o custo mais viável possível. Estes sistemas podem ter questões relacionadas ao cumprimento desses objetivos incluem estudos sobre os aspectos teóricos, ferramentas e metodologias de projeto, tradeoffs de projeto, representação e programabilidade, algoritmos e controle, confiabilidade e tolerância a falhas e projetos usando tecnologias existentes e emergentes. É tarefa do engenheiro de dados elaborar processos que refletem tais questões em um sistema com objetivos definidos.

2.1.2 Ciência de Dados

Ciência de dados como um conceito precedeu o *Big Data*, sendo um conjunto de princípios fundamentais que apoiam e orientam a extração baseada em princípios de informação e conhecimento de dados [64]. Essa definição enfatiza a estreita relação de ciência de dados com a mineração de dados e, consequentemente, com a engenharia de dados. O termo foi cunhado nos anos 60 para descrever uma nova profissão que daria suporte à compreensão e interpretação da grande quantidade de dados que estavam sendo acumulados na época [39]. Na academia, o termo começou a ser utilizado de modo mais formal a partir de 2001,

quando os títulos dos artigos científicos começaram a usar, embora artigos já estavam focando em ciência de dados e usando o termo desde os anos 60 [64] [39].

A estatística e o uso de modelos estatísticos estão profundamente enraizados no campo da ciência de dados, pois começou com estatísticas e evoluiu para incluir conceitos e práticas principalmente para aplicações de IA. À medida que mais e mais dados se tornam disponíveis, por meio de comportamentos e tendências registradas em softwares espalhados pela Internet ou mesmo por aplicações empresariais, as empresas os coletam e armazenam em quantidades cada vez maiores. Uma vez que as portas foram abertas por empresas que buscam aumentar os lucros e impulsionar melhores tomadas de decisão, seu uso, em conjunto com *Big Data*, começou a ser aplicado a outros campos, como medicina, engenharia e ciências sociais.

Um cientista de dados funcional, ao contrário de um estatístico geral, tem compreensão da arquitetura de software e entende várias linguagens de programação. O cientista de dados define o problema, identifica as principais fontes de informação e projeta a estrutura para coletar e rastrear os dados necessários. O software é normalmente responsável por coletar, processar e modelar os dados. Eles usam os princípios da ciência de dados e toda sua visão multidisciplinar para obter conhecimentos mais profundos. A ciência de dados continua a evoluir como uma disciplina que usa ciência da computação e metodologia estatística para fazer previsões úteis e obter insights em uma ampla gama de campos. Embora seja usada em áreas como astronomia e medicina, também é usada nos negócios para ajudar a tomar decisões mais inteligentes.

2.1.3 O ciclo do dado: Semelhanças e diferenças entre Engenharia de Dados

Nos tempos atuais, empresas e instituições acadêmicas utilizam Engenharia de Dados e Ciência de Dados para otimizar as suas aplicações de IA. Embora tais termos sejam distintos e seus profissionais performam tarefas distintas, na prática é frequente a função de engenharia de dados ser feita por cientistas de dados e vice-versa pois ambas se complementam muito na criação de aplicações de IA. Um bom exemplo de como estas áreas se conversam está no *paper Concise Survey of Computer Methods*, de 1974 [12]. Nele, Peter Naur detalha diversos métodos de processamento de dados e aplicações, o que poderia definir como um, mas a citação mais importante de seu *paper* está justamente no termo ciência de dados. Também define que a utilidade de um dado e de seus processos deriva de sua aplicação no desenvolvimento e manutenção de modelos da realidade [39]. Em outras palavras, a existência de um processo que trate os dados é tão importante quanto a construção de um modelo através de seus dados. Ao entender como os dados devem ser usados, é possível criar processos que auxiliem a otimizar os resultados obtidos e criar modelos cada vez mais próximos da realidade.

A principal razão para ambas as áreas se conversarem está no ciclo de vida do dado, fundamental para entender as oportunidades e os desafios de aproveitar ao máximo os dados digitais. Como um ser vivo, os dados têm um ciclo de vida, desde o nascimento, passando por uma vida ativa até alguma forma de expiração, podendo ser "imortalizado" dependendo de sua importância. Também como um organismo vivo e inteligente,

sobrevive em um ambiente que fornece suporte físico, contexto social e significado existencial [21].

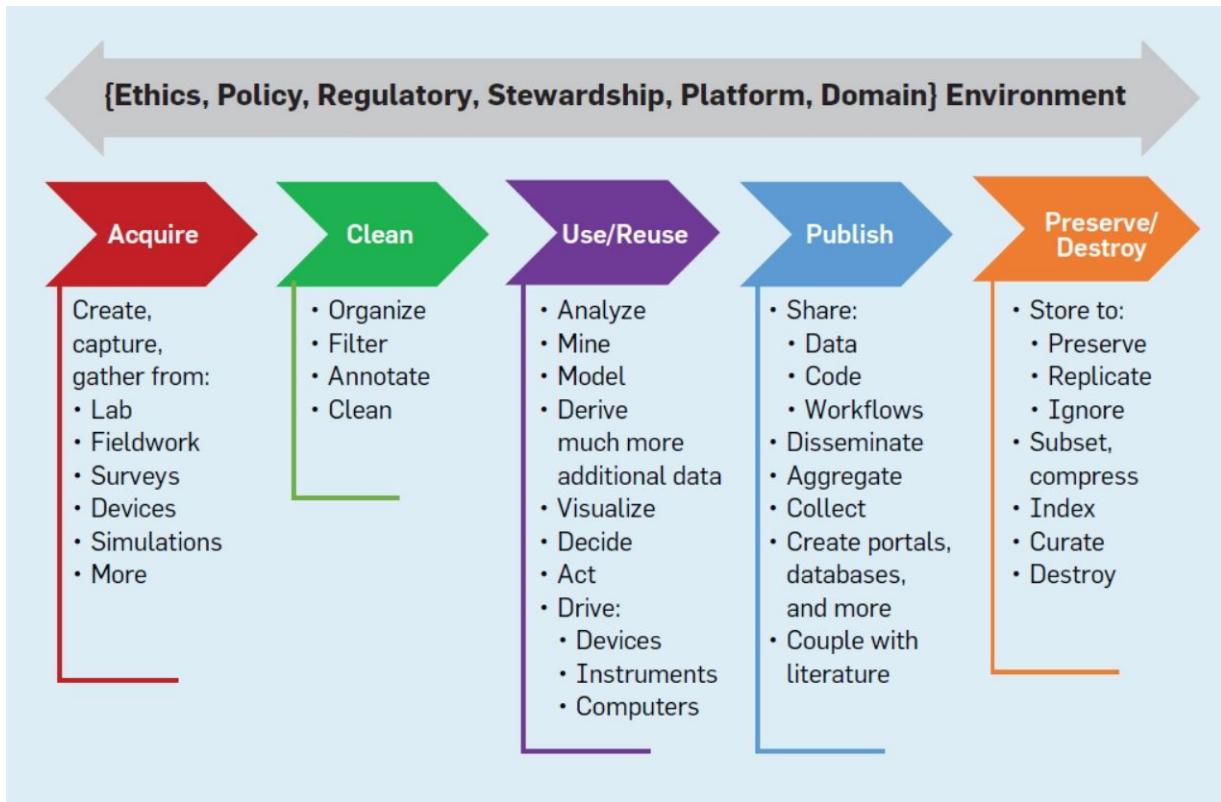


Figura 2.1: Ciclo de vida e ecossistema do dado [21].

Este ciclo, ilustrado na Figura 2.1, é dividido em 5 etapas:

- **Aquisição:** Refere-se a processos de criação e coleta de dados, através de experimentos, sensores, pesquisas, sistemas, simulações, entre outros processos
- **Limpeza:** Refere-se a processos de limpeza, organização e filtragem dos dados adquiridos.
- **Uso/Reuso:** Refere-se a aplicações que os dados podem ter para a aquisição de novos conhecimentos, como análise, mineração, modelagem, enriquecimento, sintetização de novos dados, visualização e tomadas de decisão.
- **Publicação:** Refere-se ao compartilhamento destes dados após seu uso, podendo ter como meio alguma plataforma de compartilhamento, bases de dados ou artigos acadêmicos.
- **Preservação/Destruição:** Refere-se a processos de armazenamento, curadoria e validação do dado após seu uso e publicação. No contexto de validação, se o mesmo ainda continua útil para o contexto em que foi coletado, podendo ser atualizado ou destruído em caso negativo. Também pode ser comprimido ou indexado caso espaço ou desempenho sejam considerados gargalos para os usuários que forem usar os dados publicados.

Como exemplo, é possível citar os dados para experimentos do Grande Colisor de H adrons (LHC), representando colisões de part culas dentro de um t nel de 17 milhas para testar as previsões de v rias teorias da f sica de part culas e da f sica de alta energia [21]. A maioria dos dados gerados   tecnicamente irrelevante e s o descartados, mas isso n o impede que uma enorme quantidade de dados seja influente e continua a ser analisada e preservada. Em 2012, dados sobre experimentos do LHC forneceram fortes evid ncias para o b son de Higgs, apoiando a veracidade do Modelo Padr o da F sica. Esta descoberta cient fica foi a "Descoberta do Ano" de 2012 da revista *Science* e o Pr mio Nobel de F sica em 2013.

As estimativas s o de que, em 2040, haver  de 10 exabytes a 100 exabytes de dados influentes produzidos pelo LHC. Os dados retidos do LHC s o anotados, preparados para preserv o e arquivados em mais de uma d zia de locais f sicos. O resultado desse processo   divulgado   comunidade para an lise e uso em mais de 100 outros locais de pesquisa. Al m do desenvolvimento de protocolos de administra o, dissemin o e uso de dados, o ecossistema de dados do LHC tamb m fornece um modelo econ mico que suporta de forma sustent vel os dados e sua infraestrutura. Essa combina o entre administra o dos dados e administra o econ mica permitem que os dados sejam mantidos.

O diagrama do ciclo de vida dos dados descrito na figura e o exemplo do LHC sugerem um conjunto cont nua de a es e transforma es nos dados, mas em muitas comunidades cient ficas e disciplinas hoje essas etapas s o isoladas. Os cientistas de dom nio se concentram em gerar e usar dados. Cientistas da computa o podem se concentrar em quest es de plataforma e desempenho, incluindo minera o, organiz o, modelagem e visualiza o, bem como os mecanismos para extrair significado dos dados por meio de aprendizado de m quina e outras abordagens. Estat sticos podem se concentrar na matem tica dos modelos de risco e infer ncia [21]. Engenheiros de dados podem se concentrar na administra o e preserv o de dados gerados pelo cientista de dom nio e no *backend* do pipeline, seguindo a aquisi o, decis es e a es no dom nio da publica o, arquivamento e curadoria. Cientistas de dados podem unir o trabalho dos cientistas da computa o e dos estat sticos para extrair novos conhecimentos e conhecer tomadas de decis o mais eficientes.

2.2 Machine Learning

Aprendizado de M quina (*Machine Learning*, em ingl s) pode ser definido como "a pr tica de usar algoritmos para coletar dados, aprender com eles, e ent o fazer uma determina o ou predi o sobre alguma coisa no mundo. Ent o em vez de implementar as rotinas de software manualmente, com um gama espec fica de instru es para completar uma tarefa em particular, a m quina   'treinada' usando uma quantidade grande de dados e algoritmos que d o e ela a habilidade de aprender como executar a tarefa" [3]. Com isso, o computador consegue a habilidade de realizar determinado c lculo ou tarefa sem que necessite de program o adicional ou interfer ncia humana para isso.

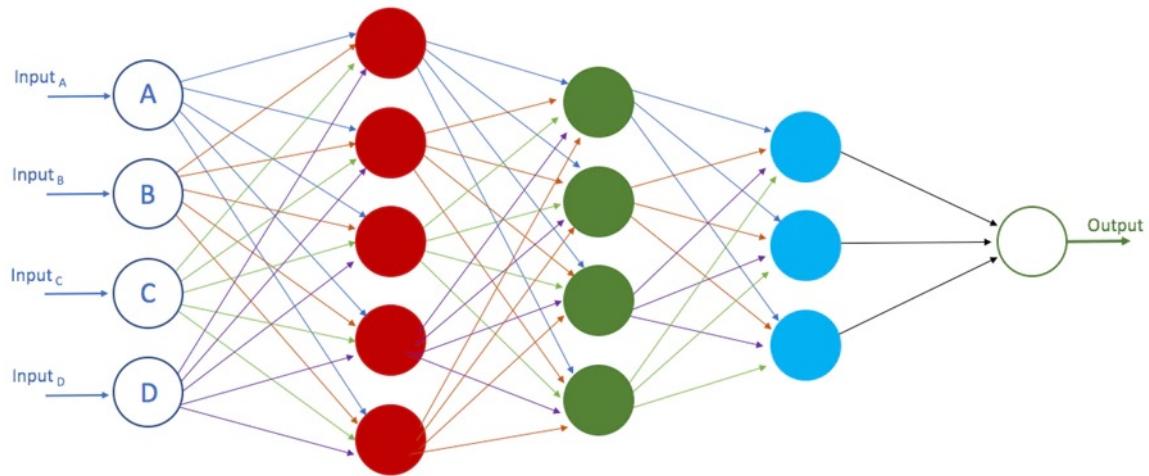
O *Machine Learning*   fortemente relacionado com a Estat stica, uma vez que seus m todos e parte de seus algoritmos, como regress es, tiveram como base modelos estat s-

ticos e a análise de seus dados. As tarefas de aprendizado podem ser classificadas em três categorias básicas [6] [5]:

- **Aprendizado supervisionado:** O treinamento é realizado por meio de exemplos rotulados, como uma entrada na qual a saída desejada é conhecida. Através de métodos como classificação, regressão e *gradient boosting*, o aprendizado supervisionado utiliza padrões para prever os valores de rótulos em dados não-rotulados adicionais. O aprendizado supervisionado é comumente empregado em aplicações nas quais dados históricos preveem eventos futuros prováveis.
- **Aprendizado não-supervisionado:** É utilizado em dados que não possuem rótulos históricos. A “resposta certa” não é informada ao sistema, o algoritmo deve descobrir o que está sendo mostrado. O objetivo é explorar os dados e encontrar alguma estrutura dentro deles. Técnicas populares incluem mapas auto-organizáveis, mapeamento por proximidade, agrupamento *k-means* e decomposição em valores singulares. Esses algoritmos também são utilizados para segmentar tópicos de texto, recomendar itens e identificar pontos discrepantes nos dados.
- **Aprendizado por reforço:** O algoritmo descobre através de testes do tipo “tentativa e erro” quais ações rendem as maiores recompensas. Este tipo de aprendizado possui três componentes principais: o agente (o aprendiz ou tomador de decisão), o ambiente (tudo com que o agente interage) e ações (o que o agente pode fazer). O objetivo é que o agente escolha ações que maximizem a recompensa esperada em um período de tempo determinado. O agente atingirá o objetivo muito mais rápido se seguir uma boa política, então o foco do aprendizado por reforço é descobrir a melhor política.

O termo *Machine Learning* se tornou muito mais evidente com a possibilidade da implementação do *Deep Learning*, que é uma técnica que utiliza Redes Neurais Artificiais para atingir seus resultados. Redes Neurais Artificiais são modelos computacionais inspirados no sistema nervoso do cérebro, onde temos neurônios divididos em camadas e conectados entre si, podendo ser abstruído conforme ilustração na Figura 2.2. Dependendo da tarefa a ser realizada, cada neurônio atribui um peso para os dados que entram e a saída final é determinada pelo total desses pesos [3]. As redes neurais utilizadas em *Deep Learning* possuem, ao menos, duas camadas de neurônios entre a camada que recebe os dados de entrada e a camada final que faz o tratamento final dos dados de saída.

Com a evolução da computação, o treino de uma tarefa passou a ser cada vez mais viável, uma vez que a execução de algoritmos de *Machine Learning* é computacionalmente muito custosa, especialmente quando redes neurais são utilizadas. E sua viabilidade é acompanhada de efetividade: Como exemplo, reconhecimento de imagens por máquinas treinadas através de deep learning em alguns cenários possuem uma taxa de acerto maior que a de humanos [3].

Figura 2.2: Exemplo de uma rede neural utilizada em *Deep Learning*.

2.2.1 Métricas de Avaliação

Tradicionalmente em um problema de classificação binária, uma previsão do classificador pode ter 4 tipos de resultados em uma matriz de confusão, presente na Tabela 2.2.1. Os Verdadeiros Positivos (VP, ou TP pelo termo em inglês *True Positives*) e Verdadeiros Negativos (VN, ou TN pelo termo em inglês *True Negatives*) são classificações corretamente previstas pelo classificador. Um Falso Positivo (FP) ocorre quando um caso previsto para estar na classe positiva quando o resultado real pertence à classe negativa, e um Falso Negativo (FN) ocorre quando um caso previsto para estar na classe negativa quando o resultado real pertence à classe positiva.

Tabela 2.1: Matriz de confusão

		Classe prevista	
		Positiva	Negativa
Classe atual	Positiva	Verdadeiros Positivos (VP/TP)	Falsos Positivos (FP)
	Negativa	Falso Negativo (FN)	Verdadeiros Negativos (VN/TN)

A taxa de sucesso, também conhecida como **acurácia**, é o número de previsões corretas dividido pelo número total de resultados:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

Ela é considerada um indicador da realização de um bom treinamento e de um bom funcionamento do modelo obtido. Entretanto, não é a melhor métrica para interpretar situações quanto a aplicação do modelo ao problema, como classes desequilibradas.

A **precisão** ajuda quando os custos de falsos positivos são altos e mostra a precisão das previsões positivas. É a fração de casos positivos previstos corretamente para estar na classe positiva de todos os casos positivos previstos no modelo:

$$P = \frac{TP}{TP + FP} \quad (2.2)$$

O **recall**, ou sensibilidade, ajuda quando o custo de falsos negativos é alto. É a fração de casos positivos previstos corretamente para estar na classe positiva de todos os casos positivos reais:

$$R = \frac{TP}{TP + FN} \quad (2.3)$$

O **F1-score** é uma medida geral da acurácia de um modelo que combina precisão e recall. Pode ser interpretado como uma ponderação entre ambos e usa a média harmônica para realizar a medição:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (2.4)$$

O **AUC** (*Area under the ROC Curve*), também chamada de precisão balanceada, pode interpretar a capacidade do classificador de evitar uma classificação falsa. Se sai melhor que a acurácia em situações que ela não é apropriada, como o desequilíbrio entre classes já citado anteriormente:

$$AUC = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (2.5)$$

2.2.2 Algoritmos

Existem diversos algoritmos de *Machine Learning* utilizados para resolver os problemas de aprendizado supervisionado, não-supervisionado e por reforço. Nesta seção serão abordados apenas os algoritmos utilizados por este trabalho de Mestrado.

Régressão Logística

A Régressão Logística é um método de classificação baseado na aplicação da técnica de Régressão Linear. Como na Régressão Linear, ele assume uma relação linear entre os recursos e calcula a soma ponderada dos recursos mais um termo de viés. Neste método, o resultado não é utilizado diretamente, mas calcula a logística do resultado. Quando w no peso das *features* e o termo de viés é b , a probabilidade estimada do modelo de régressão logística é a seguinte:

$$\hat{p} = \sigma(w^T x + b) \quad (2.6)$$

Logo após, a função logística σ é calculada:

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (2.7)$$

Uma vez que o modelo de Regressão Logística estimou a probabilidade \hat{p} , a previsão de classificação é feita facilmente considerando a seguinte regra:

$$\hat{p} = \begin{cases} 1 & \text{Se } \hat{p} \geq 0,5 \\ 0 & \text{Se } \hat{p} < 0,5 \end{cases} \quad (2.8)$$

O objetivo do treinamento de um modelo de classificação é de minimizar a diferença entre o resultado atual e o resultado previsto. Para medir o quanto próximo a função resultante do treinamento se aproxima do conjunto de dados, a seguinte função de custo é calculada conforme equação abaixo. A função utilizada pode variar, tendo como exemplos a função por entropia cruzada ou a função por diferença de quadrados.

$$J(w, b) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) \quad (2.9)$$

Durante o treinamento do modelo de Regressão Logística, é preciso encontrar os parâmetros w e b que minimizem a função de custos globais. Como a função de custo é convexa, diferentes métodos de otimização, como o gradiente descendente, garantem encontrar o mínimo global.

Para lidar com problemas com múltiplas classes para classificação, é possível calcular uma função de custo separada para cada rótulo de classe por observação e somar os resultados, técnica conhecida como *One-Vs-All*. Outra técnica que pode generalizar o método de regressão logística para suportar várias classes diretamente é chamada de Regressão Softmax, ou Regressão Logística Multinomial, utilizando a função Softmax para substituir a função de probabilidade da Regressão Logística convencional:

$$\hat{p}(Y_i = n) = \frac{e^{\beta_n \cdot \mathbf{x}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{x}_i}} \quad (2.10)$$

Support Vector Machines (SVM)

Support Vector Machines (SVM) é um algoritmo de aprendizado de máquina que pode ser usado para detecção linear, não linear, de classificação, regressão e até mesmo detecção de anomalias (*Outliers*). A ideia fundamental por trás dessa técnica é que as classes de saída são separadas com um hiperplano maximizando a margem (distância máxima entre os pontos de dados de ambas as classes) [71]. As *Support Vector Machines* criam um ou vários hiperplanos em um espaço de n dimensões. A dimensão dos hiperplanos depende do número de *features*. Quando há duas *features*, é apenas uma linha, ou quando há três *features*, é um plano bidimensional.

Para lidar com conjuntos de dados não lineares, uma abordagem é adicionar mais *features*, como um recurso polinomial, ou a outra abordagem é usar *kernels*, mapeando os dados de entrada de n dimensões para um espaço de dimensão superior, onde os dados podem ser separados linearmente.

Regressão Kernel Ridge

A Regressão Kernel Ridge (*Kernel Ridge Regression*/KRR) é um algoritmo de aprendizado de máquina proveniente da combinação de duas operações: A Regressão Ridge com o que é conhecido como "truque do kernel" [75]. A Regressão Ridge substitui a função de custo tradicional por uma com um termo de penalização incluído, conforme ilustrado na Equação 2.11, utilizando o método dos mínimos quadrados. O "truque do kernel" é uma técnica matemática que permite exemplificar problemas não-lineares de forma linear utilizando kernels, reduzindo a complexidade para uma simples operação matricial.

$$\sum_i (y_i - w^T x_i)^2 - \lambda \|w\|^2 \quad (2.11)$$

Por causa de tais operações, A Regressão Kernel Ridge exige mais processamento do que uma regressão tradicional. No entanto, é vantajoso usá-la em casos que um ajuste não-linear é desejado ou onde há mais atributos do que elementos no conjunto de treinamento. Em casos onde há mais elementos no conjunto de treinamento do que atributos, a Regressão Kernel Ridge peca por não abranger o conceito "vetores de suporte" utilizado nas *Support Vector Machines*, onde é necessário somar apenas o conjunto de vetores de suporte ao invés de todo o conjunto de treinamento. No entanto, as *Support Vector Machines* exigem mais processamento.

Árvores de Decisão (*Decision Trees*)

Árvores de decisão (*Decision Trees*) são um grupo de algoritmos de aprendizado de máquina que podem ser usados para classificação e regressão. Eles são os métodos de aprendizado mais comuns que são muito poderosos e capazes de ajustar conjuntos de dados complexos. Os algoritmos de Árvore de Decisão são baseados em uma abordagem de dividir e conquistar para os problemas de classificação [75]. Uma Árvore de Decisão é feita pelo processo contínuo de dividir o conjunto de dados nos atributos da melhor maneira possível em diferentes classes até que um critério de parada específico seja alcançado. Nas Árvores de Decisão, as observações sobre os itens são mostradas em ramificações e as conclusões das observações são mostradas nos nós. Existem três tipos diferentes de nós: os nós raiz que indicam o início do processo de decisão e não têm arestas de entrada, os nós internos que têm exatamente uma entrada e pelo menos duas arestas de saída e os nós finais (ou as folhas).

Sempre que o rótulo de classificação de destino assume valores discretos, a Árvore de Decisão é chamada de árvore de classificação e, sempre que recebe valores contínuos, é chamada de árvore de regressão. Um dos méritos do uso de Árvores de Decisão é que elas exigem pouco pré-processamento de dados e não há necessidade de dimensionamento ou centralização de dados. Treinando uma Árvore de Decisão, geralmente, é realizada com uma árvore menos complicada e mais abrangente. A complexidade de uma Árvore de Decisão pode ser controlada usando critérios de parada e métodos de poda [67], existindo quatro métricas diferentes para poder medi-la: o número total de nós, o número total de folhas, a profundidade da árvore e o número de atributos usados.

Os algoritmos usados para construir uma Árvore de Decisão a partir de um conjunto de

dados são chamados de indutores. Normalmente, o objetivo desses algoritmos é encontrar a Árvore de Decisão ótima minimizando o erro de generalização, considerando o número mínimo de nós e a profundidade mínima da árvore. Os algoritmos da Árvore de Decisão funcionam de forma *top-down*, pois escolhem a melhor variável em cada estágio que pode dividir o conjunto de dados em um atributo específico. Diferentes indutores usam critérios diferentes para encontrar a melhor variável.

Random Forest

Random Forest é um algoritmo de aprendizado de máquina que combina a simplicidade das Árvores de Decisão com a flexibilidade, resultando em melhorias na acurácia [25]. A principal ideia que o diferencia de uma Árvore de Decisão é o melhorar a redução de variância do *Bagging* por diminuição de correlação entre as árvores sem aumentar muito a variância, pois se considera apenas um subconjunto aleatório de variáveis a cada passo [25].

