MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

ROSANA LEANDRO DE OLIVEIRA

EXPLICABILIDADE DO MODELO DE APRENDIZADO DE MÁQUINA APOIADA PELA EXPLICABILIDADE DE DADOS: UMA ABORDAGEM BASEADA EM PROVENIÊNCIA

RIO DE JANEIRO
2023

ROSANA LEANDRO DE OLIVEIRA

EXPLICABILIDADE DO MODELO DE APRENDIZADO DE MÁQUINA APOIADA PELA EXPLICABILIDADE DE DADOS: UMA ABORDAGEM BASEADA EM PROVENIÊNCIA

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Kelli de Faria Cordeiro, D.Sc
Julio Cesar Duarte, D.Sc

Rio de Janeiro

2023

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 – Praia Vermelha
Rio de Janeiro – RJ CEP: 22290-270

Oliveira, Rosana Leandro de.
Explicabilidade do modelo de aprendizado de máquina apoiada pela explicabilidade de dados: Uma abordagem baseada em proveniência / Rosana Leandro de Oliveira. – Rio de Janeiro, 2023.
110 f.

Orientador(es): Kelli de Faria Cordeiro e Julio Cesar Duarte.

Dissertação (mestrado) – Instituto Militar de Engenharia, Sistemas e Computação, 2023.

1. Proveniência; PROV-DM; IA Explicável; Pré-Processamento; IA. i. Cordeiro, Kelli de Faria (orient.) ii. Duarte, Julio Cesar (orient.) iii. Título

ROSANA LEANDRO DE OLIVEIRA

Explicabilidade do modelo de aprendizado de máquina apoiada pela explicabilidade de dados: Uma abordagem baseada em proveniência

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Kelli de Faria Cordeiro e Julio Cesar Duarte.

Aprovado em Rio de Janeiro, 07 de julho 2023, pela seguinte banca examinadora:

Prof. **Julio Cesar Duarte -** D.Sc. do IME - Presidente

Profª. **Kelli de Faria Cordeiro** - D.Sc. do IME

Profª. **Maria Cláudia Reis Cavalcanti** - D.Sc. do IME

Prof. **João Luis Rebelo Moreira** - Ph.D. da University of Twente

Rio de Janeiro
2023

Este trabalho é dedicado a Deus, por todo o suporte e acolhimento, e ao meu amado filho e esposo, pela infinita compreensão e apoio durante minha ausência, que tornaram possível a dedicação para a realização desta pesquisa.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me manter saudável e fortalecer-