Bagging é uma técnica que gera uma coleção de classificadores introduzindo randomização na entrada do algoritmo, geralmente com excelentes resultados [75]. No *Random Forest*, ela é utilizada para gerar uma coleção de Árvores de Decisão simplificadas com a finalidade de generalizar o resultado no conjunto da obra.

Gradient Boosting

Gradient Boosting, também conhecido como *Gradient Boosting Machine* (GBM) ou *Gradient Boosted Regression Tree* (GBRT) [31], é um algoritmo de aprendizado de máquina que faz a classificação através da composição de pequenos modelos pela utilização do *Boosting* [41] [11]. *Gradient Boosting* faz uso de *Boosting* para gradualmente aproximar um melhor modelo, de modo a somar submodelos ao modelo composto. Árvores de Decisão tendem a gerar *overfitting*, e o *Gradient Boosting* é uma possibilidade para solucionar este problema.

Boosting é uma combinação de modelos simples, chamados de modelos fracos, onde tipicamente estes modelos são Árvores de Decisão. O algoritmo combina classificadores fracos com intuito de produzir um classificador forte. Diferente do *Bagging*, a criação de subconjuntos não é feita de maneira aleatória, e sim feita priorizando subconjuntos mal classificados [11].

Redes Neurais Artificiais

As Redes Neurais Artificiais, muitas vezes chamadas apenas de Redes Neurais, são um grupo de métodos de modelagem de dados estatísticos não lineares, que a princípio foram inspirados nos cérebros e nas estruturas dos neurônios biológicos. Elas são a base do aprendizado profundo. Eles são poderosos, escaláveis e capazes de trabalhar com grandes tarefas complexas de aprendizado de máquina, como serviços de reconhecimento de imagem e reconhecimento de fala.

As Redes Neurais Artificiais não são novas para os sistemas de computação, pois foram introduzidas pela primeira vez por Warren McCulloch e Walter Pitts em 1943 [56].

Desde então, o desenvolvimento dessas técnicas passou por altos e baixos, até ter uma adoção considerável graças aos avanços significativos no poder computacional tanto em hardware quanto em software, é possível treinar Redes Neurais complexas, e há uma enorme quantidade de dados disponíveis para uso em bancos de dados. As Redes Neurais podem ser divididas em dois grupos principais [69]:

- **Redes retroalimentadas (*Feedforward*):** Nessas redes, o fluxo de dados se move apenas na direção direta da camada de entrada para os nós de saída. Não há alimentação ou loop no sistema.
- **Redes recorrentes:** Este tipo de rede pode ter a opção de feedback e reutiliza os dados dos estágios posteriores para os estágios anteriores.

O processo simplificado de uma Rede Neural é o seguinte:

1. Os dados de entrada são fornecidos à rede, eles se propagam pelas camadas e o processo de encaminhamento produz a previsão.
2. Calcula-se o erro entre o produto previsto e o produto real (função de custo).
3. A Rede Neural usa um algoritmo de otimização para ajustar os pesos de forma a reduzir a função custo.
4. O processo de encaminhamento inicia novamente e continua até que a taxa de erro seja minimizada.

Algoritmos de otimização são métodos usados para minimizar o valor da função de custo ajustando os parâmetros internos do modelo. Algumas das técnicas de otimização mais comuns nas estruturas de aprendizado profundo são as seguintes: *Stochastic Gradient Descent* (SGD) [68], *Momentum* [63], *Nesterov Accelerated Gradient* (NAG) [73], *Adaptive Gradient* (AdaGrad) [35], *Root mean square prop* (RMSprop) [43], *Adaptive moment estimation*, (Adam) [53], *Nesterov and Adam optimizer* (Nadam) [34].

Toda Rede Neural consiste em alguns nós (neurônios), conexões ponderadas entre os nós e uma abordagem computacional chamada função de ativação usada para definir a saída de cada neurônio. Diferentes tipos de funções de ativação podem ser usados com esta técnica, como Sigmóide, tanh (tangente hiperbólica) ou ReLU (sigla de *Rectified Linear Unit*).

Redes Neurais consistem em diferentes camadas. O tipo mais simples de Rede Neural inclui uma camada de entrada que recebe informações de fontes externas, como valores de atributos do conjunto de dados de entrada. A camada de saída gera a saída da rede e as camadas ocultas que conectam a camada de entrada e a camada de saída entre si. O valor de entrada de cada nó em cada camada é calculado pela soma de todos os nós de entrada multiplicada pelo respectivo peso da interconexão entre os nós [37].

2.3 Fairness em Machine Learning

Como a coleta de dados está presente atualmente no dia-a-dia de variados setores da sociedade, o uso de *Machine Learning* é extremamente versátil para tomadas de decisão, podendo ser utilizados em problemas como admissão de universidades, contratações, análise de crédito e reconhecimento de doenças. Com o aumento dessa influência, o uso de dados sensíveis em um contexto determinado também aumentou, e temas como uma IA ética e conceitos como vieses nos dados e *Fairness* passaram a serem discutidas não apenas na Computação, mas em áreas como Direito. Algoritmos são mais objetivos, rápidos e são capazes de considerar uma grande magnitude de recursos que pessoas não são capazes. Entretanto, até o presente momento eles não são capazes de diferenciar contextos sociais, onde um resultado mais eficiente de acordo com os dados disponíveis podem amplificar as desigualdades sociais e tomar decisões de modo injusto [57].

Estes dados sensíveis, tendo como exemplos cor de pele, raça, sexo, idade e altura, são considerados atributos protegidos, que precisam ser classificados e processados antes da execução de um algoritmo de *Machine Learning*, determinarão como o algoritmo se comportará e, consequentemente, afetará suas métricas [60]. Os grupos de dados provenientes destes atributos protegidos são considerados grupos protegidos, que podem ser divididos em dois grupos: o grupo privilegiado, que possui vantagens no contexto do problema, e o grupo não-privilegiado, que possui desvantagens no contexto do problema e, portanto, sujeito a discriminação.

É possível descrever o conceito de *Fairness* no contexto de aprendizagem supervisionada, onde um modelo f pode prever um conjunto de resultados y a partir de um conjunto de *features* x , evitando discriminação injusta em relação a um atributo protegido a . É permitido, mas não exigido, que a seja um componente de x [20]. Em outras palavras, um modelo de ML considerado justo é aquele onde a correlação de seu resultado é baixa em relação a dados de entrada considerados como sensíveis a discriminações.

Geralmente, as descrições de justiça se dividem em dois grupos principais: justiça individual e justiça de grupo. O objetivo da justiça individual é que indivíduos semelhantes devem obter resultados semelhantes, enquanto na justiça de grupo, cada um dos grupos definidos pelo atributo protegido devem ser tratados igualmente. No geral, os estudos atuais costumam realizar seus experimentos em casos de justiça de grupo, uma vez que o escopo de justiça de grupo é muito mais amplo e tende a exemplificar melhor a relação entre dados, relações sociais e vieses do mundo atual.

2.3.1 Métricas de Fairness

Para avaliar a justiça de um modelo, as métricas utilizadas diferem das métricas utilizadas para avaliação do modelo, que possuem o propósito de verificar se um modelo tem previsões confiáveis ou não. As métricas de *Fairness* possuem um propósito diferente, pois verificam os dados de forma mais intimista. Elas não medem o modelo como um todo, mas o quanto os grupos e registros avaliados estão próximos dos outros. Enquanto as métricas mais tradicionais avaliam a performance do modelo e seus dados como um todo e seus resultados gerais, as métricas de *Fairness* avaliam se os resultados gerais também

se refletem em grupos específicos, para verificar se não há disparidade ou discriminação nos resultados propostos.

Assim como algumas métricas utilizadas para avaliação, muitas métricas utilizam Verdadeiros Positivos, Verdadeiros Negativos, Falsos positivos e Falsos Negativos para analisar o quanto justo o modelo é. Entretanto, diferente da acurácia, precisão e recall utilizados anteriormente, a medição das discriminações utiliza outras métricas, utilizadas ou não para avaliar a performance, para estabelecer novas métricas mais adequadas para a sua finalidade.

Exemplos de métricas utilizadas para isso são a Taxa de Verdadeiros Positivos e a Taxa de Falsos Positivos. Enquanto a **Taxa de Verdadeiros Positivos** (TVP, ou TPR) pelo termo em inglês **True Positive Rate**) é outro termo para denominar o recall, a **Taxa de Falsos Positivos** (TFP, ou FPR pelo termo em inglês **False Positive Rate**) é a fração de casos negativos previstos incorretamente como estando na classe positiva de todos os casos positivos reais:

$$FPR = \frac{FP}{FP + TN} \quad (2.12)$$

Dada essas métricas iniciais, considerando $Y = 1$ a classe positiva, $Z = 0$ o grupo não-privilegiado e $Z = 1$ o grupo privilegiado, algumas das definições de *Fairness* mais usadas são as seguintes:

- **Diferença de paridade estatística (*Statistical parity difference*), ou discriminação [77]:** Esta métrica é baseada na seguinte fórmula:

$$Pr(Y = 1|Z = 0) - Pr(Y = 1|Z = 1) \quad (2.13)$$

Aqui, o viés ou paridade estatística é a diferença entre a probabilidade de que um indivíduo aleatório retirado dos não-privilegiados seja rotulado como 1 e a probabilidade de que um indivíduo aleatório dos privilegiados seja rotulado como 1. Portanto, um valor próximo de 0 é considerado justo.

- **Diferença de oportunidade igual (*Equal opportunity difference*) [22]:** É a diferença entre a taxa positiva verdadeira do grupo não privilegiado e a taxa positiva verdadeira do grupo privilegiado:

$$TPR_{Z=0} - TPR_{Z=1} \quad (2.14)$$

Um valor próximo de 0 é considerado justo. Um classificador binário satisfaz a igualdade de oportunidades quando a taxa positiva verdadeira de ambos os grupos são iguais [44]

- **Diferença de probabilidade média (*Average odds difference*) [22]:** Essa métrica usa a taxa de falsos positivos e a taxa positiva verdadeira para calcular a tendência, calculando a igualdade de probabilidades com a fórmula:

$$\frac{1}{2}(|FPR_{Z=0} - FPR_{Z=1}| + |TPR_{Z=0} - TPR_{Z=1}|) \quad (2.15)$$

Precisa ser próximo a 0 para ser considerado justo.

- **Impacto de disparidade (*Disparate impact*) [22]:** Para esta métrica, é usada a seguinte fórmula:

$$\frac{Pr(Y = 1|Z = 0)}{Pr(Y = 1|Z = 1)}$$

Usa as mesmas probabilidades da diferença de paridade estatística, mas aqui são calculadas como proporção. Desta forma, um valor próximo de 1 é considerado justo.

- **Índice de Theil (*Theil index*) [70]:** Esta medida também é conhecida como índice de entropia generalizado, mas com α igual a 1 [70]. É calculado com a seguinte fórmula:

$$\frac{1}{n} \sum_{i=0}^n \frac{b_i}{\mu} \ln \frac{b_i}{\mu}$$

Onde $b_i = \hat{y}_i - y_i + 1$, y_i é o conjunto de saídas e \hat{y}_i é o conjunto de previsões dadas pelo modelo. Também precisa ser próximo a 0 para ser considerado justo.

2.3.2 Algoritmos para redução de vieses

Há diversos tipos de algoritmos diferentes na inteligência artificial para redução de vieses, a fim de garantir *Fairness* nos projetos de Aprendizado de Máquina. É possível classificá-los em três categorias diferentes: Algoritmos de pré-processamento, processamento e de pós-processamento.

Os algoritmos de **pré-processamento** tentam eliminar a discriminação transformando os dados, antes de executar o algoritmo de treinamento. Tais algoritmos podem ser usados caso seja permitida a modificação dos dados de treinamento [32], e a ideia por trás de tais algoritmos é que suas previsões serão mais平衡adas se o classificador for treinado com os dados já balanceados [28]. Nesta categoria se enquadram os seguintes algoritmos:

- **Reposição (*Reweighting*) [45]:** Pondera os exemplos em cada combinação de grupo e rótulo de maneira diferente para garantir a justiça antes da classificação.
- **Removedor de impacto de disparidade (*Disparate impact remover*) [38]:** Edita valores de *features* aumentando a justiça de cada grupo enquanto preserva a ordem de classificação dentro dos mesmos.
- **Aprendizado de representações justas (LFR, ou *Learning fair representations*) [77]:** Encontra uma representação latente que codifica os dados, mas ofusca informações dos atributos protegidos.

- **Pré-processamento otimizado (Optimized pre-processing) [29]:** Aprende uma transformação probabilística que edita *features* e rótulos nos dados efetuando justiça em cada grupo, distorção individual, garantindo a fidelidade de dados através de restrições.

Os algoritmos de **processamento** tentam realizar modificações nos algoritmos de treinamento para mitigar a discriminação durante o processo de treinamento do modelo. Se for permitido fazer mudanças no processo de treinamento, então os algoritmos podem ser usados incorporando mudanças na função de custo ou impondo restrições [?]. Nesta categoria se enquadram os seguintes algoritmos:

- **Remoção de viés adversário (Adversarial debiasing) [78]:** Aprende um classificador que maximiza a precisão e reduz a capacidade de um adversário de depender do atributo protegido nas previsões. Essa abordagem leva a um classificador justo, pois as previsões realizadas pelo classificador não possuem nenhuma informação de discriminação nos grupos.
- **Removedor de preconceito (Prejudice remover) [38]:** Técnica que adiciona no algoritmo escolhido um termo de regularização baseado na discriminação (No caso da biblioteca AI Fairness 360 é usado o programa da publicação, que usa a regressão logística como algoritmo base).
- **Meta-Algoritmo para classificações justas (Meta-Algorithm for Fair Classification) [30]:** Aprende um classificador compatível com uma gama grande de métricas de Fairness, sendo prático o suficiente para abrangê-las sem grande perda de performance.
- **Justiça por Subgrupos Ricos (Rich Subgroup Fairness) [48]:** Aprende um classificador que procura equalizar as taxas de falsos positivos e falsos negativos entre os dados que envolvem atributos protegidos, considerados como subgrupos.
- **Redução por Gradiente Exponencial (Exponentiated Gradient Reduction) [17]:** Aprende um classificador baseado em Gradiente Exponencial que tende a minimizar o erro de uma classificação ponderada.
- **Redução por busca em grid (Grid Search Reduction) [17] [18]:** Aprende um classificador baseado na busca em um grid de valores que tende a minimizar o erro de uma classificação ponderada. É mais simples e impreciso que a Redução por Gradiente Exponencial, mas sua escolha pode ser razoável se a quantidade de métricas de Fairness a serem consideradas for pequena.

Os algoritmos de **pós-processamento** utilizam um conjunto de validação, que não foi envolvido no processo de treinamento para melhorar a imparcialidade das previsões [32]. Quando não há possibilidade de fazer alterações nos dados de treinamento ou no treinamento do modelo, apenas algoritmos de pós-processamento podem ser usados. Nesta categoria se enquadram os seguintes algoritmos:

- **Igualdade de probabilidade calibrada (Calibrated Equalized odds) [62]:** Otimiza as previsões do classificador obtido, calibrando para alterar os rótulos de saída e obter probabilidades igualadas entre os grupos.
- **Igualdade de probabilidade (Equalized odds) [44]:** Resolve um problema linear para alterar os rótulos de saída e obter probabilidades igualadas entre os grupos.
- **Classificação baseada em Rejeição de Opções (Reject Option-based Classification) [46]:** Dá resultados favoráveis para grupos não privilegiados e resultados desfavoráveis para grupos privilegiados de acordo com uma faixa de confiança.

2.4 Engenharia de Software

2.4.1 Engenharia de Software para Aplicações de IA

Quando se fala de Arquitetura e Engenharia de Software, se fala da definição dos componentes de software, suas propriedades externas, e seus relacionamentos com outros softwares para fazer com que um sistema seja documentável, reusável e testável. A preocupação está em como um sistema deve ser organizado e com a estrutura geral desse sistema. Dado as definições sobre IA já detalhadas, é possível encaixar Machine Learning na forma de um processo bem definido, de forma que é possível sistematizar todo esse processo na forma de componentes e definir formas em que o modelo resultante do mesmo é disponibilizado para aplicações externas.

No processo, ilustrado na Figura 2.3, o conjunto de dados passa por um pré-processamento e dividido em dois conjuntos, um para treinamento e outro para teste. O conjunto de treino é utilizado para o algoritmo realizar o processo de treinamento, obtendo um modelo após o término desse processo. O conjunto de testes é utilizado para mensurar se o modelo obtido no processo de treinamento realiza previsões confiáveis ou não.

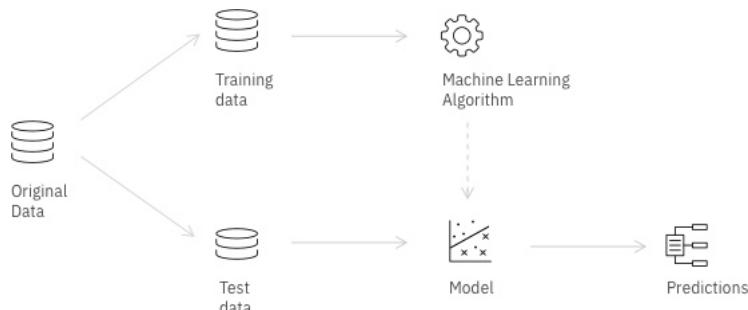


Figura 2.3: Processo padrão para aprendizado de máquina

Um exemplo de arquitetura que generaliza todo o processo, desde a necessidade de negócio até o deploy do modelo de IA, é a IBM Analytics and AI Reference Architecture [4], ilustrada na Figura 2.4. Nela, são definidos os seguintes requisitos não-funcionais: Performance, estabilidade, segurança, escalabilidade, manutenibilidade e regulamentações de privacidade/*compliance*, e pode ser classificada em 4 grupos principais envolvendo diversos tipos de processos e componentes:

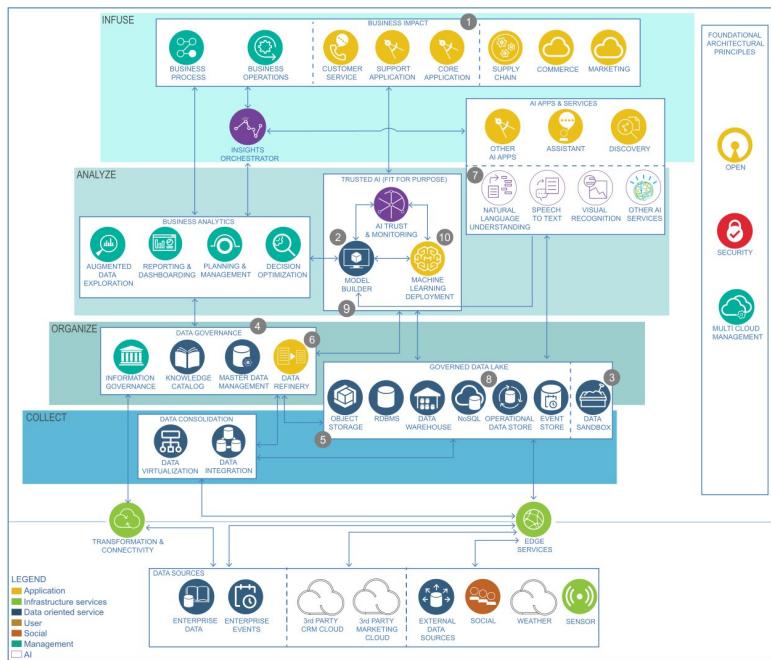


Figura 2.4: IBM Analytics and AI Reference Architecture.

- **Coleta:** Relaciona os processos de coleta, armazenamento e transformação de diversas fontes de dados, estruturadas ou não estruturadas, para determinados repositórios (*Data Lakes*).
- **Organização:** Relaciona os processos de organização e estruturação dos dados nos presentes nos *Data Lakes* e necessários para o uso das aplicações que envolvem a análise dos dados. Dependendo do uso pode-se aplicar processos de Governança.
- **Análise:** Relaciona os processos para desenvolvimento de aplicações de IA e relatórios após a organização dos dados para tomadas de decisão e uso de aplicações externas.
- **Infusão:** Relaciona os processos de disponibilização desses dados e conhecimento obtidos na fase de análise para aplicações externas.

2.5 Proveniência de Dados

O termo proveniência é comumente usado no contexto da arte para denotar a história documentada ou a cadeia de propriedade de um objeto de arte [58] [59] [36] [74]. A proveniência ajuda a determinar a autenticidade e, portanto, o valor dos objetos de arte. Passando para o contexto da computação e do processamento de dados, a proveniência dos dados, às vezes chamada de linhagem ou *pedigree*, é a descrição das origens de um dado e o processo pelo qual ele chegou a um banco de dados [26], contendo metadados informando "como", "quando", "onde", "por que" ele foi obtido e "quem" o obteve. Em outras palavras, o conceito, não somente inclui a origem do dado (identificação, responsável pelo dado, data de criação), mas também os processos aplicados a ele (algoritmos e os

parâmetros utilizados para executá-lo) [76]. Os principais benefícios da proveniência para a qualidade de dados são [24]:

- **Comunica a qualidade de dados:** confiabilidade, adequação, acurácia, atualidade, redundância;
- Melhora a interpretação do dado: em relação a função do reconhecimento da fonte e na utilização do dado para um aspecto de tomada de decisão.
- **Justificativa do uso de um determinado dado:** em relação as limitações e intenções originais do uso de um determinado dado de conjuntos de dados ambientais.
- **Redução de erros:** no quesito do juízo da precisão do dado, no acompanhamento preciso da linhagem de conjuntos de dados científicos.
- **Passos do processamento:** permite que usuários não especialistas em dados entendam a capacidade de recuperar e entender os relacionamentos entre produtos de dados, scripts ou dados gerados por programas.
- **Criação de dados científicos:** permite identificar o processo utilizado para ajudar a identificar e avaliar os componentes básicos dos sistemas que fornecem a recuperação de linhagem para produtos de dados científicos.
- **Atualização de dados:** permite a partir do desenvolvimento de estudos formais para executar rastreamento de linhagem de dados em visões relacionais.
- **Modificação de *schemas* de visões relacionais:** modelos gráficos e estudos experimentais.
- **Fontes de dados históricas:** permite identificar a origem e o subsequente histórico de processamento.

Tanto a comunidade científica quanto a empresarial adotaram o estilo de arquitetura orientada a serviços (SOA), que permite que os serviços sejam descobertos e compostos dinamicamente [59]. Os aplicativos baseados em SOA tornam-se mais dinâmicos e abertos, mas também precisam atender a novos requisitos. É preciso verificar se o processo que os trouxe resultados está em conformidade com regulamentações ou metodologias específicas, provar que os resultados são derivados independentemente de serviços ou bancos de dados com determinadas restrições de licença, e estabelecer que os dados foram capturados na fonte por instrumentos que ofereçam confiabilidade.

Como tais verificações não são automatizadas, há sempre a possibilidade de erro humano: uma etapa pode não ser feita ou feita parcialmente, por uma falha em alguma verificação, ou mesmo por falta de suporte. Os dados podem não conter as informações históricas necessárias para fazerem as verificações. Portanto, há a necessidade de capturar informações extras (documentação do processo) que descrevam o que realmente ocorreu durante a execução. A documentação do processo é para os dados o que um registro de propriedade é para uma obra de arte. Os aplicativos capazes de realizar proveniência criam a documentação do processo e a armazenam em uma base de proveniência, cujo

papel é oferecer um armazenamento seguro e persistente de longo prazo da documentação do processo, conforme ilustrado na Figura 2.5. Esse papel acomoda várias implantações: por exemplo, uma base de proveniência pode ser um serviço único e autônomo ou, para ser mais escalável, pode ser uma coleção de bases distribuídas [59].

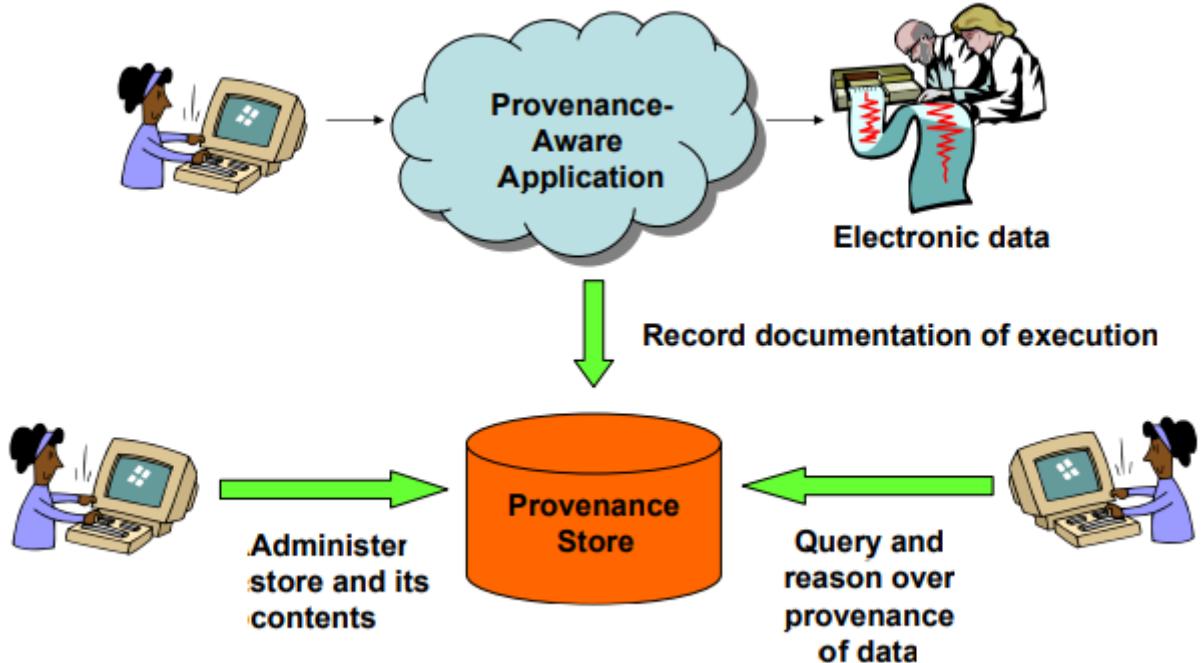


Figura 2.5: Ciclo de vida da proveniência. [59]

Uma vez registrada a documentação do processo, o resultado da proveniência dos dados pode ser recuperada consultando a base de proveniência e analisada para atender as necessidades do usuário da aplicação. Com o tempo, é preciso gerenciar e manter a base de proveniência e o conteúdo já obtido. Dado esse ponto, é possível dividir o ciclo de vida da proveniência de dados em quatro fases diferentes, sendo válido para todos os sistemas capazes de realizá-la [59]:

- Criação
- Registro
- Consulta
- Gerenciamento

É possível utilizar *workflows* para gerenciar a proveniência [33]. No domínio científico, um *workflow* é normalmente usado para executar tarefas complexas de processamento de dados. Um *workflow* pode ser pensado como um programa que é uma série de etapas de computação e etapas definidas por processos executados manualmente por uma ou mais pessoas. A proveniência do *workflow* refere-se ao registro de todo o histórico da saída final do *workflow* [74]. Nesse contexto, há duas formas distintas de proveniência [33]:

- **Prospectiva:** Trata-se da sequência de processos utilizados para a geração do dado, ou seja, a captura dos passos que devem ser seguidos para a geração de um dado produto.
- **Retrospectiva:** Trata-se das informações obtidas durante a execução dos processos de geração do dado. Compreende desde o tempo de duração de cada atividade executada até a origem dos dados de entrada. Além disso, não depende do tratamento da proveniência prospectiva para ser utilizado. Em outras palavras, é como se fosse um *log* detalhado da execução de uma tarefa.

Um importante componente da proveniência é a obtenção de informações sobre causalidade. Nesse componente é guardada a descrição do processo, ou a sequência das etapas, que, junto com dados de entrada e seus respectivos parâmetros, levam à criação de uma base de dados. Assim as dependências dos processos são usadas para documentar sua criação, bem como auxiliar na reprodução e validação desse processo. A causalidade pode ser inferida tanto para a forma prospectiva como para a forma retrospectiva de proveniência [33].

Outro componente-chave para a proveniência são as informações definidas pelo usuário, a documentação. Por serem dados, em geral, advindos dos processos de anotação, não são capturados automaticamente. Com isso, possuem diferentes níveis de granularidade e podem estar associados tanto para a forma prospectiva como para a forma retrospectiva de proveniência. Esse tipo de registro se torna muito importante, pois contém informações sobre decisões tomadas e observações feitas pelo usuário [33].

2.6 Arquitetura de Software

O consenso sobre a definição de arquitetura de software foi adotado com a adoção do padrão IEEE 1471, que define arquitetura de software como "*a organização fundamental de um sistema incorporado em seus componentes, seus relacionamentos entre si e com o meio ambiente e os princípios que orientam seu projeto e evolução*". Com esta definição, o componente e o conector são reforçados como conceitos centrais da arquitetura de software [23].

O nível de design da arquitetura de software em um projeto vai além dos algoritmos e estruturas de dados da computação. Incluem fatores como organização, protocolos de comunicação, acesso a dados, atribuição de funcionalidades, escalabilidade, performance, composição e seleção do *design* ideal [42]. É possível tratar uma arquitetura de um sistema específico como uma coleção de componentes juntamente com uma descrição dos conectores, que define as interações entre os componentes.

Um estilo de arquitetura define uma família de tais sistemas em termos de um padrão de organização estrutural, e determina o vocabulário de componentes e conectores que podem ser usados em instâncias desse estilo, juntamente com um conjunto de restrições sobre como eles podem ser combinados [42]. A decisão sobre tal estilo depende da solução e dos requisitos de um sistema. Ela pode adicionar novos componentes, incrementá-los com novos requisitos ou adicionar restrições sobre eles [23].

2.6.1 Arquitetura MAPE-K

Em 2001, Paul Horn introduziu o conceito de Computação Autônoma como alternativa a solução para a crescente complexidade dos sistemas da época, onde previa-se que os mesmos se tornariam muito grandes e complexos até mesmo para os profissionais mais qualificados configurarem e realizarem manutenção. Tal conceito qualifica sistemas de computação que podem se autogerenciar com relação aos objetivos de alto nível dados pelos administradores e é derivado da biologia, dado a grande variedade e hierarquia de sistemas autônomos presentes na natureza e na sociedade [49].

Em um ambiente autônomo e autogerenciado, os componentes de sistema podem incorporar como funcionalidade um *loop* de controle. Embora estes *loops* sejam divididos nos mesmos procedimentos, é possível categorizá-los em 4 categorias principais. Essas categorias são consideradas atributos dos componentes do sistema e são definidas como [10]:

- **Auto-configuração:** Pode se adaptar dinamicamente a mudanças no ambiente. Um componente autoconfigurável realiza esta adaptação usando políticas fornecidas pelo profissional. Tais mudanças podem incluir a implantação de novos componentes ou a remoção dos existentes, ou mudanças drásticas nas características do sistema. A adaptação dinâmica ajuda a garantir força e produtividade contínuas da infraestrutura, resultando em crescimento e flexibilidade dos negócios.
- **Auto-cura:** Pode descobrir, diagnosticar e reagir a interrupções. Um componente auto-curável pode detectar falhas no sistema e iniciar ações corretivas baseadas em políticas sem interromper o ambiente. A ação corretiva pode envolver um produto alterando seu próprio estado ou efetuando mudanças em outros componentes do ambiente. Com isso, o sistema se torna mais resiliente porque as operações cotidianas possuem menos probabilidade de falhar.
- **Auto-otimização:** Pode monitorar e ajustar recursos automaticamente. Um componente auto-otimizável pode se ajustar para atender às necessidades do usuário. As ações de ajuste podem significar realocar recursos para melhorar a utilização geral, como em resposta a cargas de trabalho que mudam dinamicamente, ou garantir que processamentos possam ser concluídos em tempo hábil. A auto-otimização ajuda a fornecer um alto padrão de serviço para quem vai utilizar o sistema. Sem funções de auto-otimização, não há uma maneira fácil de re-escalonar os recursos de infraestrutura quando um aplicativo não os usa totalmente.
- **Auto-proteção:** Pode antecipar, detectar, identificar e proteger contra ameaças de qualquer lugar. Um componente de autoproteção pode detectar comportamentos hostis à medida que ocorrem e tomar ações corretivas para se tornarem menos vulneráveis. Os comportamentos hostis podem incluir acesso e uso não autorizados, infecção e proliferação de vírus e ataques de negação de serviço. Os recursos de autoproteção permitem que as empresas apliquem consistentemente políticas de segurança e privacidade.

Para a Computação Autônoma acontecer, é implementado um Elemento Autônomo [16], um componente de software que gerencia partes do sistema baseando-se em um *loop*

MAPE-K (*Monitor, Analyze, Plan, Execute, and Knowledge*), ilustrado na Figura 2.6. O MAPE-K é um conceito que constitui um *loop* de controle, usado para monitorar e controlar um ou mais elementos gerenciados. Um elemento gerenciado (*Managed Element*) pode ser um hardware, como uma impressora, um software, como um banco de dados, outro Elemento Autônomo ou funções específicas, como balanceamento de carga. Um *loop* de controle MAPE-K é dividido da seguinte forma:

- **Monitoramento (Monitor):** Esta parte é responsável por monitorar os recursos gerenciados e coletar, agregar e filtrar dados. O monitoramento é feito por meio de um sensor (*Sensor*) ou mais sensores.
- **Análise (Analyze):** Analisa os dados relatados pela parte do monitor. A análise visa compreender qual é o estado atual do sistema e se há medidas para serem tomadas.
- **Planejamento (Plan):** Um plano de ação é preparado no base dos resultados da análise. O plano é uma série de medidas que irão mover o sistema de seu estado atual para um estado desejado.
- **Execução (Execute):** O plano é executado e controlado. Um efetor (*Effector*) ou mais executam as ações planejadas no recurso.
- **Conhecimento (Knowledge):** A base de conhecimento é central e acessível por todas as partes do *loop*. Separado a partir de dados coletados e analisados, ele contém conhecimento adicional, como modelos de arquitetura, modelos de metas, políticas e planos de mudança.

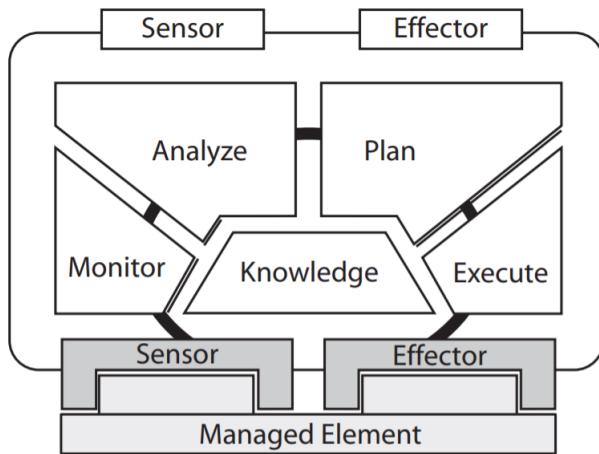


Figura 2.6: Diagrama de funcionamento da arquitetura MAPE-K [16].

2.6.2 Arquitetura *Pipe-and-Filter*

Em uma arquitetura *Pipe-and-Filter*, ilustrado na Figura 2.7, cada componente tem um conjunto de entradas e um conjunto de saídas. Um componente lê *streams* (ou fluxos) de

dados em suas entradas e produz *streams* de dados em suas saídas, abstraindo a entrega de um resultado como um todo. O *stream* de entrada é transformado de modo que a saída comece a ser produzida antes da entrada ser completamente consumida. Por isso, os componentes são chamados de filtros (*filters*). Os conectores deste estilo servem como condutores para os *streams*, transmitindo as saídas de um filtro para as entradas de outro. Por isso, os conectores são chamados de tubos (*pipes*) [42].

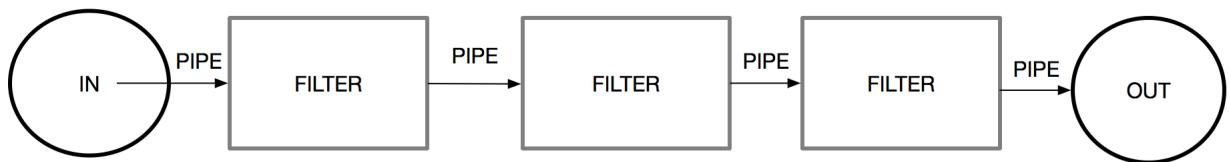


Figura 2.7: Diagrama de uma arquitetura *Pipe-and-Filter*.

Os filtros devem ser independentes, isto é, não devem compartilhar estado com outros filtros e, durante a programação, o ideal é que os filtros independam da ordem de processamento. Suas especificações podem restringir os dados transportados nos *pipes* de entrada e nos *pipes* de saída, mas não conhecem quaisquer outros filtros, ou componentes, conectados por seus *pipes*. Especializações comuns desse estilo incluem *pipelines*, que restringem as topologias a sequências lineares de filtros; *Pipes* limitados, que restringem a quantidade de dados que podem ser passados em um *pipe*, e *pipes* tipados, que exigem que os dados passados entre dois filtros tenham um tipo bem definido.

O uso da arquitetura *Pipe-and-Filter* possui como vantagens:

- Permitem que o Arquiteto de Software/Desenvolvedor entendam o comportamento geral de entrada/saída de um sistema como uma composição simples dos comportamentos dos filtros individuais.
- Suportam a reutilização: quaisquer dois filtros podem ser conectados, desde que concordem com os dados que estão sendo transmitidos entre eles.
- Os sistemas podem ser facilmente mantidos e aprimorados: novos filtros podem ser adicionados a sistemas existentes e filtros antigos podem ser substituídos por outros melhorados.
- Permitem certos tipos de análise especializada, como análise de rendimento e de impasse.
- Naturalmente suportam a execução simultânea: Cada filtro pode ser implementado como uma tarefa separada e potencialmente executado em paralelo com outros filtros.

Como desvantagens, é possível citar:

- Geralmente levam a uma organização de processamento em lote. Embora os filtros possam processar dados de forma incremental, uma vez que os filtros são inerentemente independentes, o Arquiteto de Software é forçado a pensar em cada filtro como fornecendo uma transformação completa dos dados de entrada em dados de saída.
- Por sua natureza de transformação de dados, os sistemas que usam a arquitetura *Pipe-and-Filter* normalmente não são bons para lidar com aplicativos interativos. Esse problema é mais grave quando são necessárias atualizações de exibição incrementais, porque o padrão de saída para atualizações incrementais é radicalmente diferente do padrão para saída de filtro.
- Podem ser prejudicados por terem que manter correspondências entre dois *streams* separados, mas relacionados.
- Podem forçar um resultado médio na transmissão de dados em situações onde muitos filtros sejam encadeados com um único filtro, resultando em trabalho adicional para cada filtro separar seus dados e analisar o que for necessário. Isso, por sua vez, pode levar tanto à perda de desempenho quanto ao aumento da complexidade na escrita dos próprios filtros.

Capítulo 3

Metodologia

3.1 Detalhamento do processo

Dado as etapas da AI Reference Architecture, foram definidos os seguintes papéis onde um projeto de Machine Learning pode ter atuação e onde se encaixariam:

- **Especialista de domínio:** É a pessoa que detém de todo o conjunto de regras do qual a aplicação deve respeitar. Pode não ter conhecimento técnico, embora esse conhecimento possa ajudar na comunicação das regras com os demais papéis. Está presente nas fases de Coleta, Organização e Infusão do ciclo.
- **Engenheiro de Dados:** É a pessoa responsável pelos processos de coleta e transformação dos dados para o uso em outros processos, sejam eles de Softwares tradicionais ou aplicações de Machine Learning. Pode aplicar processos de governança antes de definir que o dado esteja pronto para ser usado por outras pessoas. Está presente nas fases de Coleta e Organização do ciclo.
- **Cientista de Dados:** É a pessoa responsável pela análise dos dados e do desenvolvimento do processo de Machine Learning após a transformação e tratamento dos dados. Pode realizar tratamentos próprios antes do treinamento, como *Encoding* (*Label Encoding/One-hot Encoding*), normalização, processos de regularização como aumento de dados, para melhorar a performance do mesmo. Está presente nas fases de Organização e Análise do ciclo.
- **Engenheiro de Software:** É a pessoa responsável por usar o modelo de Machine Learning obtido na fase de Análise em aplicações que façam sentido para seu uso, como assistentes, automações, dashboards e relatórios. Está presente apenas na fase de Infusão, mas pode ser considerado na fase de Análise para verificar com o Cientista de Dados como está o andamento dos modelos desenvolvidos e desenhar alternativas caso os mesmos não estejam prontos para uso.

Dado esses papéis e suas respectivas funções, foi desenhado um diagrama de atividades, presente na Figura 3.1, determinando como eles se encaixariam no processo. Como o foco está na parte de Machine Learning, o detalhamento maior ficará na parte responsável pelo Cientista de Dados.

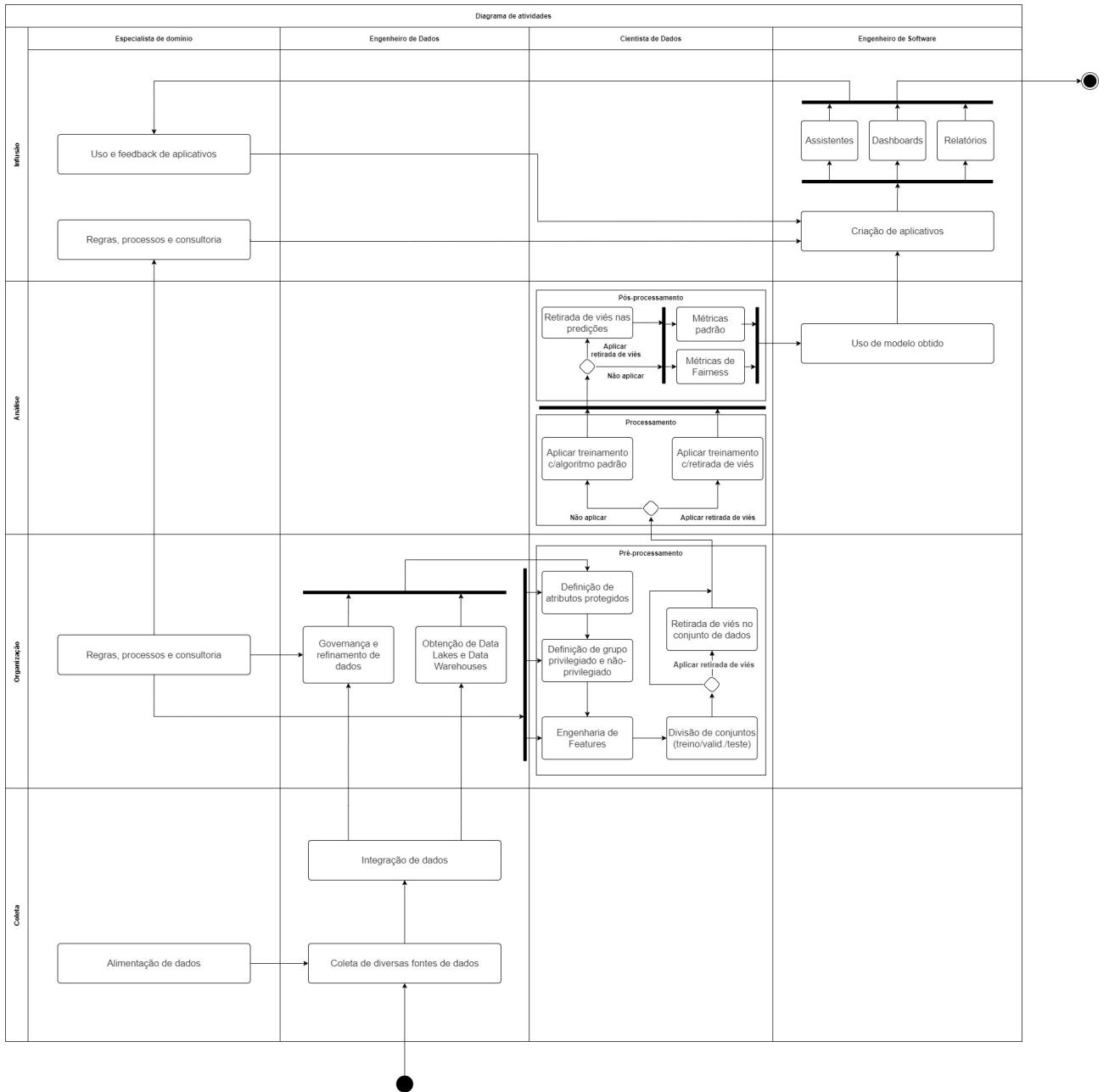


Figura 3.1: Diagrama de atividades com subdivisão de cada papel em uma aplicação de IA utilizando métricas de Fairness, com base na IBM AI Reference Architecture [4]

O desenvolvimento realizado neste trabalho pode ser dividido em 5 etapas maiores, que podem ser detalhadas e sub-divididas:

- **Obtenção dos conjuntos de dados:** Foi realizada uma pesquisa dos conjuntos de dados utilizados como experimento envolvendo algoritmos de redução de vieses e, após uma quantidade de conjuntos de dados, os mesmos foram encontrados e obtidos para serem analisados.
- **Transformação dos conjuntos de dados:** Como alguns dos conjuntos de dados obtidos diferentes valores qualitativos e codificados de acordo com o atributo,

primeiro foi realizado uma transformação dos dados para garantir uma melhor legibilidade, o que já equivaleria minimamente ao trabalho do Engenheiro de Dados presente neste trabalho de Mestrado. Posteriormente, novas transformações foram realizadas devido a necessidade de separar grupo privilegiado ao grupo não-privilegiado, mas neste trabalho já equivaleria a parte do Cientista de Dados.

- **Desenvolvimento do Pipeline:** Como Machine Learning possui um processo muito bem definido, é possível subdividir este processo em etapas e subdividir cada uma dessas etapas em componentes isolados. O objetivo do desenvolvimento de um Pipeline é sistematizar todo esse processo de modo que supostas atualizações sejam incrementais e que seja possível realizar uma execução de forma autônoma a partir da próxima etapa. Para simplificar o processo, foi utilizada a arquitetura *Pipe-and-Filter*, que é simples de ser entendida e pode ser encontrada em diferentes publicações.
- **Desenvolvimento dos processos de autonomia do Pipeline:** Após o desenvolvimento do Pipeline, foi desenvolvido um elemento autônomo para trabalhar em conjunto com o Pipeline como elemento gerenciado. Através das métricas obtidas em execuções anteriores, é possível inferir sugestões de melhores combinações para execução e garantia de um melhor resultado de forma mais eficiente. Para garantir uma maior flexibilidade nos resultados, foram implementadas diversas estratégias para garantir que essa sugestão seja customizada pelo Cientista de Dados e/ou pelo Especialista de Domínio se isso for necessário para as necessidades do problema.
- **Interface Humano-Computador:** Para facilitar o entendimento dos arquivos utilizados para configuração e utilização das estratégias da arquitetura MAPE-K, foi criada uma interface onde é possível realizar a configuração pela mesma, podendo também realizar uma execução do Pipeline isolada e uma execução com sugestões, acompanhando a análise realizada no gerenciador MAPE-K.

Para definição dos objetivos do Pipeline e da sua autonomia, foi realizado um detalhamento na forma de *Assurance Cases*, onde o objetivo principal, que é a obtenção do modelo mais equilibrado entre métricas de avaliação, onde para melhor diferenciação de seu objetivo serão denominadas como métricas de Performance, e métricas de Fairness, é subdividido em diferentes estratégias, novamente subdividido em novos objetivos, e através de evidências é possível verificar o progresso do objetivo principal.

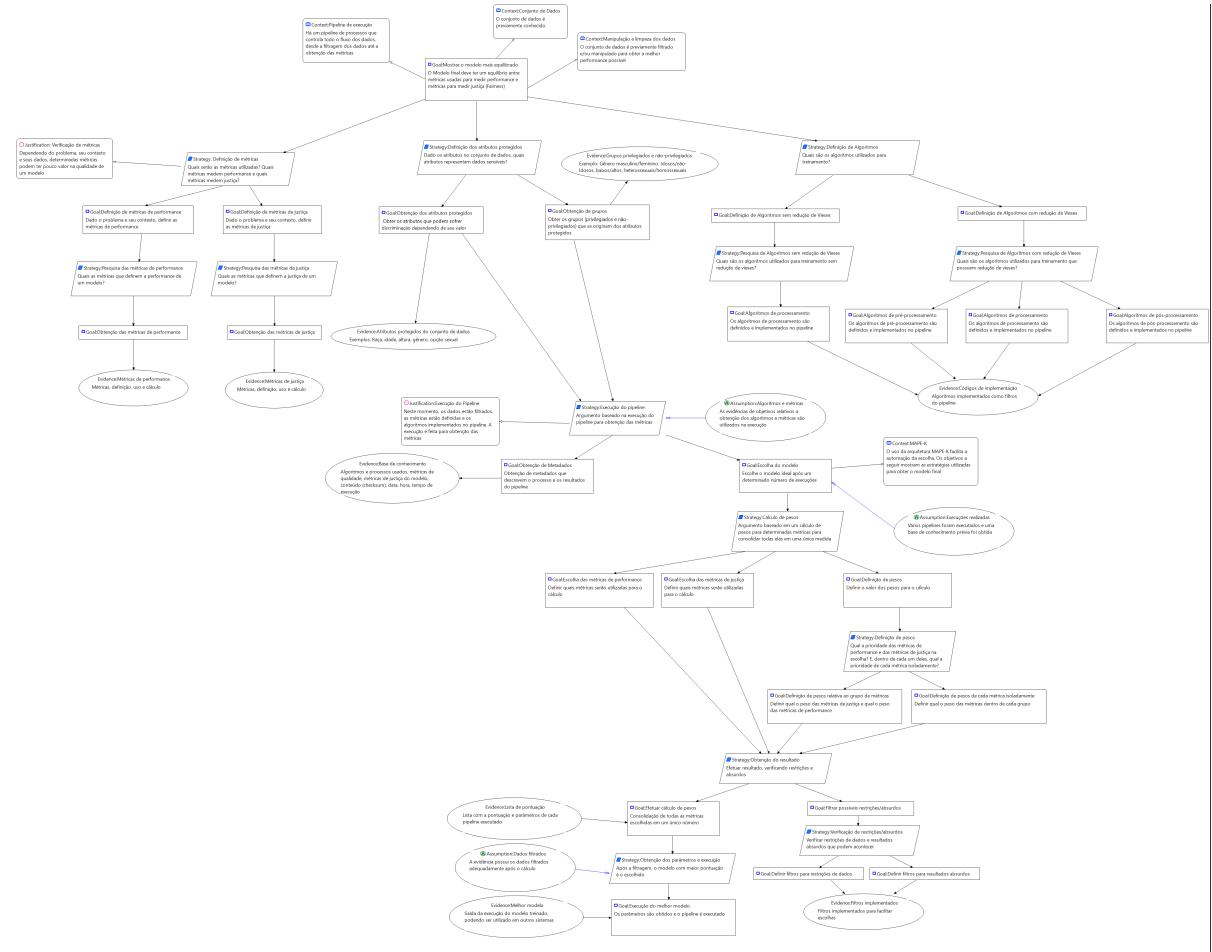


Figura 3.2: *Assurance Cases* feitos para detalhar os objetivos necessários para executar o Pipeline mais adequado

3.2 Arquitetura do código

3.2.1 Transformação dos conjuntos de dados

Nesta etapa, os conjuntos de dados obtidos com atributos qualitativos codificados com uma documentação própria, como o *German Credit Dataset* [9], são transformados com nomenclaturas mais legíveis e colocados em arquivos CSV. O Trecho 3.1, colocado abaixo, ilustra um exemplo dessas operações.

```

1 def german_data_to_csv():
2     pd.set_option('display.max_columns', None)
3     df = pd.read_csv('datasets/german.data', sep=' ')
4     df.info()
5     print(df)
6
7     df['checking_account'] = df['checking_account'].map(
8         {'A11': '<0', 'A12': '0<=x<200', 'A13': '>=200', 'A14': 'None'})
9     .astype(str)
10    df['credit_history'] = df['credit_history'].map(
11        {'A30': 'no_credits_taken', 'A31': 'all_credits_paid_bank',
```

```

11     'A32': 'existing_credits_paid', 'A33': 'delay_in_past', 'A34':
12     'critical'}}).astype(str)
13     df['purpose'] = df['purpose'].map( 'radio/tv',
14         'A44': 'domestic_appliances', 'A45': 'repairs', 'A46': ,
15     education', 'A47':
16         {'A40': 'car_new', 'A41': 'car_used', 'A42': 'furniture/
17     equipment', 'A43': 'vacation', 'A48': 'retraining',
18     'A49': 'business', 'A410': 'others'}).astype(str)
19     df['savings_account'] = df['savings_account'].map(
20         {'A61': '<100', 'A62': '100<=x<500', 'A63': '500<=x<1000', 'A64':
21 : '>=1000', 'A65': 'unknown'}).astype(str)
22     df['present_employment_since'] = df['present_employment_since'].map(
23         {'A71': 'unemployed', 'A72': '<1', 'A73': '1<=x<4', 'A74': '4<=x
24 <7', 'A75': '>=7'}).astype(str)
25     df['personal_status_sex'] = df['personal_status_sex'].map(
26         {'A91': 'male_divorced/separated', 'A92': 'female_divorced/
27 separated/married', 'A93': 'male_single',
28     'A94': 'male_married/widowed', 'A95': 'female_single'}).astype(
29     str)
30     df['other_debtors_guarantors'] = df['other_debtors_guarantors'].map(
31         {'A101': 'None', 'A102': 'co-applicant', 'A103': 'guarantor'}).
32     astype(str)
33     df['property'] = df['property'].map(
34         {'A121': 'real_estate', 'A122': 'savings_insurance', 'A123': ,
35     car_other', 'A124': 'unknown'}).astype(str)
36     df['installment_plans'] = df['installment_plans'].map(
37         {'A141': 'bank', 'A142': 'stores', 'A143': 'None'}).astype(str)
38     df['housing'] = df['housing'].map(
39         {'A151': 'rent', 'A152': 'own', 'A153': 'for_free'}).astype(str)
40     df['job'] = df['job'].map(
41         {'A171': 'unemployed', 'A172': 'unskilled', 'A173': 'skilled', ,
42     A174': 'management'}).astype(str)
43     df['telephone'] = df['telephone'].map(
44         {'A191': 'none', 'A192': 'yes'}).astype(str)
45     df['foreign'] = df['foreign'].map(
46         {'A201': 'yes', 'A202': 'no'}).astype(str)
47     df['risk'] = df['risk'].map(
48         {1: 'good', 2: 'bad'}).astype(str)
49
50     print(df)
51     df.to_csv('datasets/german_credit_data.csv')

```

Código 3.1: Transformações do conjunto de dados *German Credit Dataset*

Nele, cada uma das colunas presentes no conjunto de dados que atribuem uma qualidade é renomeada de acordo com o equivalente presente em sua documentação. Esta etapa simularia uma transformação realizada por um Engenheiro de Dados para facilitar o trabalho de análise e implementação de um Cientista de Dados.

3.2.2 Pipeline

Desenvolvimento de *Framework*

Para facilitar o desenvolvimento do Pipeline e deixar seu código mais legível, foi vista durante seu desenvolvimento a necessidade de desenvolver um pequeno *Framework* baseado na arquitetura *Pipe-and-Filter* onde foram feitas adaptações devido a características e a limitações da linguagem *Python*.

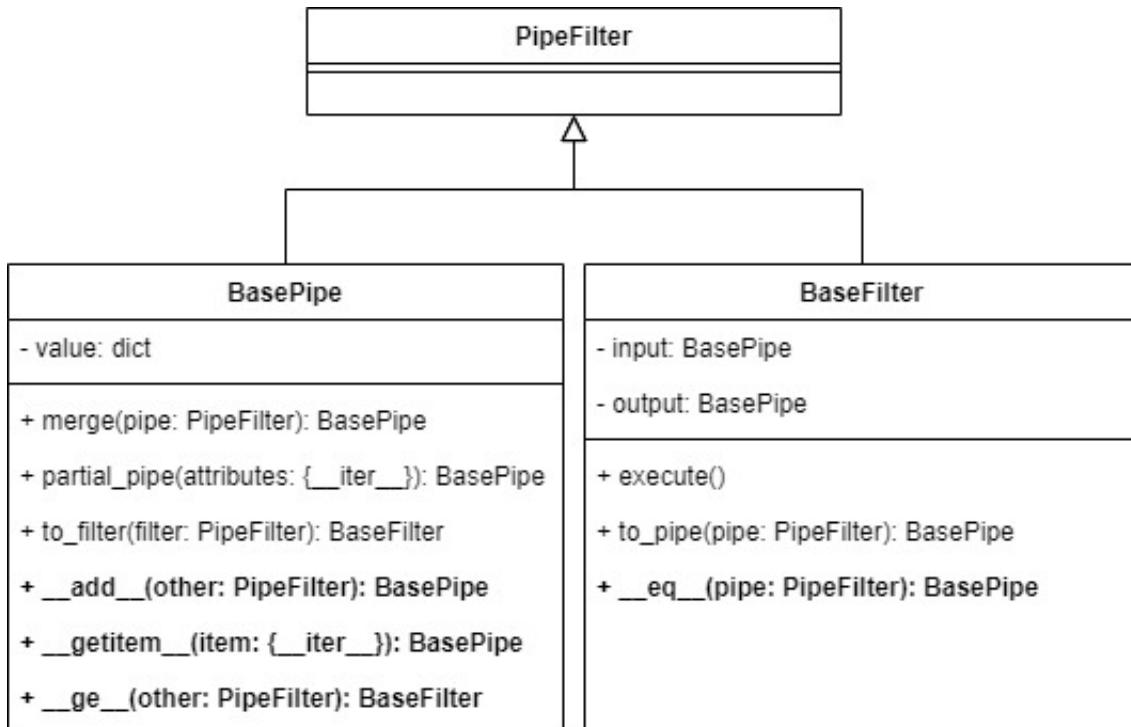


Figura 3.3: Diagrama de classes do *Framework* baseado na arquitetura *Pipe-and-Filter*

Conforme ilustrado na Figura 3.3, o *Framework* possui 3 classes: **PipeFilter**, **BasePipe** e **BaseFilter**. A classe **PipeFilter** foi criada para ser herdada pelas classes **BasePipe** e **BaseFilter** por duas limitações: Inexistência de interfaces no *Python* e evitar erro de recursão de heranças acusado no momento da compilação. A classe **BasePipe** possui apenas um atributo **value** que é um dicionário que simboliza os dados que trafegam por esse Pipe. Como é um dicionário, ele é totalmente livre de quantos atributos o Pipe trafega, sendo assim a sua validação é responsabilidade da aplicação. A classe **BaseFilter** possui dois atributos **input** e **output** que simbolizam os Pipes de entrada e de saída do Filtro, e o uso desses dois Pipes mais a execução do Filtro vai estruturar o Pipeline como um todo.

O Pipeline é realizado através de encadeamentos do método **to_filter** com o método **to_pipe**. O método **to_filter** prepara um Filtro para utilizar o Pipe, setando o atributo **input** presente na classe de Filtro passada como parâmetro, e o método **to_pipe** obtém o Pipe resultante do Filtro, chamando o método **execute** e obtendo o atributo **output** presente no Filtro para setar no atributo **value** presente na classe de Pipe passada como parâmetro. Isso forma encadeamentos de métodos que se assemelham muito com os de

uma *Fluent API* [40], presente no Apêndice A.2, criando uma DSL (*Domain-Specific Language*) cujo domínio é a estruturação de um Pipeline. O Trecho 3.2 mostra um breve exemplo de como é possível encadeá-los, onde **Pipe1**, **Pipe2** e **Pipe3** são classes que herdam **BasePipe** e **Filter1** e **Filter2** são classes que herdam **BaseFilter**.

```

1 root_pipe = Pipe1()
2
3 pipe_with_filter_transformations = root_pipe.to_filter(Filter1())
4                                     .to_pipe(Pipe2())
5                                     .to_filter(Filter2())
6                                     .to_pipe(Pipe3())

```

Código 3.2: Exemplo de uso do *Framework* para encadeamento dos Pipelines

Como a execução dos próprios filtros já é executada dentro do método **to_pipe**, basta apenas realizar a instância das classes e executar os métodos **to_filter** e **to_pipe** para a execução do pipeline. Com isso, os algoritmos de Machine Learning ficam encapsulados em classes que herdam a classe **BaseFilter** e utilizam o método **execute** para executá-los. Desta forma, acontece a separação de interesses entre o componente que gerencia a execução do Pipeline e os componentes que gerenciam cada passo do Pipeline especificamente, garantindo um baixo acoplamento entre os componentes uma vez que é possível trocar de componentes que executam diferentes métodos sem afetar a execução do Pipeline. Consequentemente, conforme novos métodos são descobertos, adicioná-los ao Pipeline é extremamente simples e envolve poucas mudanças no código original. Ferramentas como o Apache Airflow [2] realizam uma abordagem semelhante através da classe **BaseOperator**, que pode ser herdada para implementar um operador customizado e executá-lo através de DAGs (*Direct Acyclic Graphs*).

Utilizando o recurso de métodos especiais presente no Python [7], é possível manipular alguns *tokens* como *proxies* para executar os métodos já presentes no *Framework*, simplificando a DSL e também simplificando o entendimento do Pipeline caso o Cientista de Dados tenha ciência do recurso. Para o método **to_filter** foi escolhido o *token* `>=`, pela semelhança com um funil, que é geralmente assemelhado a filtros em sistemas atuais, e para o *token* o método `__ge__` é implementado como um *proxy* do mesmo. Para o método **to_pipe** foi escolhido o *token* `==`, pela semelhança com um cano (pipe, em inglês), e para este *token* o método `__eq__` é implementado como um *proxy*.

```

1 root_pipe = Pipe1()
2
3 pipe_with_filter_transformations = root_pipe >= Filter1() == Pipe2() >=
4                               Filter2() == Pipe3()

```

Código 3.3: Trecho 3.2 adaptado com métodos especiais da linguagem *Python*

Também foram implementados métodos para manipular os dicionários presentes nos Pipes conforme necessário. O método **merge(pipe)** junta os atributos presentes nos dicionários dos Pipes em um único Pipe, enquanto o método **partial_pipe(attributes)** pega apenas os atributos presentes nos dicionários dos Pipes que forem especificados em uma lista de atributos definida no parâmetro **attributes**, conforme ilustrado no Trecho 3.4.

Como métodos especiais, para o método **merge** foi escolhido o *token* `+`, pela semelhança com uma operação de adição, e para este *token* o método `__add__` é implementado como um *proxy*. E para o método **partial_pipe** foi escolhido os *tokens* `[]`, colocados após o dicionário, pela semelhança com a operação de procurar com um item no dicionário (neste caso, é possível procurar um ou mais itens), e para este *token* o método `__getitem__` é implementado como um *proxy*. O Trecho 3.5 mostra exemplos da modificação.

```

1 pipe1 = Pipe1()
2 pipe1.value = {'attribute1': 1, 'attribute2': 2}
3
4 pipe2 = Pipe2()
5 pipe2.value = {'attribute3': 3, 'attribute4': 4}
6
7
8 pipe3 = pipe1.merge(pipe2)
9 print(pipe3) # Output - {'attribute1': 1, 'attribute2': 2, 'attribute3':
    3, 'attribute4': 4}
10 pipe4 = pipe3.partial_pipe(['attribute1', 'attribute3'])
11 print(pipe4) # Output - {'attribute1': 1, 'attribute3': 3}
```

Código 3.4: Manipulações para Pipes presentes no *Framework*

```

1 pipe1 = Pipe1()
2 pipe1.value = {'attribute1': 1, 'attribute2': 2}
3
4 pipe2 = Pipe2()
5 pipe2.value = {'attribute3': 3, 'attribute4': 4}
6
7
8 pipe3 = pipe1 + pipe2
9 print(pipe3) # Output - {'attribute1': 1, 'attribute2': 2, 'attribute3':
    3, 'attribute4': 4}
10 pipe4 = pipe3['attribute1', 'attribute3']
11 print(pipe4) # Output - {'attribute1': 1, 'attribute3': 3}
```

Código 3.5: Manipulações para Pipes com métodos especiais

Após o desenvolvimento deste *Framework*, foi realizada uma refatoração com a sua utilização e o desenvolvimento do Pipeline foi continuado.

Arquitetura e desenvolvimento do Pipeline

O Pipeline segue uma estrutura das fases de Pré-processamento, processamento e pós-processamento presentes em um processo para desenvolvimento de um modelo de Machine Learning, ilustrados na Figura 3.4. Os pontos de decisão entre redução de viés ou não presentes na figura são transparentes no código devido a implementação do *Framework* já explicado na seção anterior e foram colocados para ilustrar melhor onde atuam os algoritmos de redução dos vieses.

Além de tratar os dados, treinar o modelo e realizar testes sobre o mesmo, o Pipeline possui etapas para obter os metadados durante o processo e salvá-los junto com as métricas para obter a proveniência dos dados necessária para integrá-lo a um elemento autônomo.

Para a execução dos algoritmos com redução de viés, é utilizada a biblioteca Ai Fairness 360, ou AIF360 [1], biblioteca da IBM que compila diversos algoritmos para este fim e facilita o cálculo das métricas de Fairness. Para os algoritmos sem redução de viés, é usado o scikit-learn [8].

Devido a natureza deste trabalho de Mestrado ser voltada à Engenharia de Software e a métricas de Fairness, detalhes do processo que também poderiam ser utilizados como estratégias mais sofisticadas para validação dos modelos, como *k-fold Cross Validation*, e estratégias para regularização foram desconsiderados para evitar complexidade.

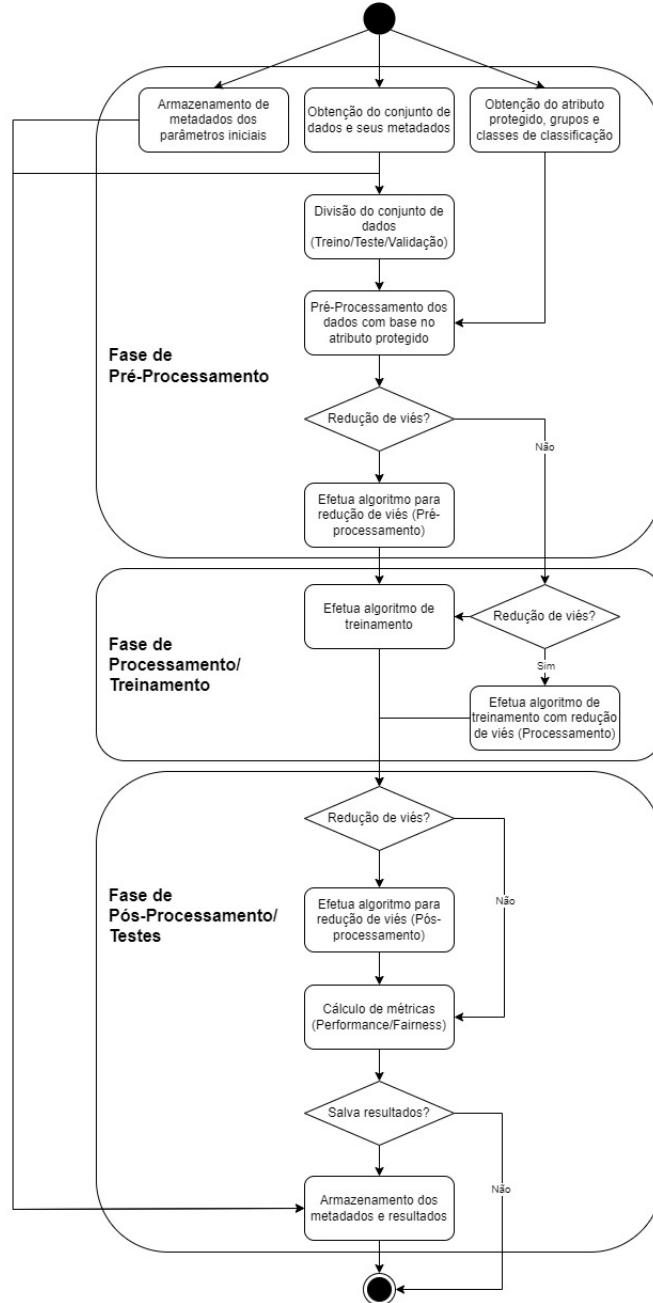


Figura 3.4: Fases e etapas do Pipeline implementado

Cada etapa ilustrada na figura se comporta da seguinte maneira:

- **Fase de Pré-Processamento:**

- **Armazenamento de metadados dos parâmetros iniciais:** Os parâmetros que irão determinar quais serão os conjuntos de dados, atributos protegidos e algoritmos utilizados em cada fase do Pipeline são armazenados em um Pipe e guardados para uma gravação em arquivo caso necessário.
- **Obtenção do conjunto de dados e seus metadados:** De acordo com o parâmetro passado para o Pipeline, um Pipe relativo ao conjunto de dados em questão é selecionado, e os metadados presentes neste Pipe podem ser utilizados para uma gravação em arquivo caso necessário.
- **Obtenção do atributo protegido, grupos e classes de classificação:** De acordo com o parâmetro passado para o Pipeline, um Pipe relativo ao atributo protegido em questão é selecionado, e informações de grupos privilegiados e não-privilegiados e classes para classificação, necessárias para executar os algoritmos presentes no AIF360, também estão presentes neste Pipe.
- **Divisão do conjunto de dados (Treino/Validação/Testes):** Um Filtro relativo a divisão do conjunto de dados (entre conjunto de treino, conjunto de validação e conjunto de testes) é executado e suas divisões resultantes são colocadas em um Pipe.
- **Pré-Processamento dos dados com base no atributo protegido:** As informações de atributo protegido são passadas para um Filtro para realizar um pré-processamento no conjunto de dados, preparando o conjunto de dados para o AIF360.
- **Execução(ou não) de algoritmo para redução de viés:** De acordo com o parâmetro passado para o Pipeline, um algoritmo para redução de viés na fase de Pré-Processamento é colocado ou não para ser executado no Pipeline.

- **Fase de Processamento/Treinamento:**

- **Execução de algoritmo de treinamento com(ou sem) redução de viés:** De acordo com o parâmetro passado para o Pipeline, um algoritmo de treinamento, podendo ter ou não ter redução de viés, é colocado ou não para ser executado no Pipeline.

- **Fase de Pós-Processamento/Testes:**

- **Execução(ou não) de algoritmo para redução de viés:** De acordo com o parâmetro passado para o Pipeline, um algoritmo para redução de viés na fase de Pós-Processamento é colocado ou não para ser executado no Pipeline.
- **Cálculo de métricas (Performance/Fairness):** É realizado uma predição do algoritmo treinado com o conjunto de teste, e as previsões são comparadas com os valores verdadeiros para obtenção das métricas.
- **Armazenamento dos metadados e resultados:** Caso seja habilitado a gravação dos resultados, os metadados obtidos em etapas anteriores e o resultado das métricas são combinados e gravados em um arquivo JSON.

3.2.3 Componente MAPE-K

Os arquivos JSON gravados na última etapa a cada execução de Pipeline vão ser utilizados como insumos para análises elemento autônomo, seguindo o modelo de arquitetura MAPE-K. Em um Pipeline de Machine Learning, é possível guardar alguns metadados que podem servir de Base de Conhecimento inicial para a etapa de análise do componente a ser desenvolvido e que ajudariam na escolha do melhor modelo possível para execução:

- **Fase de Pré-Processamento:** Parâmetros utilizados para execução do Pipeline (Conjuntos de dados, Atributos protegidos e Algoritmos utilizados) e "assinatura"/checksum do conjunto de dados utilizado.
- **Fase de Processamento:** Parâmetros utilizados para execução dos algoritmos de treinamento.
- **Fase de Pós-Processamento:** Métricas do modelo resultante (Metrics de Performance e métricas de Fairness).

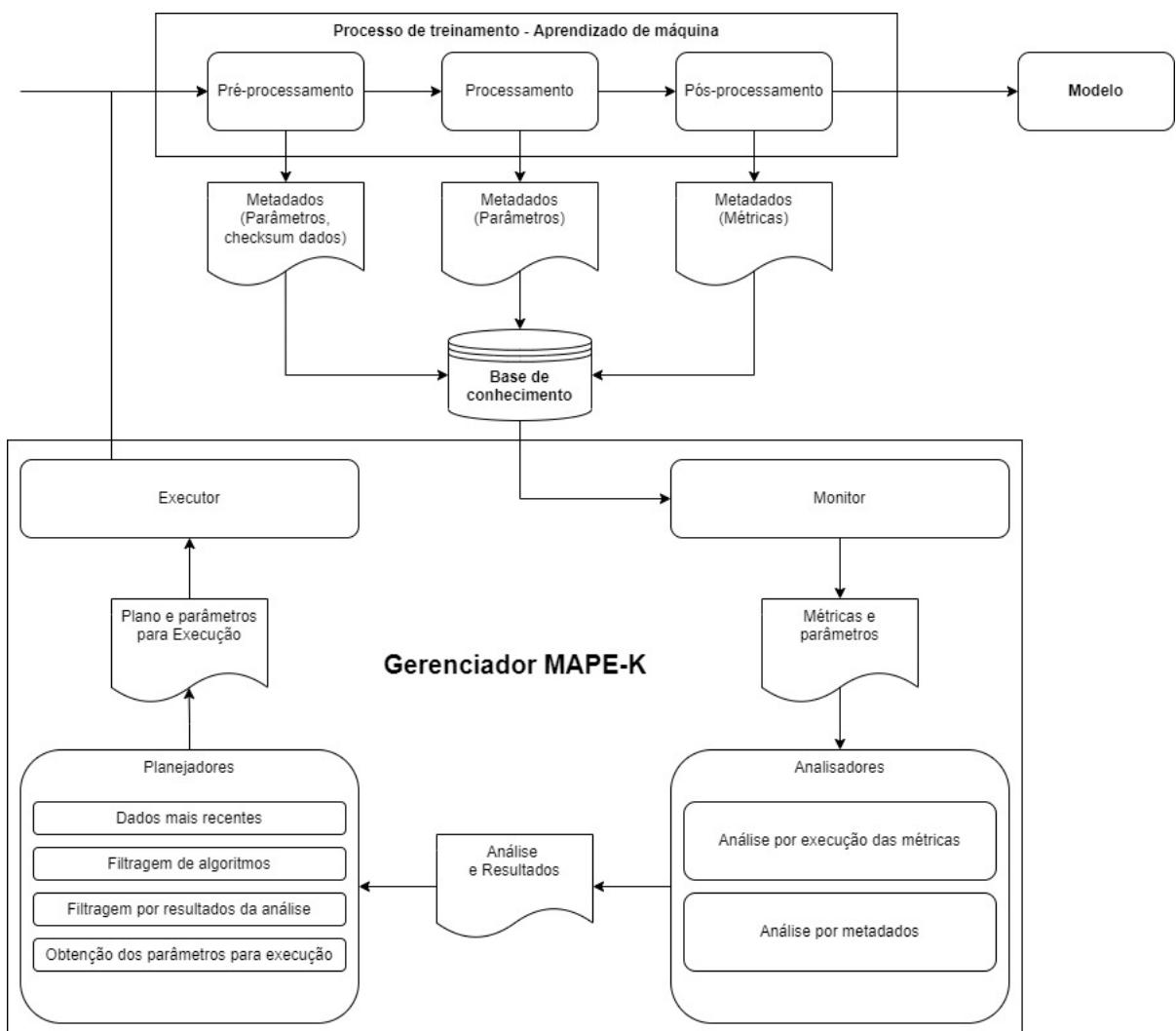


Figura 3.5: Adequação do Pipeline ao gerenciador MAPE-K

Após a obtenção de tal Base de Conhecimento, é possível verificar como as etapas de cada parte do MAPE-K serão desenvolvidas. Conforme ilustrado na Figura 3.5, as etapas de Análise e Planejamento foram subdivididas em diferentes estratégias para o MAPE-K ser configurável durante o processo de escolha, enquanto as etapas de monitoria e execução foram configuradas com uma única estratégia. As estratégias das etapas de Análise e Planejamento podem ser ativadas ou desativadas conforme o que é definido como *Feature Toggles* [65], presente no Apêndice A.3, por meio de arquivos JSON presentes no projeto conforme ilustrado no Trecho 3.6.

```

1 {
2     "ml_algorithm_validation": true,
3     "ml_data_checksum": false,
4     "ml_pipeline": false,
5     "ml_pipeline_threshold": true
6 }
```

Código 3.6: Exemplo de arquivo utilizando ([ver termo acima](#))

Monitoria

Para a etapa de monitoria, os dados são obtidos em cada arquivo presente na base de conhecimento e separados em dois conjuntos diferentes: Um contendo os metadados e informações relativas a execução do Pipeline (checksum do conjunto de dados, estatísticas e parâmetros de execução do Pipeline) e outro contendo as métricas obtidas pelo modelo resultante dessa execução, ambos possuindo um identificador do arquivo para estabelecer consistência entre os dados dos dois conjuntos. Os arquivos podem ser filtrados pelo conjunto de dados e pelo atributo protegido, ou não possuir filtro obtendo todos os dados possíveis.

Análise

Para a etapa de análise, é realizado um cálculo de pesos baseado em uma seleção livre das métricas. O motivo de existir esse cálculo é mensurar o contexto do problema de acordo com uma análise prévia do Cientista de Dados e do Especialista de Domínio, e consolidar todas as métricas para simplificar as estratégias de planejamento. É possível encontrar abordagens similares e com objetivos similares, porém em diferentes contextos, pesquisando em artigos acadêmicos [?].

```

1 {
2     "metrics_groups": [
3         {
4             "group_name": "standard",
5             "weight": 0.5,
6             "metrics": {
7                 "accuracy": {
8                     "weight": 0.5,
9                     "normalize": null
10                },
11                 "f1_score": {
12                     "weight": 0.5,
```

```

13     "normalize": null
14   }
15 }
16 ],
17 {
18   "group_name": "fairness",
19   "weight": 0.5,
20   "metrics": {
21     "statistical_parity_difference": {
22       "weight": 0.33,
23       "normalize": "diff"
24     },
25     "equal_opportunity_difference": {
26       "weight": 0.33,
27       "normalize": "diff"
28     },
29     "average_abs_odds_difference": {
30       "weight": 0.34,
31       "normalize": "diff"
32     }
33   }
34 }
35 ]
36 }
```

Código 3.7: Exemplo de arquivo contendo informações sobre as métricas, grupos e seus respectivos pesos

As métricas são divididas em dois grupos (Métricas de performance e Metrics de Fairness), e dentro desse grupo pode-se colocar quantas métricas forem necessárias, desde que seja respeitado o contexto de cada grupo. A cada grupo é atribuído pesos diferentes, e a cada métrica desse grupo também é atribuído pesos diferentes, conforme ilustrado no Trecho 3.7. Primeiro, normaliza-se as métricas m_{F_i} para m'_{F_i} , referentes às métricas de Fairness, para todas ficarem em um intervalo de 0 a 1, conforme exibido na equação 3.1. Dessa forma, seus resultados ficam uniformes e é possível aplicar os pesos sem haver distorções no cálculo. No caso das métricas de Performance, todas possuem a mesma escala, por isso as métricas m_{P_i} não são normalizadas. Depois, multiplica-se cada uma por seus pesos correspondentes w_{P_i} e w_{F_i} , e realiza-se uma média ponderada dentro do grupo para atribuir uma pontuação S_P para o grupo das Métricas de Performance e S_F para o grupo das Metrics de Fairness, conforme exibido na equação 3.2. Para facilitar a visualização das pontuações, multiplica-se as pontuações por um fator $X = 1000$ para o intervalo da pontuação ser de 0 a 1000 e arredonda-se o número. Após tais pontuações serem obtidas, a pontuação geral S é calculada multiplicando-as por seus pesos correspondentes w_P e w_F e realizando a média ponderada, conforme exibido na equação 3.3.

$$m'_{F_i} = \begin{cases} 1 - |m_{F_i}| & \text{caso } m_{F_i} \text{ envolva diferença e } -1 < m_{F_i} < 1 \\ 0 & \text{caso } m_{F_i} \text{ envolva diferença, e } m_{F_i} \geq 1 \text{ ou } m_{F_i} \leq -1 \\ 1 - |\frac{1}{m_{F_i}} - 1| & \text{caso } m_{F_i} \text{ envolva razão e } m_{F_i} > 1 \\ 1 - |m_{F_i} - 1| & \text{caso } m_{F_i} \text{ envolva razão e } m_{F_i} \leq 1 \\ m_{F_i} & \text{caso contrário} \end{cases} \quad (3.1)$$

$$S_F = \left[X \times \frac{\sum_{i=1}^{n_F} w_{m'_{F_i}} \times m'_{F_i}}{\sum_{i=1}^n w_{m'_{F_i}}} \right] \quad (3.2)$$

$$S_P = \left[X \times \frac{\sum_{i=1}^{n_P} w_{m_{P_i}} \times m_{P_i}}{\sum_{i=1}^n w_{m_{P_i}}} \right]$$

$$S = \frac{w_F \times S_F + w_P \times S_P}{w_F + w_P} \quad (3.3)$$

Nesta etapa, foram desenvolvidas as seguintes estratégias, cujos motivos para existir e funcionamento estão explicados abaixo:

- **Análise por execução das métricas:** É atribuída a cada execução presente no conjunto uma pontuação conforme o cálculo explicado acima, e as pontuações são passadas adiante para a fase de planejamento.
- **Análise por metadados:** Para o desenvolvimento de melhores estratégias de planejamento, alguns metadados, como data de execução, são analisados e o conjunto resultante é enriquecido com esses metadados mais específicos.

Planejamento

Para a etapa de planejamento, foram desenvolvidas as seguintes estratégias, cujos motivos para existir e funcionamento estão explicados abaixo, para a escolha dos melhores modelos:

- **Dados mais recentes:** Os dados podem ser modificados de acordo com a execução e a qualidade das métricas são melhor definidas de acordo com a qualidade dos dados. Nesta estratégia, é realizado um filtro baseado na assinatura do dado com a execução mais recente, que só é possível de ser obtida se a análise dos metadados for executada.
- **Filtragem de algoritmos:** Alguns algoritmos podem estar mal implementados ou seus modelos podem estar com métricas que necessitam uma melhor análise do Cientista de Dados para serem consideradas confiáveis. Para isso não acontecer, é possível realizar um filtro de acordo com as combinações de algoritmos consideradas confiáveis antes de selecionar os modelos ideais.
- **Fase por resultados da análise:** Pelo mesmo motivo da estratégia anterior, métricas não confiáveis significam uma distorção na pontuação final. Para esse caso, foi criado um limiar de pontuação mínimo e máximo para determinar pontuações que podem ser consideradas confiáveis para avaliação.

- **Obtenção dos parâmetros para execução:** Alguns dos modelos restantes são selecionados de acordo com as maiores pontuações e os parâmetros presentes nestes modelos selecionados são obtidos.

Após a execução das mesmas, caso estejam ativadas, os parâmetros selecionados são passados para a etapa de execução.

Execução

Para a etapa de execução, os parâmetros selecionados na fase de planejamento são configurados como parâmetros para a execução do Pipeline, e esta execução pode ampliar a Base de Conhecimento ou não dependendo do valor da opção de gravação dos resultados.

3.2.4 Interface Humano-Computador

Durante o desenvolvimento do Pipeline e de seu componente autônomo, foi notado que o número de configurações e a complexidade das mesmas era muito grande, ocasionando problemas na hora de documentar e detalhar todo o processo executado. Para facilitar tais configurações, foi criada uma Interface Humano-Computador onde é condensada toda a organização das configurações e dos arquivos utilizados para a execução simples e autônoma do Pipeline, além da própria realização destas execuções. A interface foi dividida em uma parte Frontend e outra parte Backend para ter mais flexibilidade, poder ter uma escolha de ferramentas mais adequada a cada parte e próxima ao padrão de aplicações atuais.

O Backend foi desenvolvido em Python para reusar códigos já desenvolvidos em etapas anteriores, usando o framework Flask para construir as requisições web e foi dividido em 3 camadas. A camada *web* corresponde às requisições que constroem a ponte entre Frontend e Backend, a camada *service* corresponde às funcionalidades e casos de uso que serão chamados pelas requisições, e a camada *repo* corresponde às operações de leitura e escrita que serão realizadas nos arquivos do Pipeline.

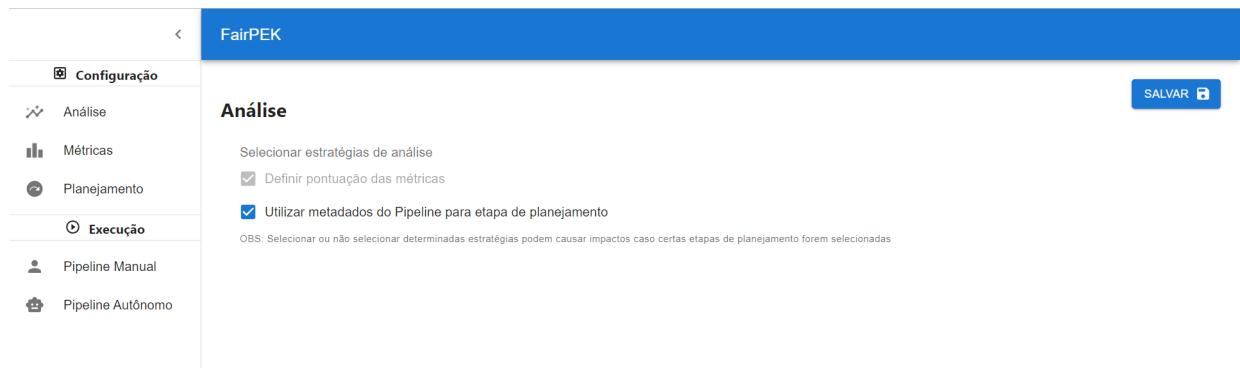
O Frontend foi desenvolvido em JavaScript devido à facilidade e à robustez para construir interfaces com a linguagem e ao acervo grande de ferramentas para facilitar o desenvolvimento, usando as bibliotecas React e Redux e a biblioteca de componentes Material UI para economizar tempo com componentes já prontos, uma vez que o foco deste trabalho não exige componentes específicos para a interface. O uso da Material UI permite um *look-and-feel* similar a aplicativos desenvolvidos para o Sistema Operacional para dispositivos móveis Android, uma vez que é baseado nas *guidelines* do Material Design elaborados pelo Google.

A interface foi nomeada de FairPEK, junção dos termos Fairness e MAPE-K, e possui os seguintes detalhes:

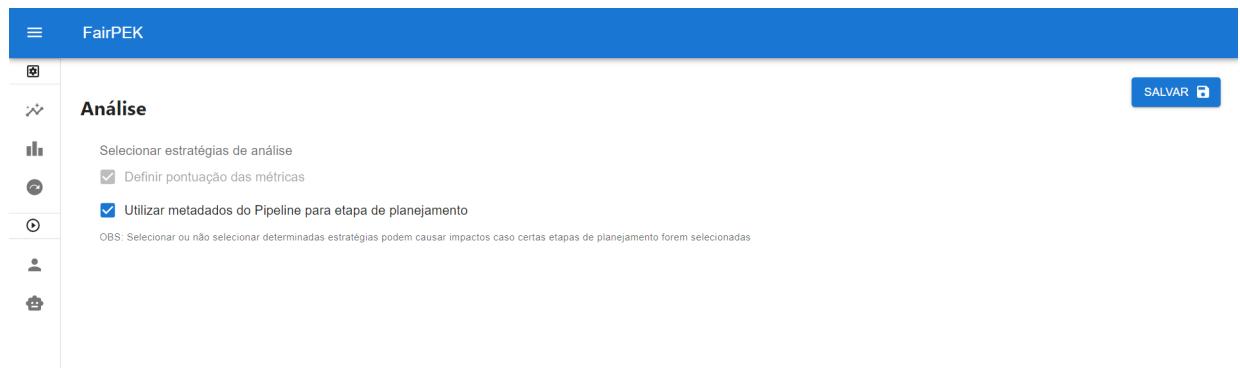
Opções do Menu

Conforme ilustrado na Figura 3.6, o menu pode ser expandido e recolhido, onde suas opções são selecionáveis independente da configuração, e foram colocados ícones ao lado

do nome de sua opção para que a localização das opções seja acessível mesmo com o menu recolhido.



(a) Opções de seleção de menu expandidas.



(b) Opções de seleção de menu recolhidas.

Figura 3.6: Comportamento das opções de menu.

No menu, as opções são divididas em Configuração e Execução. Em Configuração, há as opções Análise, Métricas e Planejamento relativas às configurações presentes no Componente MAPE-K. Em Execução, há as opções Pipeline Manual e Pipeline Autônomo relativas às maneiras de como realizar uma execução do Pipeline.

Configurações para Análise

Ao clicar a opção do menu "Análise", é exibida a tela ilustrada na Figura 3.7. Ela possui apenas duas opções, relativas às estratégias desenvolvidas na parte de análise do componente MAPE-K. Ao clicar no botão "Salvar" localizado no canto superior direito, é chamada uma requisição que salva o arquivo com as opções selecionadas e é exibida uma indicação de sucesso no canto inferior esquerdo.

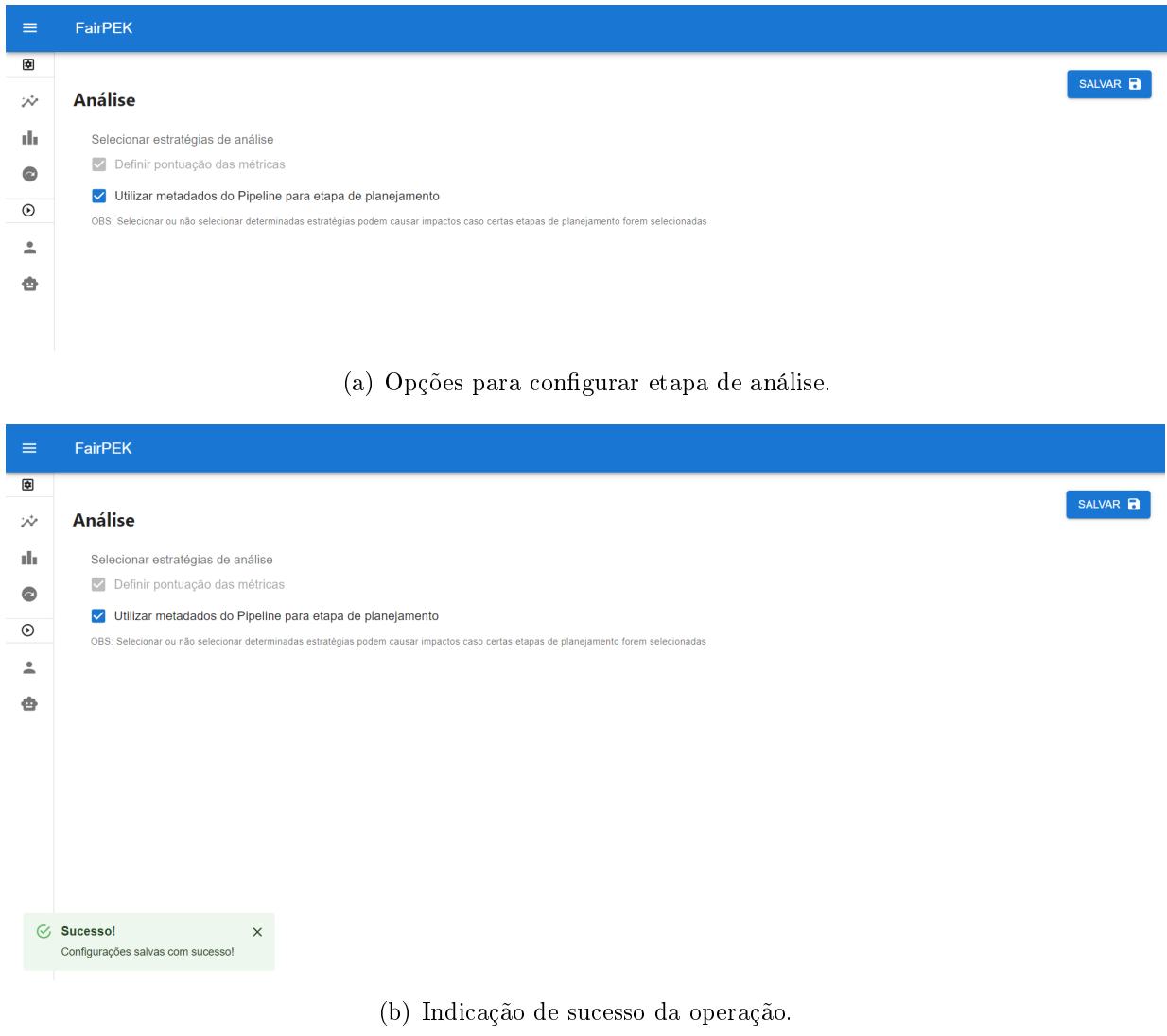
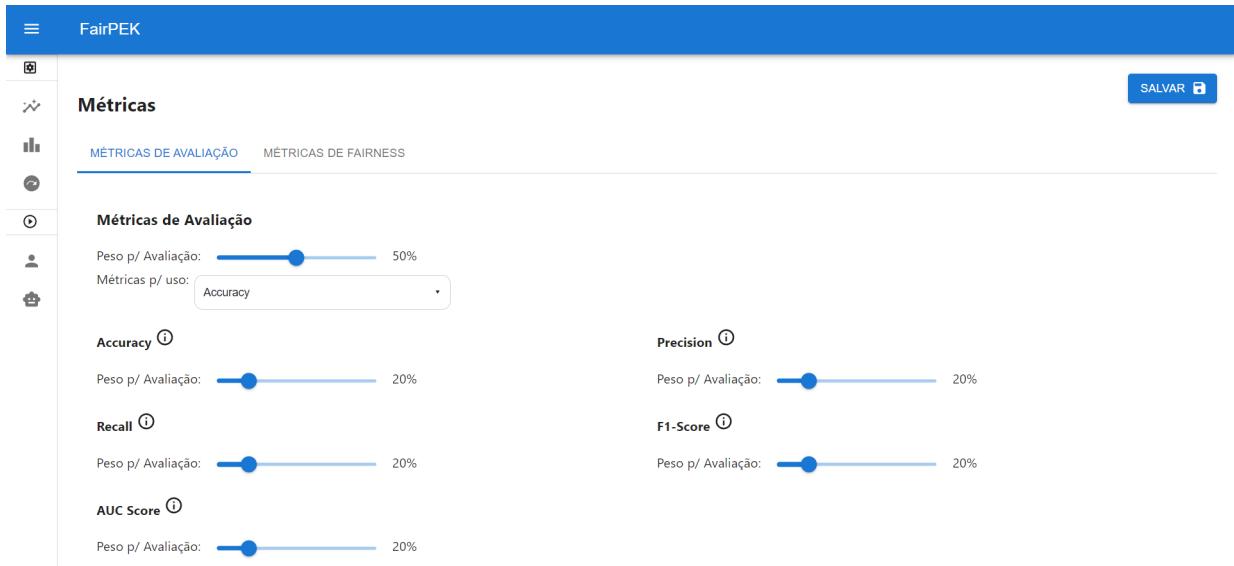


Figura 3.7: Configuração da etapa de análise para o pipeline autônomo.

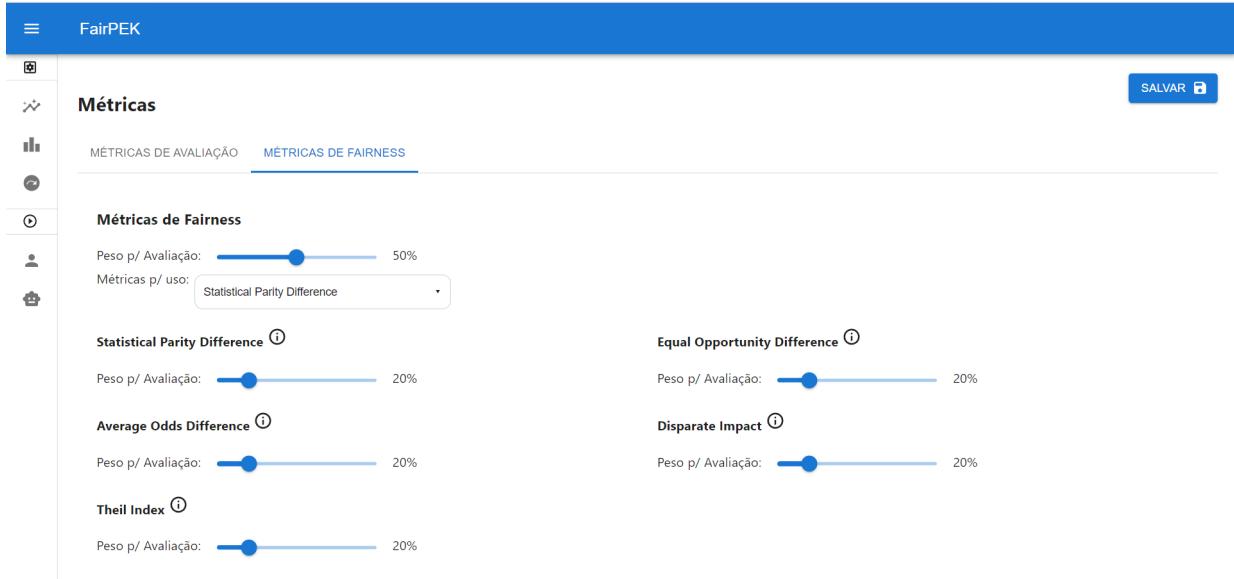
Uma das duas estratégias está desabilitada pois foi desenvolvida como a estratégia padrão usada pelo componente MAPE-K. Se a estratégia de utilizar metadados não for habilitada, certas opções marcadas na etapa de planejamento podem não funcionar.

Configurações das Métricas

Ao clicar a opção do menu "Métricas", é exibida a tela ilustrada na Figura 3.8. Ela é dividida em duas abas: Métricas de Avaliação e Métricas de Fairness, relativas aos grupos de métricas divididos no componente MAPE-K. Em ambas as abas, há os campos de peso para avaliação, que simboliza o peso no cálculo da pontuação final, e o campo de métricas para uso, que determina quais métricas serão utilizadas para o cálculo da pontuação de cada grupo. Uma vez que há apenas duas abas, a alteração do campo de peso para avaliação de uma aba automaticamente alterará o campo de peso para avaliação da outra aba para complementar a soma de 100% sem precisar de validações.



(a) Configuração para Métricas de Performance.



(b) Configuração para Métricas de Fairness.

Figura 3.8: Configuração das métricas para etapa de análise do pipeline autônomo.

A alteração do campo de métricas para uso implicará na presença de mais ou menos métricas para alterar os pesos para o cálculo da pontuação para o grupo de métricas. Uma vez que eu posso ter mais de duas métricas nesse caso, implementar a soma de todas as métricas automaticamente para 100% seria uma tarefa com maior complexidade, e por isso foi substituída por uma validação, conforme ilustrado na Figura 3.9. Ao clicar no botão "Salvar" localizado no canto superior direito, é realizada a validação da soma de todas as métricas e, em caso positivo, chamada uma requisição que salva o arquivo com as opções selecionadas e seus respectivos valores. Após este procedimento, será exibida uma indicação de sucesso ou erro de validação no canto inferior esquerdo indicando se o arquivo foi salvo ou não.

The figure consists of two screenshots of the FairPEK application interface, labeled (a) and (b), illustrating different outcomes of metric configuration.

Screenshot (a): Erro de validação

This screenshot shows the 'Métricas' (Metrics) section with the 'MÉTRICAS DE AVALIAÇÃO' tab selected. Under 'Métricas de Avaliação', there are four sliders for 'Accuracy', 'Precision', 'Recall', and 'F1-Score', each set to 20%. Above these, a global slider for 'Peso p/ Avaliação' is set to 50%, and a dropdown 'Métricas p/ uso' is set to 'Accuracy'. A red warning box at the bottom left states: 'Erro de validação' (Validation error) and 'A soma das pesos para as métricas de Fairness e de Performance estão maiores do que 100%' (The sum of weights for Fairness and Performance metrics are greater than 100%).

Screenshot (b): Sucesso!

This screenshot shows the same 'Métricas' section. The configuration is identical to (a), but a green success message box at the bottom left says: 'Sucesso!' (Success!) and 'Configurações salvas com sucesso!' (Configurations saved successfully!).

(a) Erros de validação ao concluir a operação.
(b) Indicação de sucesso da operação.

Figura 3.9: Cenários possíveis na configuração das métricas.

Configurações para Planejamento

Ao clicar a opção do menu "Planejamento", é exibida a tela ilustrada na Figura 3.10. Ela possui quatro opções, relativas às estratégias desenvolvidas na parte de análise do componente MAPE-K. Ao clicar no botão "Salvar" localizado no canto superior direito, são chamadas requisições que salvam os arquivos necessários e é exibida uma indicação de sucesso no canto inferior esquerdo.

Planejamento

Selecionar estratégias de planejamento

Escolher o modelo com a melhor pontuação

Escolher o modelo de acordo com os dados mais recentes encontrados em execuções prévias

Restringir conjunto de algoritmos escolhidos

Sem método
Logistic Regression
Sem método

Sem método
Logistic Regression
Equalized Odds

Sem método
Logistic Regression
Calibrated Equalized Odds

Restringir limiar de pontuação

500 ————— 881

OBS. Selecionar ou não selecionar determinadas estratégias podem causar impactos caso certas etapas de análise não forem selecionadas

Sucesso!
Configurações salvas com sucesso!

(a) Opções para configurar etapa de planejamento.

(b) Indicação de sucesso da operação.

Figura 3.10: Configuração da etapa de planejamento para o pipeline autônomo.

Uma das quatro estratégias está desabilitada pois foi desenvolvida como a estratégia padrão usada pelo componente MAPE-K. Se a estratégia de restrição de algoritmos for selecionada, é exibido um submenu contendo as combinações de algoritmos que podem ser selecionadas que são salvadas em um arquivo separado para serem filtrados na etapa de planejamento. Se a estratégia de restrição por um limiar de pontuação for selecionada, é exibido um slider com uma pontuação mínima e uma pontuação máxima, e ambas as pontuações são salvadas em um arquivo separado para serem filtrados na etapa de planejamento.

Execução simples do Pipeline

Ao clicar a opção do menu "Pipeline Manual", é exibida a tela ilustrada na Figura 3.11. Ela possui indicações de etapas divididas em Parametrização, Execução e Resultados, campos para selecionar o conjunto de dados e o atributo protegido e opções para selecionar onde a redução de viés será executada. Dependendo da opção selecionada, aparecem campos para selecionar o algoritmo de treinamento e o algoritmo de redução de viés. Ao clicar no botão "Executar" localizado no canto superior direito, é feita uma requisição para executar o Pipeline com as opções selecionadas, é exibida uma indicação de sucesso no canto inferior esquerdo e a indicação de etapa é atualizada para a etapa de execução.

The screenshot shows the FairPEK Pipeline Manual interface. At the top, there's a blue header bar with the FairPEK logo and a sidebar with various icons. The main area is titled "Pipeline Manual" and has three numbered steps: 1. Parametrização, 2. Execução, and 3. Resultado. Step 1 is active. Under "Parametrização", there are dropdown menus for "Conjunto de dados" (set to "German Credit Dataset") and "Atributo Protegido" (set to "Idade (-25 anos/+25 anos)"). Below these, a section titled "Redução de viés será aplicada em qual etapa do Pipeline?" has four radio button options: "Pré-Processamento/Dados" (selected), "Processamento/Treinamento", "Pós-Processamento/Avaliação", and "Nenhum (Executar treinamento convencional)". Further down are dropdown menus for "Algoritmo de treinamento" (set to "Logistic Regression") and "Algoritmo de redução de viés" (set to "Reweighting"). A note below the "Redução de viés" section states: "Algoritmo de treinamento sem redução de viés". A large blue "EXECUTAR" button is located at the top right of the main area.

(a) Opções para executar o pipeline manualmente.

This screenshot shows the same interface as above, but step 1 ("Parametrização") now has a checked checkbox icon next to it, indicating it's completed. Step 2 ("Execução") is active and shows the status "Executando Pipeline...". In the bottom left corner, there's a green notification box with a checkmark and the word "Sucesso!", with the subtext "Execução será realizada em alguns segundos".

(b) Pipeline em execução.

Figura 3.11: Execução simples e manual do Pipeline.

Após a execução ser concluída, a indicação de etapa é atualizada para a etapa de resultados, conforme ilustração nas Figuras 3.12 e 3.13. Nela, os parâmetros gravados são organizados em 4 grupos: Execução, relativos aos parâmetros utilizados e estatísticas da execução, Métricas de Performance, relativas aos resultados das métricas de Performance,

Métricas de Fairness, relativas aos resultados das métricas de Fairness, e Pontuação, relativas ao cálculo realizado com as configurações utilizadas na parte de métricas.

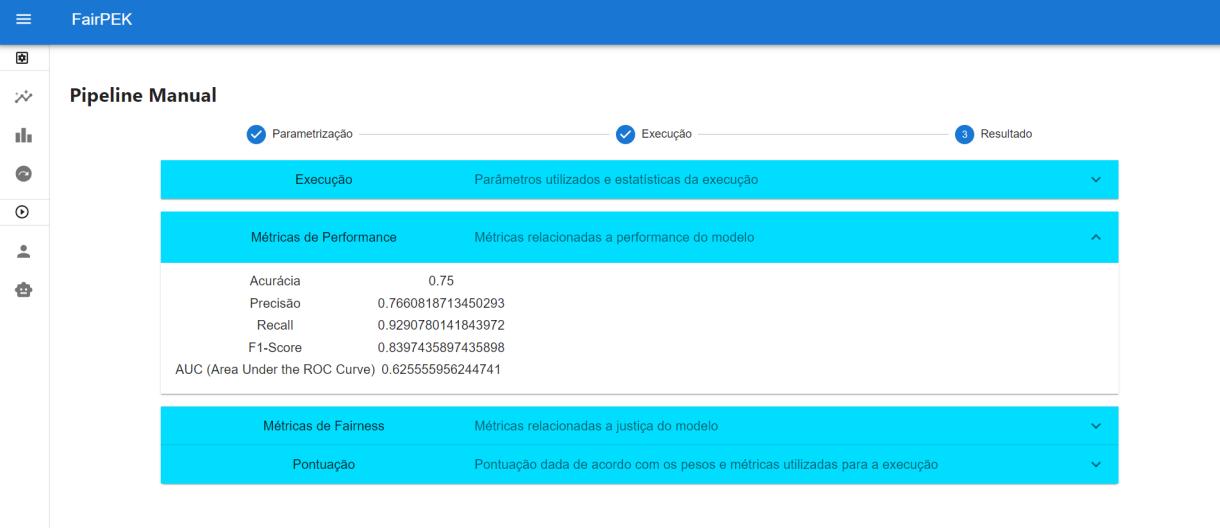
	Parâmetros utilizados e estatísticas da execução
Checksum do conjunto de dados	fcde189dd32a1631a474e1087a7f4835525194741309493e0475bce7cb3451d0a499c70007db6c4bf5e33cb4ecb2da9feb859b57cfa21255573103d92e6b497d
Conjunto de dados utilizado	Datasets.GERMAN_CREDIT
Atributo protegido	Preprocessors.AGE
Algoritmo de redução de viés no dado	UnbiasDataAlgorithms.NOTHING
Algoritmo de treinamento (redução de viés: Sim)	UnbiasInProcAlgorithms.META_FAIR_CLASSIFIER
Algoritmo de redução de viés no pós-processamento	UnbiasPostProcAlgorithms.NOTHING
Data de inicio da execução	02/05/2022 00:42:02.878556
Data de fim da execução	02/05/2022 00:42:04.598311
Tempo de Execução	1719 ms

(a) Exibição dos parâmetros no resultado do pipeline.

	Pontuação dada de acordo com os pesos e métricas utilizadas para a execução
Pontuação das métricas de performance	782
Pontuação das métricas de fairness	963
Pontuação geral	872

(b) Exibição das pontuações no resultado do pipeline.

Figura 3.12: Informações do resultado do pipeline.

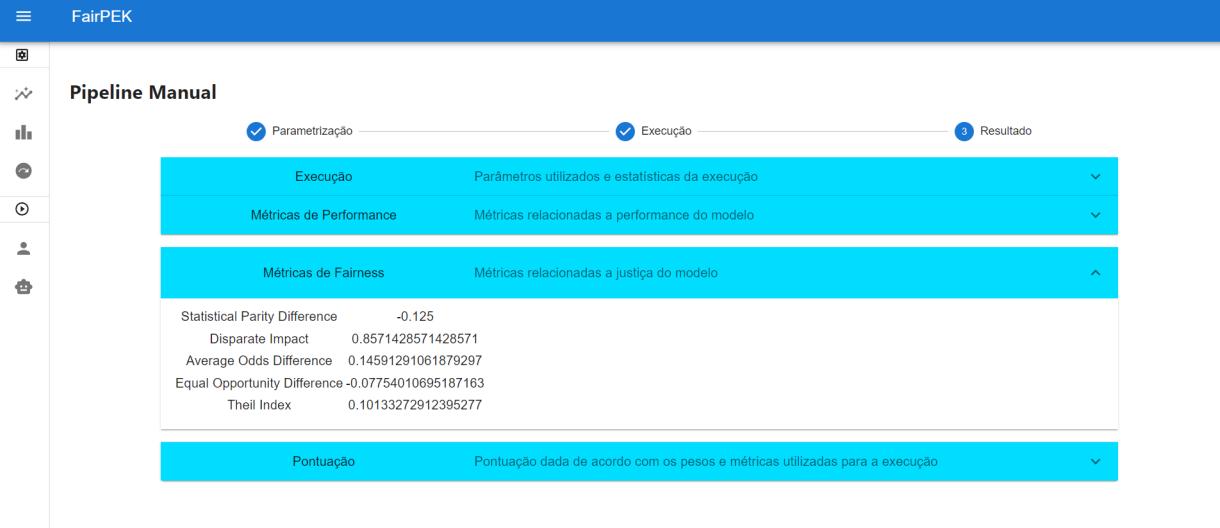


The screenshot shows the FairPEK Pipeline Manual interface. The top navigation bar includes icons for file, search, refresh, and user. The main title is "Pipeline Manual". Below it, a horizontal bar indicates the current step: "Parametrização" (checked), "Execução" (checked), and "Resultado" (with a value of 3). The "Resultado" section is expanded, showing two tabs: "Execução" (selected) and "Parâmetros utilizados e estatísticas da execução". Under "Execução", the "Métricas de Performance" tab is selected, displaying the following data:

Acurácia	0.75
Precisão	0.7660818713450293
Recall	0.9290780141843972
F1-Score	0.8397435897435898
AUC (Area Under the ROC Curve)	0.625555956244741

Below this, the "Métricas de Fairness" tab is shown, with the "Pontuação" sub-tab selected, indicating that the score is based on weights and metrics used for execution.

(a) Exibição das métricas de performance no resultado do pipeline.



This screenshot shows the same interface as above, but the "Métricas de Fairness" tab is selected under the "Execução" section. It displays the following data:

Statistical Parity Difference	-0.125
Disparate Impact	0.8571428571428571
Average Odds Difference	0.14591291061879297
Equal Opportunity Difference	-0.07754010695187163
Theil Index	0.10133272912395277

Below this, the "Pontuação" sub-tab is selected, indicating that the score is based on weights and metrics used for execution.

(b) Exibição das métricas de fairness no resultado do pipeline.

Figura 3.13: Métricas do resultado do pipeline.

Execução autônoma do Pipeline

Ao clicar a opção do menu "Pipeline Autônomo", é exibida a tela ilustrada na Figura 3.14. Ela possui indicações de etapas divididas em Parametrização, Análise, Opções, Execução e Resultados e um campo para selecionar o conjunto de dados. Ao clicar no botão "Executar" localizado no canto superior direito, é feita uma requisição para escolher o melhor conjunto de parâmetros baseado em execuções anteriores, é exibida uma indicação de sucesso no canto inferior esquerdo e a indicação de etapa é atualizada para a etapa de análise.

(a) Opções para configurar o pipeline de forma autônoma.

(b) Análise do componente MAPE-K em execução.

Figura 3.14: Execução autônoma do Pipeline.

Após a etapa de análise ser concluída, a indicação de etapa é atualizada para a etapa de opções, conforme ilustração na Figura 3.15. Nela, os parâmetros sugeridos são organizados nos mesmos grupos presentes nas Figuras 3.12 e 3.13 e podem ser consultados para selecionar a melhor escolha possível das 5 melhores sugestões de acordo com a pontuação calculada, podendo contestar ou não a melhor escolha sugerida pelo componente MAPE-K.

(a) Opções para seleção de pipeline para execução.

(b) Parâmetros a serem utilizados para execução.

Figura 3.15: Seleção do pipeline após análise.

Ao clicar novamente no botão "Executar" localizado no canto superior direito, é feita uma requisição para executar o Pipeline com a opção selecionada, e a indicação de etapa é atualizada para a etapa de execução. Após a execução ser concluída, a indicação de etapa é atualizada para a etapa de resultados, conforme ilustração na Figura 3.16. Nela, os parâmetros sugeridos são organizados nos mesmos grupos presentes na Figura 3.15, mas desta vez refletem as métricas e pontuação da execução realizada pelo Pipeline.

,

The figure consists of two screenshots of the FairPEK software interface.

(a) Pipeline em execução: This screenshot shows the pipeline execution process. The top navigation bar is blue with the text "FairPEK". Below it, a sidebar on the left contains icons for search, refresh, user, and help. The main area is titled "Pipeline Autônomo" and shows a horizontal progress bar with five steps: "Parametrização" (checked), "Análise" (checked), "Opções" (checked), "Execução" (not checked), and "Resultado" (not checked). Below the progress bar, the status is "Executando melhor Pipeline...".

(b) Resultados da execução realizada: This screenshot shows the results of the executed pipeline. It has the same layout as (a), with the "FairPEK" header and sidebar. The main area is titled "Pipeline Autônomo" and shows the completed steps: "Parametrização", "Análise", "Opções", "Execução" (checked), and "Resultado" (not checked). A detailed table under "Resultado" provides performance metrics:

Execução	Parâmetros utilizados e estatísticas da execução
Métricas de Performance	Métricas relacionadas a performance do modelo
Métricas de Fairness	Métricas relacionadas a justiça do modelo
Pontuação	Pontuação dada de acordo com os pesos e métricas utilizadas para a execução
Pontuação das métricas de performance	825
Pontuação das métricas de fairness	952
Pontuação geral	888

Figura 3.16: Execução do pipeline após seleção.

3.3 Experimentos e limitações

Foram realizados testes com o seguinte conjunto de dados e dado o seguinte objetivo:

- **Objetivo:** Obter classificação de crédito (boa ou ruim), através de uma série de *features*
- **Conjunto de dados:** German Credit Dataset, presente em [9].
- **Atributos protegidos:** Idade, ou Nacionalidade
- **Grupo privilegiado:** Idade maior ou igual a 25 anos; Nacionalidade Alemã
- **Grupo não-privilegiado:** Idade menor que 25 anos; Nacionalidade diferente da Alemã (Estrangeiro)

Para realizar este experimento, foi considerado o seguinte objetivo e consideradas as seguintes limitações:

- **Experimento:** Execução de um *pipeline* de ML para obtenção de dados iniciais na base de conhecimento e discussão de um *pipeline* de ML utilizando MAPE-K como facilitador da escolha de modelo.

- **Medição:** Serão medidas as pontuações de cada agrupamento de métricas (Performance e Fairness), e tais métricas também serão comparadas com o valor de cada uma para verificar seu impacto na pontuação.
- **Pré-condição 1:** Houveram *pipelines* que já foram pré-executados e já formaram uma base de conhecimento. Para o experimento, foi considerado ao menos 1 execução para cada combinação de conjunto de dados, atributo protegido e algoritmos utilizados.
- **Pré-condição 2:** A base de conhecimento será capaz de explicar a imparcialidade/justiça de um modelo, mas não será capaz de explicar a influência de cada *feature* aplicada no modelo.
- **Pré-condição 3:** Podem existir ruídos nos resultados finais devido ao German Credit Dataset ser uma base de dados com poucas amostras (apenas 1000), mas eles serão desconsiderados uma vez que o conjunto de dados ainda é considerado como *benchmark* em alguns trabalhos acadêmicos [45] [38] [30] e seus dados exemplificam muito bem uma situação real onde dados sensíveis podem ser utilizados e são possíveis de afetar a decisão de um modelo e, consequentemente, a situação de vida de uma pessoa.
- **Restrição 1:** O algoritmo de retirada de viés é feito em apenas uma parte das etapas (Pré-processamento/Processamento/Pós-processamento). Isto foi decidido pois não foram verificadas referências onde a aplicação desses algoritmos em duas ou em todas as três etapas impacta no desempenho.
- **Restrição 2:** Por não conseguirem rodar com sucesso, foram retirados os Pipelines que rodaram os seguintes algoritmos: Pré-processamento Otimizado.
- **Restrição 3:** Pipelines também foram retirados por apresentarem métricas com valores máximos, para evitar análises caso exista alguma falha não detectada na implementação. Como exemplo, os que possuíam Classificação baseada em Rejeição de Opções.
- **Restrição 4:** Para evitar pipelines com conjuntos de algoritmos ruins e também pelo mesmo motivo da restrição anterior, o intervalo de pontuação para análise foi limitado de 500 a 950.

Capítulo 4

Resultados e Discussões

A execução do Pipeline utilizando o componente MAPE-K foi realizada com 3 pesagens diferentes na pontuação geral:

- 50% para métricas de Performance e 50% para métricas de Fairness, para uma configuração equilibrada.
- 75% para métricas de Performance e 25% para métricas de Fairness, para uma configuração que prioriza a performance em detrimento da justiça.
- 25% para métricas de Performance e 75% para métricas de Fairness, para uma configuração que prioriza a justiça em detrimento da performance.

Em todas as execuções são utilizadas as métricas Acurácia, Precisão, *Recall*, *F1-Score* e AUC como métricas de Performance e as métricas *Statistical Parity Difference*, *Equal Opportunity Difference*, *Average Odds Difference*, *Disparate Impact* e *Theil Index* como métricas de Fairness, todas com pesagens iguais em seu respectivo agrupamento.

Os resultados baseados nas pré-condições e restrições já comentadas no capítulo anterior estão presentes abaixo nas Tabelas 4.1, 4.2 e 4.3:

Tabela 4.1: Melhores opções escolhidas pelo modelo MAPE-K
Todos os métodos - 50% Performance/50% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Régressão Logística	Equalized Odds	968	860	914
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	902	922	912
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	870	925	898
Idade	Nenhum	Gradient Boosting	Equalized Odds	927	862	894
Idade	Reweighting	Gradient Boosting	Nenhum	804	931	868

Tabela 4.2: Melhores opções escolhidas pelo modelo MAPE-K
Todos os métodos - 75% Performance/25% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Regressão Logística	Equalized Odds	968	860	941
Idade	Nenhum	Gradient Boosting	Equalized Odds	927	862	910
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	902	922	907
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	870	925	883
Idade	Nenhum	Random Forest	Equalized Odds	898	799	874

Tabela 4.3: Melhores opções escolhidas pelo modelo MAPE-K
Todos os métodos - 25% Performance/75% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	928
Nacionalidade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	928
Idade	Nenhum	Adversarial Debiasing	Nenhum	742	979	920
Nacionalidade	Reweighting	Support Vector Machines	Nenhum	755	972	918
Nacionalidade	Learning Fair Representations	Support Vector Machines	Nenhum	755	972	918

Nessas execuções, surpreende 2 observações. A primeira é o fato da predominância de algoritmos com redução de viés no pós-processamento/resultado especialmente em configurações que priorizavam performance, contrariando o esperado de que os algoritmos com redução de viés aumentavam justiça em detrimento da performance. A segunda é a predominância de algoritmos com redução de viés no pré-processamento/dado em configurações que priorizavam justiça, principalmente pois todos as execuções usavam *Support Vector Machines* como algoritmo de treinamento.

Diante destas 2 predominâncias envolvendo todos os pipelines executados, novos experimentos com restrições adicionais foram realizados para obter observações mais detalhadas a respeito dos resultados:

- Uso apenas de pipelines com o uso de algoritmos com redução de viés no pré - processamento/dado.
- Uso apenas de pipelines com o uso de algoritmos com redução de viés no processamento/treinamento.
- Uso apenas de pipelines com o uso de algoritmos com redução de viés no pós - processamento/resultado.
- Uso apenas de pipelines sem o uso de algoritmos com redução de viés.

As pré-condições, restrições e pesagens nas pontuações usadas anteriormente foram mantidas e seus resultados estão presentes abaixo nas Tabelas 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 e 4.15:

Tabela 4.4: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no dado - 50% Performance/50% Fairness

Atributo protegido	Pipeline			Pontuação		
	Pré-processamento	Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Reweighting	Gradient Boosting	Nenhum	804	931	868
Idade	Learning Fair Representations	Gradient Boosting	Nenhum	804	931	868
Idade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	868
Nacionalidade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	868
Nacionalidade	Learning Fair Representations	Support Vector Machines	Nenhum	755	972	864

Tabela 4.5: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no dado - 75% Performance/25% Fairness

Atributo protegido	Pipeline			Pontuação		
	Pré-processamento	Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Reweighting	Gradient Boosting	Nenhum	804	931	836
Idade	Learning Fair Representations	Gradient Boosting	Nenhum	804	931	836
Nacionalidade	Learning Fair Representations	Gradient Boosting	Nenhum	811	878	828
Nacionalidade	Reweighting	Gradient Boosting	Nenhum	811	878	828
Nacionalidade	Learning Fair Representations	Random Forest	Nenhum	801	883	821

Tabela 4.6: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no dado - 25% Performance/75% Fairness

Atributo protegido	Pipeline			Pontuação		
	Pré-processamento	Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	928
Nacionalidade	Disparate Impact Remover	Support Vector Machines	Nenhum	747	989	928
Nacionalidade	Learning Fair Representations	Support Vector Machines	Nenhum	755	972	918
Nacionalidade	Reweighting	Support Vector Machines	Nenhum	755	972	918
Idade	Learning Fair Representations	Support Vector Machines	Nenhum	755	969	916

Nos pipelines utilizando apenas algoritmos com redução de viés no dado, percebe-se a predominância dos algoritmos de treinamento *Gradient Boosting* e *Support Vector Machines*, sendo o *Gradient Boosting* predominante em configurações priorizando performance e o *Support Vector Machines* predominante em configurações priorizando justiça, o que começa a explicar a sua predominância também presente no resultado geral. Também é possível perceber mais 2 observações: A primeira observação é que a análise de apenas uma categoria de algoritmos dá mais clareza em ver como o cálculo utilizado nas 3 configurações faz com que o equilíbrio de ambas as métricas se torna mais importante do que a prioridade apenas em performance ou apenas em justiça, uma vez que há exemplos de conjuntos de algoritmos com pontuações ligeiramente maiores em performance que acabaram sendo pior avaliados pois a pontuação em *Fairness* está bem menor, e vice-versa. A segunda observação é que o uso de um atributo protegido diferente (e, por consequência, com tratamento de dados diferente) e de um algoritmo de treinamento parecem impactar tanto quanto ou até mais que o próprio algoritmo com redução de viés no dado.

Tabela 4.7: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no treinamento - 50% Performance/50% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Adversarial Debiasing	Nenhum	742	979	860
Nacionalidade	Nenhum	Grid Search Reduction	Nenhum	789	895	845
Idade	Nenhum	Meta Fair Classifier	Nenhum	776	910	843
Idade	Nenhum	Exponentiated Gradient Reduction	Nenhum	811	869	840
Nacionalidade	Nenhum	Rich Subgroup Fairness	Nenhum	791	856	824

Tabela 4.8: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no treinamento - 75% Performance/25% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Exponentiated Gradient Reduction	Nenhum	811	869	826
Nacionalidade	Nenhum	Grid Search Reduction	Nenhum	789	895	820
Idade	Nenhum	Meta Fair Classifier	Nenhum	776	910	809
Nacionalidade	Nenhum	Exponentiated Gradient Reduction	Nenhum	807	810	808
Nacionalidade	Nenhum	Rich Subgroup Fairness	Nenhum	791	856	807

Tabela 4.9: Melhores opções escolhidas pelo modelo MAPE-K
Apenas com redução de viés no treinamento - 25% Performance/75% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Adversarial Debiasing	Nenhum	742	979	920
Idade	Nenhum	Meta Fair Classifier	Nenhum	776	910	876
Nacionalidade	Nenhum	Grid Search Reduction	Nenhum	795	895	870
Idade	Nenhum	Exponentiated Gradient Reduction	Nenhum	811	869	854
Nacionalidade	Nenhum	Prejudice Remover	Nenhum	770	874	848

Nos pipelines utilizando apenas algoritmos com redução de viés no treinamento, percebe-se uma variedade maior nos algoritmos, até porque não há usos de redução de viés em um pré ou um pós-processamento, com destaque para o *Adversarial Debiasing* que foi bem avaliado pela pontuação alta em *Fairness*. Nestes pipelines, as 2 observações percebidas nos pipelines utilizando apenas algoritmos com redução de viés no dado são reforçadas por uma maior variedade de pontuações e pelo algoritmo *Exponentiated Gradient Reduction* com 2 exemplos diferentes na Tabela 4.8, onde o uso da Nacionalidade como atributo protegido possui pontuações de performance e *Fairness* piores que a Idade e conclundo que o processamento utilizado no atributo protegido pode afetar todas as métricas.

Tabela 4.10: Melhores opções escolhidas pelo modelo MAPE-K

Apenas com redução de viés no resultado - 50% Performance/50% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Ressonância Magnética	Equalized Odds	968	860	914
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	902	922	912
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	870	925	898
Idade	Nenhum	Gradient Boosting	Equalized Odds	927	862	894
Idade	Nenhum	Random Forest	Equalized Odds	898	799	849

Tabela 4.11: Melhores opções escolhidas pelo modelo MAPE-K

Apenas com redução de viés no resultado - 75% Performance/25% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Idade	Nenhum	Ressonância Magnética	Equalized Odds	968	860	941
Idade	Nenhum	Gradient Boosting	Equalized Odds	927	862	910
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	902	922	907
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	870	925	883
Idade	Nenhum	Random Forest	Equalized Odds	898	799	874

Tabela 4.12: Melhores opções escolhidas pelo modelo MAPE-K

Apenas com redução de viés no resultado - 25% Performance/75% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	902	922	917
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	870	925	911
Idade	Nenhum	Ressonância Magnética	Equalized Odds	968	860	887
Idade	Nenhum	Gradient Boosting	Equalized Odds	927	862	878
Idade	Nenhum	Random Forest	Equalized Odds	898	799	824

Nos pipelines utilizando apenas algoritmos com redução de viés no resultado, surpreende o fato de que os pipelines obtiveram as melhores pontuações em Performance e as piores pontuações em Fairness, podendo indicar uma característica dos algoritmos *Equalized Odds* e *Calibrated Equalized Odds*. Entretanto, por conta da grande melhora por parte das métricas de performance, estes pipelines possuem um maior equilíbrio entre Performance e Fairness e acabam garantindo maiores pontuações na média, justificando as melhores pontuações nas primeiras execuções onde foram considerados todos os métodos. Fora isto, as demais observações anteriores também se aplicam nestas execuções.

Tabela 4.13: Melhores opções escolhidas pelo modelo MAPE-K
Apenas sem redução de viés - 50% Performance/50% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Nacionalidade	Nenhum	Support Vector Machines	Nenhum	755	972	864
Idade	Nenhum	Support Vector Machines	Nenhum	755	969	862
Nacionalidade	Nenhum	Random Forest	Nenhum	802	885	844
Nacionalidade	Nenhum	Régressão Logística	Nenhum	782	865	824
Nacionalidade	Nenhum	Gradient Boosting	Nenhum	817	784	800

Tabela 4.14: Melhores opções escolhidas pelo modelo MAPE-K
Apenas sem redução de viés - 75% Performance/25% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Nacionalidade	Nenhum	Random Forest	Nenhum	802	885	823
Nacionalidade	Nenhum	Gradient Boosting	Nenhum	817	784	809
Nacionalidade	Nenhum	Support Vector Machines	Nenhum	755	972	809
Idade	Nenhum	Support Vector Machines	Nenhum	755	969	808
Idade	Nenhum	Gradient Boosting	Nenhum	817	778	807

Tabela 4.15: Melhores opções escolhidas pelo modelo MAPE-K
Apenas sem redução de viés - 25% Performance/75% Fairness

Atributo protegido	Pré-processamento	Pipeline		Pontuação		
		Treinamento	Pós-processamento	Performance	Fairness	Geral
Nacionalidade	Nenhum	Support Vector Machines	Nenhum	755	972	918
Idade	Nenhum	Support Vector Machines	Nenhum	755	969	916
Nacionalidade	Nenhum	Random Forest	Nenhum	802	885	864
Nacionalidade	Nenhum	Régressão Logística	Nenhum	782	865	844
Idade	Nenhum	Régressão Logística	Nenhum	782	802	797

Olhando os pipelines sem algoritmos com redução de viés, é curioso notar que, ao comparar com pipelines equivalentes mas com algoritmos usando redução de viés no dado, é possível notar que a hipótese principal se confirma em sua grande maioria: As pontuações em Performance são ligeiramente maiores e as pontuações em *Fairness* são ligeiramente menores. É possível notar uma exceção no pipeline envolvendo o algoritmo *Random Forest*, mas nos outros casos a hipótese é verificada com sucesso. Ao comparar com pipelines usando algoritmos com redução de viés no treinamento tal hipótese também se confirma, entretanto o uso de *Support Vector Machines* parece ser uma exceção a regra, implicando que o uso do algoritmo possibilita modelos mais justos para o conjunto de dados utilizado. Ao comparar com pipelines usando algoritmos com redução de viés no resultado, a hipótese não se confirma devido a observação da grande melhora por parte das métricas de performance nos pipelines usando algoritmos com redução de viés no resultado, mas ao verificar a pontuação em *Fairness* é possível notar uma ligeira melhora. Há a exceção de pipelines envolvendo *Support Vector Machines* que são melhores nos

pipelines sem algoritmos com redução de viés, mas no uso de outros algoritmos ocorre melhora na pontuação em *Fairness*.

Após a verificação das pontuações, foram catalogadas todas as métricas de todos os conjuntos de algoritmos em seus valores máximo, mínimo e médio para verificar a eficácia destas pontuações, definidas nas Tabelas 4.16 e 4.17 exibidas abaixo.

Tabela 4.16: Métricas de performance das execuções de Pipeline

Atributo protegido	Pré-processamento	Treinamento	Pos-processamento	Métricas						
				Acurácia	Precisão	Recall	F1-Score	AUC		
Nacionalidade	Nenhum	Support Vector Machines	Nenhum	Mínimo	0,715	Mínimo	0,7143	Mínimo	0,9929	Mínimo
				Máximo	0,715	Máximo	0,7143	Máximo	0,9929	Máximo
				Média	0,715	Média	0,7143	Média	0,9929	Média
Idade	Nenhum	Support Vector Machines	Nenhum	Mínimo	0,715	Mínimo	0,7143	Mínimo	0,9929	Mínimo
				Máximo	0,715	Máximo	0,7143	Máximo	0,9929	Máximo
				Média	0,715	Média	0,7143	Média	0,9929	Média
Nacionalidade	Nenhum	Random Forest	Nenhum	Mínimo	0,76	Mínimo	0,7784	Mínimo	0,9007	Mínimo
				Máximo	0,79	Máximo	0,8037	Máximo	0,9291	Máximo
				Média	0,774	Média	0,7933	Média	0,9192	Média
Nacionalidade	Nenhum	Regressão Logística	Nenhum	Mínimo	0,75	Mínimo	0,7661	Mínimo	0,9291	Mínimo
				Máximo	0,75	Máximo	0,7661	Máximo	0,9291	Máximo
				Média	0,75	Média	0,7661	Média	0,9291	Média
Nacionalidade	Nenhum	Gradient Boosting	Nenhum	Mínimo	0,79	Mínimo	0,8194	Mínimo	0,9007	Mínimo
				Máximo	0,79	Máximo	0,8194	Máximo	0,9007	Máximo
				Média	0,79	Média	0,8194	Média	0,9007	Média
Idade	Nenhum	Gradient Boosting	Nenhum	Mínimo	0,79	Mínimo	0,8235	Mínimo	0,8936	Mínimo
				Máximo	0,79	Máximo	0,8235	Máximo	0,8936	Máximo
				Média	0,79	Média	0,8235	Média	0,8936	Média
Idade	Nenhum	Regressão Logística	Nenhum	Mínimo	0,75	Mínimo	0,7661	Mínimo	0,9291	Mínimo
				Máximo	0,75	Máximo	0,7661	Máximo	0,9291	Máximo
				Média	0,75	Média	0,7661	Média	0,9291	Média
Idade	Reweighting	Gradient Boosting	Nenhum	Mínimo	0,75	Mínimo	0,8	Mínimo	0,9078	Mínimo
				Máximo	0,775	Máximo	0,8	Máximo	0,9078	Máximo
				Média	0,775	Média	0,8	Média	0,9078	Média
Idade	Learning Fair Representations	Gradient Boosting	Nenhum	Mínimo	0,775	Mínimo	0,8	Mínimo	0,9078	Mínimo
				Máximo	0,775	Máximo	0,8	Máximo	0,9078	Máximo
				Média	0,775	Média	0,8	Média	0,9078	Média
Idade	Disparate Impact Remover	Support Vector Machines	Nenhum	Mínimo	0,705	Mínimo	0,705	Mínimo	1	Mínimo
				Máximo	0,705	Máximo	0,705	Máximo	1	Máximo
				Média	0,705	Média	0,705	Média	1	Média
Nacionalidade	Disparate Impact Remover	Support Vector Machines	Nenhum	Mínimo	0,705	Mínimo	0,705	Mínimo	1	Mínimo
				Máximo	0,705	Máximo	0,705	Máximo	1	Máximo
				Média	0,705	Média	0,705	Média	1	Média
Nacionalidade	Learning Fair Representations	Support Vector Machines	Nenhum	Mínimo	0,715	Mínimo	0,7143	Mínimo	0,9929	Mínimo
				Máximo	0,715	Máximo	0,7143	Máximo	0,9929	Máximo
				Média	0,715	Média	0,7143	Média	0,9929	Média
Nacionalidade	Learning Fair Representations	Gradient Boosting	Nenhum	Mínimo	0,785	Mínimo	0,8063	Mínimo	0,9149	Mínimo
				Máximo	0,785	Máximo	0,8063	Máximo	0,9149	Máximo
				Média	0,785	Média	0,8063	Média	0,9149	Média
Nacionalidade	Reweighting	Gradient Boosting	Nenhum	Mínimo	0,785	Mínimo	0,8063	Mínimo	0,9149	Mínimo
				Máximo	0,785	Máximo	0,8063	Máximo	0,9149	Máximo
				Média	0,785	Média	0,8063	Média	0,9149	Média
Nacionalidade	Learning Fair Representations	Random Forest	Nenhum	Mínimo	0,76	Mínimo	0,7831	Mínimo	0,8936	Mínimo
				Máximo	0,79	Máximo	0,8113	Máximo	0,922	Máximo
				Média	0,772	Média	0,7919	Média	0,9121	Média
Nacionalidade	Reweighting	Support Vector Machines	Nenhum	Mínimo	0,715	Mínimo	0,7143	Mínimo	0,9929	Mínimo
				Máximo	0,715	Máximo	0,7143	Máximo	0,9929	Máximo
				Média	0,715	Média	0,7143	Média	0,9929	Média
Idade	Learning Fair Representations	Support Vector Machines	Nenhum	Mínimo	0,715	Mínimo	0,7143	Mínimo	0,9929	Mínimo
				Máximo	0,715	Máximo	0,7143	Máximo	0,9929	Máximo
				Média	0,715	Média	0,7143	Média	0,9929	Média
Idade	Nenhum	Adversarial Debiasing	Nenhum	Mínimo	0,295	Mínimo	0	Mínimo	0	Mínimo
				Máximo	0,705	Máximo	0,7097	Máximo	1	Máximo
				Média	0,6317	Média	0,5888	Média	0,8168	Média
Nacionalidade	Grid Search Reduction	Nenhum	Nenhum	Mínimo	0,75	Mínimo	0,7791	Mínimo	0,9007	Mínimo
				Máximo	0,78	Máximo	0,7939	Máximo	0,9362	Máximo
				Média	0,765	Média	0,7857	Média	0,9167	Média
Idade	Nenhum	Meta Fair Classifier	Nenhum	Mínimo	0,75	Mínimo	0,7278	Mínimo	0,9929	Mínimo
				Máximo	0,75	Máximo	0,7278	Máximo	0,9929	Máximo
				Média	0,75	Média	0,7278	Média	0,9929	Média
Idade	Nenhum	Exponentiated Gradient Reduction	Nenhum	Mínimo	0,75	Mínimo	0,7791	Mínimo	0,9007	Mínimo
				Máximo	0,785	Máximo	0,8063	Máximo	0,9149	Máximo
				Média	0,785	Média	0,7863	Média	0,9149	Média
Nacionalidade	Nenhum	Rich Subgroup Fairness	Nenhum	Mínimo	0,76	Mínimo	0,808	Mínimo	0,8653	Mínimo
				Máximo	0,76	Máximo	0,808	Máximo	0,8653	Máximo
				Média	0,76	Média	0,808	Média	0,8653	Média
Nacionalidade	Nenhum	Exponentiated Gradient Reduction	Nenhum	Mínimo	0,775	Mínimo	0,8038	Mínimo	0,9007	Mínimo
				Máximo	0,785	Máximo	0,8101	Máximo	0,9078	Máximo
				Média	0,7788	Média	0,8057	Média	0,9043	Média
Nacionalidade	Nenhum	Prejudice Remover	Nenhum	Mínimo	0,735	Mínimo	0,7683	Mínimo	0,8936	Mínimo
				Máximo	0,735	Máximo	0,7683	Máximo	0,8936	Máximo
				Média	0,735	Média	0,7683	Média	0,8936	Média
Idade	Nenhum	Regressão Logística	Equalized Odds	Mínimo	0,965	Mínimo	0,9589	Mínimo	0,9929	Mínimo
				Máximo	0,965	Máximo	0,9589	Máximo	0,9929	Máximo
				Média	0,965	Média	0,9589	Média	0,9929	Média
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	Mínimo	0,82	Mínimo	0,7966	Mínimo	1	Mínimo
				Máximo	0,93	Máximo	0,9097	Máximo	1	Máximo
				Média	0,8925	Média	0,87	Média	1	Média
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	Mínimo	0,875	Mínimo	0,8494	Mínimo	1	Mínimo
				Máximo	0,875	Máximo	0,8494	Máximo	1	Máximo
				Média	0,855	Média	0,8296	Média	1	Média
Idade	Nenhum	Gradient Boosting	Equalized Odds	Mínimo	0,903	Mínimo	0,9769	Mínimo	0,8723	Mínimo
				Máximo	0,915	Máximo	0,9919	Máximo	0,9007	Máximo
				Média	0,9083	Média	0,9869	Média	0,8818	Média
Nacionalidade	Nenhum	Random Forest	Equalized Odds	Mínimo	0,84	Mínimo	0,8150	Mínimo	1	Mínimo
				Máximo	0,875	Máximo	0,8494	Máximo	1	Máximo
				Média	0,8733	Média	0,8798	Média	1	Média

Tabela 4.17: Métricas de Fairness das execuções de Pipeline

Atributo protegido	Pré-processamento	Treinamento	Pós-processamento	Métricas							
				Statistical Parity Difference	Equal Opportunity Difference	Average Odds Difference	Disparate Impact	Theil Index			
Nacionalidade	Nenhum	Support Vector Machines	Nenhum	Mínimo -0,209	Mínimo -0,0075	Mínimo -0,296	Mínimo 0,791	Mínimo 0,0615			
				Máximo 0,209	Máximo -0,0075	Máximo 0,791	Máximo 0,970	Máximo 0,0615			
				Média 0,209	Média -0,0075	Média 0,791	Média 0,970	Média 0,0615			
Idade	Nenhum	Support Vector Machines	Nenhum	Mínimo 0,228	Mínimo 0,0084	Mínimo 0,348	Mínimo 1,924	Mínimo 0,0615			
				Máximo 0,228	Máximo 0,0084	Máximo 0,348	Máximo 1,924	Máximo 0,0615			
				Média 0,228	Média 0,0084	Média 0,348	Média 1,924	Média 0,0615			
Nacionalidade	Nenhum	Random Forest	Nenhum	Mínimo -0,0831	Mínimo 0,273	Mínimo -0,2191	Mínimo 0,8894	Mínimo 0,0955			
				Máximo 0,0839	Máximo 0,1899	Máximo -0,1378	Máximo 1,301	Máximo 0,1173			
				Média 0,0830	Média 0,1733	Média -0,1806	Média 0,9428	Média 0,1045			
Nacionalidade	Nenhum	Regressão Logística	Nenhum	Mínimo -0,1518	Mínimo -0,0752	Mínimo -0,2014	Mínimo 0,8482	Mínimo 0,1013			
				Máximo -0,1518	Máximo -0,0752	Máximo -0,2014	Máximo 0,8482	Máximo 0,1013			
				Média -0,1518	Média -0,0752	Média -0,2014	Média 0,8482	Média 0,1013			
Nacionalidade	Nenhum	Gradient Boosting	Nenhum	Mínimo 0,1134	Mínimo 0,2923	Mínimo -0,1211	Mínimo 1,1702	Mínimo 0,1137			
				Máximo 0,1134	Máximo 0,2923	Máximo -0,1211	Máximo 1,1702	Máximo 0,1137			
				Média 0,1134	Média 0,2923	Média -0,1211	Média 1,1702	Média 0,1137			
Idade	Nenhum	Gradient Boosting	Nenhum	Mínimo -0,2111	Mínimo -0,1971	Mínimo -0,2337	Mínimo 0,7	Mínimo 0,1183			
				Máximo -0,2111	Máximo -0,1971	Máximo -0,2337	Máximo 0,7	Máximo 0,1183			
				Média -0,2111	Média -0,1971	Média -0,2337	Média 0,7	Média 0,1183			
Idade	Nenhum	Regressão Logística	Nenhum	Mínimo -0,1518	Mínimo -0,0752	Mínimo -0,2014	Mínimo 0,8482	Mínimo 0,1013			
				Máximo -0,1518	Máximo -0,0752	Máximo -0,2014	Máximo 0,8482	Máximo 0,1013			
				Média -0,1518	Média -0,0752	Média -0,2014	Média 0,8482	Média 0,1013			
Idade	Reweighting	Gradient Boosting	Nenhum	Mínimo -0,0505	Mínimo -0,0523	Mínimo -0,0517	Mínimo 0,9265	Mínimo 0,1118			
				Máximo -0,0505	Máximo -0,0523	Máximo -0,0517	Máximo 0,9265	Máximo 0,1118			
				Média -0,0505	Média -0,0523	Média -0,0517	Média 0,9265	Média 0,1118			
Idade	Learning Fair Representations	Gradient Boosting	Nenhum	Mínimo -0,0505	Mínimo -0,0523	Mínimo -0,0517	Mínimo 0,9265	Mínimo 0,1118			
				Máximo -0,0505	Máximo -0,0523	Máximo -0,0517	Máximo 0,9265	Máximo 0,1118			
				Média -0,0505	Média -0,0523	Média -0,0517	Média 0,9265	Média 0,1118			
Idade	Disparate Impact Remover	Support Vector Machines	Nenhum	Mínimo 0	Mínimo 0	Mínimo 0	Mínimo 1	Mínimo 0,0573			
				Máximo 0	Máximo 0	Máximo 1	Máximo 0,0573	Máximo 0,0573			
				Média 0	Média 0	Média 1	Média 0,0573	Média 0,0573			
Nacionalidade	Disparate Impact Remover	Support Vector Machines	Nenhum	Mínimo 0	Mínimo 0	Mínimo 0	Mínimo 1	Mínimo 0,0573			
				Máximo 0	Máximo 0	Máximo 1	Máximo 0,0573	Máximo 0,0573			
				Média 0	Média 0	Média 1	Média 0,0573	Média 0,0573			
Nacionalidade	Learning Fair Representations	Support Vector Machines	Nenhum	Mínimo -0,0209	Mínimo -0,0075	Mínimo -0,296	Mínimo 0,9701	Mínimo 0,0615			
				Máximo -0,0209	Máximo -0,0075	Máximo -0,296	Máximo 0,9701	Máximo 0,0615			
				Média -0,0209	Média -0,0075	Média -0,296	Média 0,9701	Média 0,0615			
Nacionalidade	Learning Fair Representations	Gradient Boosting	Nenhum	Mínimo -0,0931	Mínimo 0,0423	Mínimo -0,2302	Mínimo 0,8953	Mínimo 0,1055			
				Máximo -0,0931	Máximo 0,0423	Máximo -0,2302	Máximo 0,8953	Máximo 0,1055			
				Média -0,0931	Média 0,0423	Média -0,2302	Média 0,8953	Média 0,1055			
Nacionalidade	Reweighting	Gradient Boosting	Nenhum	Mínimo -0,0931	Mínimo 0,0423	Mínimo -0,2302	Mínimo 0,8953	Mínimo 0,1055			
				Máximo -0,0931	Máximo 0,0423	Máximo -0,2302	Máximo 0,8953	Máximo 0,1055			
				Média -0,0931	Média 0,0423	Média -0,2302	Média 0,8953	Média 0,1055			
Nacionalidade	Learning Fair Representations	Random Forest	Nenhum	Mínimo -0,0983	Mínimo 0,0197	Mínimo -0,2289	Mínimo 0,8894	Mínimo 0,1025			
				Máximo -0,0983	Máximo 0,0197	Máximo -0,1405	Máximo 0,8894	Máximo 0,1025			
				Média -0,0983	Média 0,0197	Média -0,1405	Média 0,8894	Média 0,1025			
Nacionalidade	Reweighting	Support Vector Machines	Nenhum	Mínimo -0,0209	Mínimo -0,0075	Mínimo -0,296	Mínimo 0,9701	Mínimo 0,0615			
				Máximo -0,0209	Máximo -0,0075	Máximo -0,296	Máximo 0,9701	Máximo 0,0615			
				Média -0,0209	Média -0,0075	Média -0,296	Média 0,9701	Média 0,0615			
Idade	Learning Fair Representations	Support Vector Machines	Nenhum	Mínimo -0,0238	Mínimo 0,0084	Mínimo -0,0080	Mínimo 1	Mínimo 0,0573			
				Máximo -0,0238	Máximo 0,0084	Máximo -0,0080	Máximo 1	Máximo 0,0573			
				Média -0,0238	Média 0,0084	Média -0,0080	Média 1	Média 0,0573			
Idade	Nenhum	Adversarial Debiasing	Nenhum	Mínimo 0	Mínimo 0	Mínimo -0,0080	Mínimo 1	Mínimo 0,0573			
				Máximo 0,0104	Máximo 0,0412	Máximo -0,0017	Máximo 1	Máximo 0,0573			
				Média 0,0104	Média 0,0412	Média -0,0017	Média 1	Média 0,0573			
Nacionalidade	Nenhum	Grid Search Reduction	Nenhum	Mínimo -0,0836	Mínimo 0,0273	Mínimo -0,1933	Mínimo 0,9071	Mínimo 0,0932			
				Máximo -0,0836	Máximo 0,0273	Máximo -0,1670	Máximo 0,9124	Máximo 0,1204			
				Média -0,0836	Média 0,0273	Média -0,1827	Média 0,9128	Média 0,1076			
Idade	Nenhum	Meta Fair Classifier	Nenhum	Mínimo -0,1548	Mínimo -0,0913	Mínimo -0,1849	Mínimo 0,8289	Mínimo 0,0652			
				Máximo -0,1548	Máximo -0,0913	Máximo -0,0406	Máximo 0,8783	Máximo 0,1013			
				Média -0,1548	Média -0,0913	Média -0,1186	Média 0,8979	Média 0,0802			
Idade	Nenhum	Exponentiated Gradient Reduction	Nenhum	Mínimo -0,1339	Mínimo -0,1146	Mínimo -0,1328	Mínimo 0,8387	Mínimo 0,1055			
				Máximo -0,1339	Máximo -0,1146	Máximo -0,1328	Máximo 0,8387	Máximo 0,1055			
				Média -0,1339	Média -0,1146	Média -0,1328	Média 0,8387	Média 0,1055			
Nacionalidade	Nenhum	Rich Subgroup Fairness	Nenhum	Mínimo -0,1402	Mínimo -0,0103	Mínimo -0,2638	Mínimo 0,8423	Mínimo 0,1427			
				Máximo -0,1402	Máximo -0,0103	Máximo -0,2638	Máximo 0,8423	Máximo 0,1427			
				Média -0,1402	Média -0,0103	Média -0,2638	Média 0,8423	Média 0,1427			
Nacionalidade	Nenhum	Exponentiated Gradient Reduction	Nenhum	Mínimo -0,2199	Mínimo -0,1653	Mínimo -0,2080	Mínimo 0,7801	Mínimo 0,1101			
				Máximo -0,2147	Máximo -0,0977	Máximo -0,2093	Máximo 0,7853	Máximo 0,1165			
				Média -0,2186	Média -0,1015	Média -0,2093	Média 0,7814	Média 0,1135			
Nacionalidade	Nenhum	Prejudice Remover	Nenhum	Mínimo 0,0442	Mínimo 0,1523	Mínimo -0,1049	Mínimo 1,057	Mínimo 0,1274			
				Máximo 0,0442	Máximo 0,1523	Máximo -0,1049	Máximo 1,057	Máximo 0,1274			
				Média 0,0442	Média 0,1523	Média -0,1049	Média 1,057	Média 0,1274			
Idade	Nenhum	Regressão Logística	Equalized Odds	Mínimo -0,1726	Mínimo 0,0084	Mínimo 0,3042	Mínimo 1,2458	Mínimo 0,1159			
				Máximo -0,1726	Máximo 0,0084	Máximo 0,3042	Máximo 1,2458	Máximo 0,1159			
				Média -0,1726	Média 0,0084	Média 0,3042	Média 1,2458	Média 0,1159			
Nacionalidade	Nenhum	Random Forest	Calibrated Equalized Odds	Mínimo -0,1193	Mínimo 0	Mínimo -0,1207	Mínimo 0,8658	Mínimo 0,02303			
				Máximo -0,0841	Máximo 0	Máximo 0,3103	Máximo 0,9594	Máximo 0,046			
				Média -0,0841	Média 0	Média 0,1853	Média 0,9594	Média 0,0314			
Nacionalidade	Nenhum	Gradient Boosting	Calibrated Equalized Odds	Mínimo -0,0617	Mínimo 0	Mínimo 0,2155	Mínimo 0,9306	Mínimo 0,0362			
				Máximo -0,0205	Máximo 0	Máximo 0,2759	Máximo 0,9719	Máximo 0,0428			
				Média -0,0407	Média 0	Média 0,25	Média 0,9512	Média 0,0401			
Idade											

métricas com contextos completamente diferentes ainda se torna difuso diante da grande quantidade de métricas e conjuntos de algoritmos utilizados. Além disso, a diferença entre as métricas é extremamente pequena e dificulta ainda mais a escolha. Nesse contexto, a consolidação das métricas em grupos simplifica a visualização de quais *pipelines* são mais equilibrados, e o uso de pesos para cada métrica e para cada grupo pode calibrar qual o melhor equilíbrio desejado para determinada situação.

Deste modo, pode-se concluir também que, em um contexto de desenvolvimento, o processo simplifica a decisão do Cientista de Dados e reduz显著mente o tempo para obtenção e implantação de um modelo otimizado, pois não exigirá execuções em diversos algoritmos uma vez que já há uma base de conhecimento prévia. Além disso, poderá poupar processamento e custos para a resolução de diversos outros problemas, uma vez que as execuções economizadas pelas equipes que utilizariam esse processo abrem margem para que outras equipes utilizem esse processamento.

O uso da AI Reference Architecture para definição dos papéis permite visualizar com clareza quais pessoas, quais etapas e projetos para desenvolvimento de funcionalidades e aplicações são necessários para ir da obtenção dos dados, passando pela implantação do modelo até chegar ao consumo pelo cliente final. Deste modo, é possível traçar melhores planejamentos para desenvolvimento de uma aplicação baseada em Inteligência Artificial.

O uso de *Assurance Cases* permitiu uma melhor visualização em no contexto da aplicação a ser implementada, com suas funcionalidades, objetivos e algoritmos a serem implementados e cumpridos. Desta forma, também é possível dividir melhor as tarefas para uma equipe implementá-la e permitir que todos os membros tenham uma visão de todos os detalhes a serem implementados e testados.

```

1  def data_postprocess(self, test_pipe, prediction_pipe, fairness_pipe
2      , unbias_postproc_algorithm):
3      unbias_postproc_options = [
4          (UnbiasPostProcAlgorithms.EQUALIZED_ODDS,
5           EqualizedOddsFilter()),
6          (UnbiasPostProcAlgorithms.CALIBRATED_EQUALIZED_ODDS,
7           CalibratedEqualizedOddsFilter()),
8          (UnbiasPostProcAlgorithms.REJECT_OPTION_CLASSIFICATION,
9           RejectOptionClassificationFilter())
10         ]
11
12
13     for option, filter in unbias_postproc_options:
14         if unbias_postproc_algorithm == option:
15             init_pipe = test_pipe + prediction_pipe + fairness_pipe[
16                 'unprivileged_group', 'privileged_group']
17             init_pipe >= filter == prediction_pipe
18             break
19
20
21     return prediction_pipe

```

Código 4.1: Método para escolha do algoritmo com redução de viés no pós-processamento

Quanto a Arquitetura de Software, o uso da arquitetura *Pipe-and-Filter* permite o encapsulamento dos algoritmos e a separação de interesses de forma simples, fazendo com que a parte de código existente para o *Pipeline* possa ter escolhas mais interessantes e

elegantes para um bom *Design* do código. Como exemplo disso há o Trecho 4.1, onde há grande flexibilidade para configurar o método de pós-processamento e ele pode ser expandido conforme novas classes de filtro forem implementadas sem grande esforço. A maior clareza do código levou a uma maior rapidez para seu entendimento e para novas implementações, uma vez que a separação dos interesses permite um código mais coeso, onde contextos diferentes são entendidos de maneira separada e é possível notar rapidamente qual a parte defeituosa caso um diagnóstico de *bugs* seja necessário.

Ao usar a arquitetura MAPE-K para possibilitar uma escolha autônoma de algoritmos e processamento do conjunto de dados foram notadas algumas vantagens. É um modelo de organização conhecido, o que facilita a manutenção do desenvolvedor que já possui conhecimento desta arquitetura. Ela permite separar muito bem as etapas para executar um plano e, com isso, reconfigurar o *pipeline* para executar as opções com melhores resultados. Esta separação também auxilia na manutenção, uma vez que o ciclo de monitoria, análise, planejamento, execução e obtenção de conhecimento é um *workflow* muito bem definido para análise de dados e execução de ações, facilitando o entendimento de seu funcionamento e, consequentemente, de seu código.

Embora o MAPE-K permita o desenvolvimento de uma aplicação autônoma, isso não significa que ela seja completamente automatizada para qualquer problema relacionado a *Machine Learning*, necessitando de ação humana para funcionar. A principal limitação por parte do desenvolvimento é que a biblioteca AIF360 suporta apenas problemas de classificação binária, e para evoluções e novos métodos é provável que ocorram refatorações no *pipeline*. Também há de se considerar que, mesmo que o pipeline evolua para abrigar outros tipos de problemas, o contexto do problema é importante ao se avaliar se o modelo é considerado bom ou não. O uso de pesos para as métricas e diferentes estratégias nas fases de análise e planejamento do MAPE-K ajudam a definir o contexto para uma avaliação, mas ainda vai depender de um Cientista de Dados e/ou de um especialista de Domínio para entender quais as necessidades do problema analisado e se os resultados são aceitáveis para a publicação de um modelo otimizado.

Ao usar uma interface humano-computador para intermediar as interações entre o componente MAPE-K e as escolhas de um usuário, houve uma maneira completamente diferente de como enxergar as soluções que podem ser propostas. Inicialmente, teve-se a ideia de que a autonomia seria contínua, em uma espécie de *pipeline* completamente "online", onde era possível obter resultados imediatos com o deploy de novos modelos através de um componente orquestrador das etapas do MAPE-K. Entretanto, foi notado que tal abordagem dificultava o entendimento de seu funcionamento, motivando a elaboração da interface. Após o desenvolvimento da mesma, foi possível notar e explicar melhor como funcionam as etapas, como funciona o cálculo para análise e como as configurações presentes para a etapa de planejamento afetam o resultado final, implicando em resultados mais eficientes para o treinamento de novos usuários.

O uso da interface também levou a uma conclusão inesperada: Com alguns ajustes, é possível adaptar o sistema para processos de MLOps (ver Apêndice A.4), uma vez que a base de conhecimento gerada pode ajudar na decisão de retornar modelos mais antigos, porém com menos ruídos em seus dados e, consequentemente, melhores métricas no geral. Isso necessitaria de algumas adaptações, como um sistema para versionamento

dos conjuntos de dados para armazenamento e economia de espaço, funcionalidades como opções para notificação em casos como piora das métricas e opções de visualização dos dados, e modificações no *pipeline* para deixá-lo mais flexível, robusto e com suporte a técnicas bastante utilizadas em *Machine Learning* como *Data Augmentation* e *K-Fold Cross-Validation*.

Capítulo 5

Conclusões

- Falar da possibilidade, viabilidade e resumir discussões do capítulo anterior
 - Mostrar trade-offs
 - Mostrar situações de uso e não uso
 - Como trabalho futuro, citar possíveis ajustes para evolução do processo

Referências Bibliográficas

- [1] Ai fairness 360. URL <https://aif360.mybluemix.net>.
- [2] Statlog (german credit data) data set. URL <https://airflow.apache.org/docs/apache-airflow/stable/index.html>.
- [3] What's the difference between artificial intelligence, machine learning and deep learning? URL <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>.
- [4] IBM - Analytics and AI architecture. URL <https://www.ibm.com/cloud/architecture/architectures/aiAnalyticsArchitecture/reference-architecture/>.
- [5] Machine learning: o que é e qual sua importância? URL https://www.sas.com/pt_br/insights/analytics/machine-learning.html.
- [6] Aprendizado de máquina. URL https://pt.wikipedia.org/wiki/Aprendizado_de_máquina.
- [7] Python - special method names. URL <https://docs.python.org/3/reference/datamodel.html#special-method-names>.
- [8] scikit-learn. URL <https://scikit-learn.org>.
- [9] Statlog (german credit data) data set. URL [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)).
- [10] An architectural blueprint for autonomic computing. Technical report, IBM, 2005. URL <https://www-03.ibm.com/autonomic/pdfs/AC%20Blueprint%20White%20Paper%20V7.pdf>.
- [11] *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009. URL <https://hastie.su.domains/Papers/ESLII.pdf>.
- [12] Data engineering - introduction and epochs. Technical report, panoply.io, 2017. URL <https://www-03.ibm.com/autonomic/pdfs/AC%20Blueprint%20White%20Paper%20V7.pdf>.

- [13] Machine learning: things are getting intense, 2018. URL <https://www2.deloitte.com/content/dam/Deloitte/global/Images/infographics/technologymediatelecommunications/gx-deloitte-tmt-2018-intense-machine-learning-report.pdf>.
- [14] Brasil se destaca com 42% das iniciativas de IA na América Latina, em 2020, 2021. URL <https://cio.com.br/tendencias/brasil-se-destaca-com-42-das-iniciativas-de-ia-na-america-latina-em-2020/>.
- [15] Proteção de Dados - LGPD, 2021. URL <https://www.gov.br/defesa/pt-br/acesso-a-informacao/lei-geral-de-protecao-de-dados-pessoais-lgpd>.
- [16] Nadeem Abbas, Jesper Andersson, and Welf Löwe. Autonomic software product lines. pages 324–331, 2010.
- [17] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.
- [18] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129, 2019.
- [19] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.
- [20] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. Explainability for fair machine learning, 2021.
- [21] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, and Alexander S. Szalay. Realizing the potential of data science. *Communications of the ACM*, 61:67–72, 2018.
- [22] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. page 642–653, 2020.
- [23] Jan Bosch. Software architecture: The next step. pages 194–199, 2004.
- [24] Rajendra Bose and James Frew. Lineage retrieval for scientific data processing: A survey. *ACM Comput. Surv.*, page 1–28, 2005.
- [25] L Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

- [26] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where: A characterization of data provenance. In *Proceedings of the 8th International Conference on Database Theory*, page 316–330, 2001.
- [27] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, 2018.
- [28] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.
- [29] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Nateesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 2017. URL <https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf>.
- [30] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.
- [31] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *22nd ACM SIGKDD International Conference*, pages 785–794, 2016.
- [32] Brian d’Alessandro, Cathy O’Neil, and Tom LaGatta. Conscientious classification: A data scientist’s guide to discrimination-aware classification. *Big Data*, pages 120–134, 2017.
- [33] Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, page 1345–1350, 2008.
- [34] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.
- [35] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- [36] Luc (Editor, Beth Plale, Simon Miles, Carole Goble, Paolo Missier, Roger Barga, Yogesh Simmhan, Joe Futrelle, Robert McGrath, James Myers, Patrick Paulson, Shawn Bowers, Bertram Ludäscher, Natalia Kwasnikowska, Jan Van den Bussche, Tommy Ellqvist, Juliana Freire, and Paul Groth. The open provenance model (v1.01). 2009.
- [37] R. J. Erb. Introduction to backpropagation neural network computation. *Pharmaceutical Research*, 10:165–170, 1993.

- [38] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015. URL <https://doi.org/10.1145/2783258.2783311>.
- [39] Keith D. Foote. A brief history of data science, 2021. URL <https://www.dataversity.net/brief-history-data-science/>.
- [40] Martin Fowler. Fluentinterface, 2005. URL <https://martinfowler.com/bliki/FluentInterface.html>.
- [41] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 2000.
- [42] David Garlan and Mary Shaw. An introduction to software architecture. In *Advances in Software Engineering and Knowledge Engineering*, 1993.
- [43] Alex Graves. Generating sequences with recurrent neural networks. 2013.
- [44] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016. URL <https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf>.
- [45] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 2011. URL https://www.researchgate.net/publication/228975972_Data_Pre-Processing_Techniques_for_Classification_without_Discrimination.
- [46] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.
- [47] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda B. Viégas, and Michael Terry. XRAI: better attributions through regions. In *IEEE/CVF International Conference on Computer Vision*, pages 4947–4956, 2019.
- [48] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.
- [49] Jeffrey Kephart and D.M. Chess. The vision of autonomic computing. pages 41 – 50, 2003.
- [50] Bahador Khalegi. The how of explainable ai: Explainable modelling, 2019. URL <https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed>.

- [51] Bahador Khalegi. The how of explainable ai: Post-modelling explainability, 2019. URL <https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f>.
- [52] Bahador Khalegi. The how of explainable ai: Pre-modelling explainability, 2019. URL <https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4>.
- [53] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.
- [54] Rezvan Mahdavi-hezaveh, Jacob Dremann, and Laurie Williams. Software development with feature toggles: practices used by practitioners. *Empirical Software Engineering*, 2021.
- [55] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. 2013.
- [56] Warren Mcculloch and Walter Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127–147, 1943.
- [57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, pages 1–35, 2021.
- [58] Luc Moreau, Bertram Ludäscher, Ilkay Altintas, Roger Barga, Shawn Bowers, Steven Callahan, George Chin, Ben Clifford, Shirley Cohen, Sarah Cohen-Boulakia, Susan Davidson, Ewa Deelman, Luciano Digiampietri, Ian Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, and Jun Zhao. The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 2007.
- [59] Luc Moreau, Paul Groth, Simon Miles, Javier Vázquez-Salceda, John Ibbotson, Jian Sheng, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, and László Varga. The provenance of electronic data. *Communications of the ACM*, pages 52–58, 2008.
- [60] Carlos Mougan, José Manuel Álvarez, Gourab K. Patro, Salvatore Ruggieri, and Steffen Staab. Fairness implications of encoding protected categorical attributes. 2022.
- [61] Tomoki Nakamaru and Shigeru Chiba. Generating a generic fluent api in java. *The Art, Science, and Engineering of Programming*, 2020. URL <http://dx.doi.org/10.22152/programming-journal.org/2020/4/9>.
- [62] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017. URL <https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526ffffbeb2d39ab038d1cd7-Paper.pdf>.

- [63] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 1964.
- [64] Daphne Raban and Avishag Gordon. The evolution of data science and big data research: A bibliometric analysis. *Scientometrics*, 122:1563–1581, 2020.
- [65] Md Tajmilur Rahman, Louis-Philippe Querel, Peter C. Rigby, and Bram Adams. Feature toggles: Practitioner practices and a case study. In *Proceedings of the 13th International Conference on Mining Software Repositories*, page 201–211. Association for Computing Machinery, 2016.
- [66] C.v Ramamoorthy and Benjamin Wah. Knowledge and data engineering. *IEEE Transactions on Knowledge and Data Engineering*, 1:9 – 16, 1989.
- [67] Lior Rokach and Oded Maimon. Top-down induction of decision trees classifiers—a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35:476 – 487, 2005.
- [68] Mark Schmidt, Nicolas Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162, 2013.
- [69] Yashpal Singh and Alok Singh Chauhan. Neural networks in data mining. *Journal of Theoretical & Applied Information Technology*, 5, 2009.
- [70] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness. 2018. URL <http://dx.doi.org/10.1145/3219819.3220046>.
- [71] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008. URL <http://www.lamda.nju.edu.cn/zhangt/seminar/RKHS/pdf/Support%20Vector%20Machines%20-%20Ingo%20Steinwart,%20Andreas%20Christmann.pdf>.
- [72] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 2017.
- [73] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. *30th International Conference on Machine Learning, ICML 2013*, pages 1139–1147, 01 2013.
- [74] Wang-chiew Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, pages 3–12, 2007.
- [75] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. URL <https://www.wi.hs-wismar.de/~cleve/vorl/projects/dm/ss13/HierarClustern/Literatur/WittenFrank-DM-3rd.pdf>.

- [76] Allison Woodruff and Michael Stonebraker. Supporting fine-grained data lineage in a database visualization environment. Technical report, EECS Department, University of California, Berkeley, 1997.
- [77] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333, 2013. URL <https://proceedings.mlr.press/v28/zemel13.html>.
- [78] Brian Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. pages 335–340, 2018. URL https://www.researchgate.net/publication/330299272_Mitigating_Unwanted_Biases_with_Adversarial_Learning.

Apêndice A

Conceitos complementares

A.1 AI Explainability

AI Explainability é um conceito em IA que propõe a criação de um conjunto de técnicas de Aprendizado de Máquina (ou Machine Learning/ML) que produz modelos mais explicáveis, mantendo qualidade em suas métricas, e permite que os humanos entendam, confiem e gerenciem aplicações baseadas em IA. XAI também absorve conceitos das Ciências Sociais e considera a psicologia da explicaçāo [19]. Um algoritmo de ML explicável precisa não apenas mostrar tomadas de decisão, mas também mostrar o processo que o levou ao tomar tal decisão, de modo que seja compreensível e transparente para humanos.

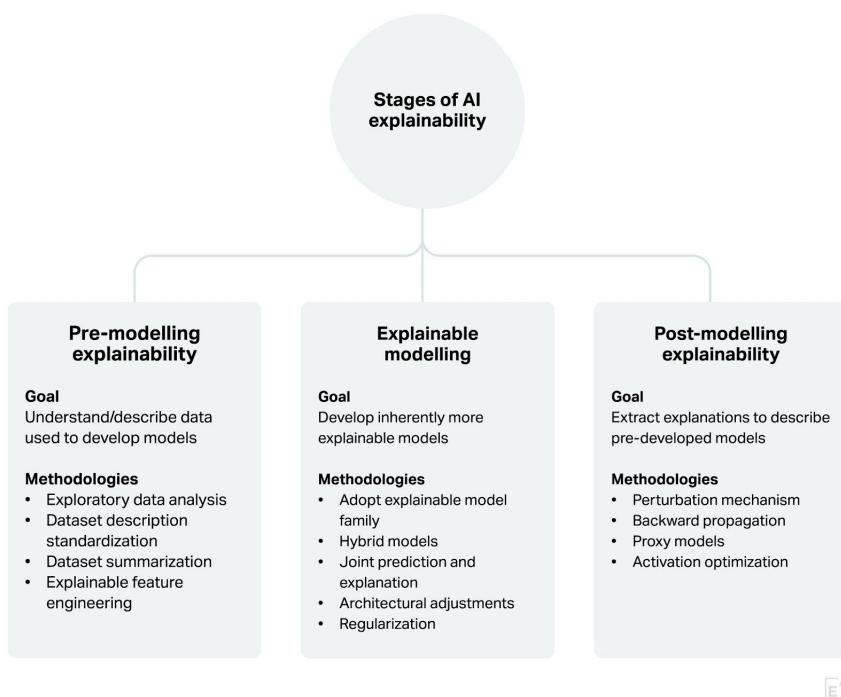


Figura A.1: Fases de um processo baseado em IA explicável.

Conforme ilustrado na Figura A.1, o um processo baseado em IA explicável pode ser classificado em 3 fases:

- ***Pre-Modelling Explainability:*** Esta fase se caracteriza por obter explicações no conjunto de dados usado para treino e validação, tendo como principal motivação o fato do comportamento de um modelo depender muito dos dados que o alimentam. Análises e visualizações dos dados, ou mesmo através de documentações do conjunto de dados se enquadram neste estágio. [52]
- ***Explainable Modelling:*** Esta fase se caracteriza por obter explicações por modelos considerados transparentes (ver adiante). Ao adotar modelos deste tipo, ele já pode ser considerado como explicável por um humano. Também é possível obter a explicação através de regularizadores, ou adotar abordagens onde explicação e resultado sejam obtidos simultaneamente, sejam via junção de dados ou sendo intrínseco a arquitetura do modelo. [50]
- ***Post-Modelling Explainability:*** Também chamada de *Post-hoc Explainability* e onde a maioria dos trabalhos disponíveis baseados em XAI tenta focar, esta fase se dedica a explicar modelos desenvolvidos anteriormente, ou modelos que não são considerados como transparentes. Pode-se explicar um modelo por métodos como árvores de decisão e estimativas como valores de Shapley e propagação inversa. [51]

Entretanto, estas 3 fases não necessariamente são executadas de maneira sequencial: Por ser realizada no conjunto de dados e não no modelo, a fase de *Pre-modelling Explainability* pode ser opcional, e a aplicação das fases de *Explainable Modelling* e *Post-Modelling Explainability* e seus métodos dependem exclusivamente se o modelo é classificado como transparente ou não, e tal critério é determinado pela adoção de uma (ou mais) das seguintes características [19]:

- **Simulabilidade:** Denota a capacidade de um modelo de ser simulado ou pensado estritamente por um humano. A complexidade assume um lugar dominante nesta classe: Sistemas com uma grande quantidade de regras, por mais simples que essas sejam, já não podem ser classificados como tal. O modelo que se adequa a essa característica precisa ser autocontido o suficiente para que um ser humano pense e raciocine sobre ele como um todo.
- **Decomponibilidade:** Também considerado como inteligibilidade, significa a capacidade de explicar cada uma das partes de um modelo. Se o modelo não for simples o suficiente, ele precisa ser divisível em várias pequenas partes e cada parte do modelo deve ser compreensível por um ser humano, sem a necessidade de ferramentas adicionais.
- **Transparência do algoritmo:** Trata-se da habilidade do usuário de entender o processo seguido pelo modelo para produzir qualquer saída dada a partir de seus dados de entrada. Colocando de outra forma, um modelo linear é considerado transparente porque sua superfície de erro pode ser entendida e fundamentada, permitindo ao usuário entender como o modelo irá agir em cada situação que pode enfrentar.

A.2 *Fluent API*

O conceito de *Fluent Interface* [40] (também conhecido como *Fluent API* [61]) é uma abstração de código que o organiza de modo a criar uma *Domain Specific Language* (DSL) interna, tendo como prioridade em seu *design* a legibilidade e a fluidez. Embora a construção de uma *Fluent API* consuma tempo, esse tempo pode ser recuperado com o tempo devido a maior facilidade no desenvolvimento e na manutenção dos componentes presentes no sistema.

Seu desenvolvimento lembra o *Design Pattern Builder*, onde seus métodos realizam ações específicas e retornam a referência do Objeto. Sua diferença está no objetivo em que foi desenvolvido: Enquanto no *Builder* o objetivo está na construção da instância de um Objeto, na *Fluent API* o objetivo está em desenvolver uma série de ferramentas e ações de modo a resolver um problema específico.

A melhor forma de descrever e entender as diferenças e objetivos da *Fluent API* é através de um exemplo: O Código A.1 representa uma implementação padrão de um sistema, com instanciações de novos Objetos e métodos determinando atribuições e adição do elemento em uma lista.

```

1  private void makeNormal(Customer customer) {
2      Order o1 = new Order();
3      customer.addOrder(o1);
4      OrderLine line1 = new OrderLine(6, Product.find("TAL"));
5      o1.addLine(line1);
6      OrderLine line2 = new OrderLine(5, Product.find("HPK"));
7      o1.addLine(line2);
8      OrderLine line3 = new OrderLine(3, Product.find("LGV"));
9      o1.addLine(line3);
10     line2.setSkippable(true);
11     o1.setRush(true);
12 }
```

Código A.1: Implementação padrão presente em um sistema [40]

O Código A.2 representa o mesmo sistema usando uma implementação com *Fluent API*. Além na diminuição no número de linhas, a implementação é mais legível devido ao menor número de elementos presentes no código: Todas as chamadas e instanciações do Código A.1 estão ímplicitas, e o encadeamento das funções possui uma progressão lógica, mais sucinta e de simples entendimento.

```

1  private void makeFluent(Customer customer) {
2      customer.newOrder()
3          .with(6, "TAL")
4          .with(5, "HPK").skippable()
5          .with(3, "LGV")
6          .priorityRush();
7 }
```

Código A.2: Implementação com o uso de *Fluent API* [40]

A.3 Feature Toggles

Feature toggles é um conceito antigo e conceitualmente simples: Basicamente, é uma variável usada em uma instrução condicional para proteger blocos de código, com o objetivo de habilitar ou desabilitar o código de uma feature nesses blocos para teste ou liberação [65]. Inicialmente elas foram pensadas para serem colocadas em tempo de compilação, excluindo a execução das features no binário de um aplicativo. Atualmente, é possível implementar *feature toggles* que permitem que as *features* sejam ativadas ou desativadas em tempo de execução, geralmente através de um arquivo ou através de uma variável introduzida no Sistema Operacional do *software* executado. O Código A.3 mostra um exemplo de *feature toggle*, onde a escolha dinâmica de um algoritmo de pesquisa depende do valor da *toggle* `useNewAlgorithm`. Se o valor dessa *toggle* for `true`, o novo algoritmo de pesquisa será usado, caso contrário, o método `Search` chamará o algoritmo de pesquisa antigo.

```

1 function Search() {
2     var useNewAlgorithm = false;
3     if(useNewAlgorithm){
4         return newSearchAlgorithm();
5     }else{
6         return oldSearchAlgorithm();
7     }
8 }
```

Código A.3: Implementação de uma *Feature Toggle* [54]

O uso de *feature toggles* é uma técnica frequentemente usada em contextos de integração contínua (CI) e entrega contínua (CD) que permite às equipes integrar e testar uma nova *feature* de forma incremental, mesmo quando ela não está pronta para ser lançada [54]. Os desenvolvedores também usam *feature toggles* para outros fins, como implantação gradual e experimentos. No entanto, a *feature toggle* pode se transformar em débito técnico. O uso de *feature toggles* adiciona mais pontos de decisão ao código, o que adiciona mais complexidade. Essa maior complexidade leva à necessidade de remover *toggles* quando a *feature* estiver concluída, ou olhar com mais atenção para os testes do sistema a ser implementado com a técnica.

Embora agilize testes ou coleta de dados importantes para as empresas, o uso de *feature toggles* sem seguir boas práticas pode ser prejudicial, levando a grandes prejuízos: Em 2012, os desenvolvedores do Knight Capital Group, uma empresa americana de serviços financeiros globais, atualizaram seu roteador algorítmico automatizado de alta velocidade que, inadvertidamente, reprovou um *feature toggle*, ativando a funcionalidade que não era utilizada há 8 anos. Em 2 minutos, os desenvolvedores perceberam que o código implantado se comportou incorretamente, mas levaram 45 minutos para interromper o sistema. Durante esse período, a Knight Capital perdeu quase 400 milhões de dólares, o que fez com que o grupo falisse [54].

Por ser uma técnica simples de ser implementada, as linguagens de programação fornecem o necessário para implementar *feature toggles* há muito tempo. No entanto, o primeiro uso desta técnica para suportar CI/CD foi no Flickr em 2009 [54]. Atualmente, empresas como Google, Facebook e Netflix utilizam esta técnica, podendo auxiliar na redução do

tempo de atualização de seus aplicativos para poucas semanas ou até diariamente [65]. Nos *softwares*, as *feature toggles* podem ser categorizadas em cinco tipos [54]:

- **Toggles de lançamento:** *Toggles* usadas para adicionar novas features em um contexto de *trunk-based development*. No *trunk-based development*, todos os desenvolvedores fazem o *commit* das alterações para uma *branch* compartilhada. Usando *toggles* de lançamento no desenvolvimento baseado em tronco suporta CI/CD para features parcialmente concluídas.
- **Toggles de experimento:** *Toggles* usadas para realizar experimentação no *software*, para avaliar novas alterações de recursos e observar sua influência no comportamento do usuário.
- **Toggles de operação:** *Toggles* usadas para controlar o aspecto operacional do comportamento do sistema. Quando uma nova feature é implantada, os operadores do sistema podem desabilitar o recurso rapidamente se ela apresentar alguma inconformidade.
- **Toggles de permissão:** *Toggles* usadas para fornecer a funcionalidade apropriada para um usuário, por exemplo *features* especiais para usuários *premium* ou pagos. *Toggles* de permissão também são chamados de *toggles* de negócios de longo prazo.
- **Toggles de desenvolvimento:** *Toggles* usados para habilitar ou desabilitar certos recursos para testar e depurar código.

Toggles de permissão, *toggles* de operação e *toggles* de desenvolvimento são *toggles* de longa duração com base em sua finalidade de uso no código. As *toggles* de lançamento e *toggles* de experimento são *toggles* de curta duração [54].

A.4 MLOps

Blablabla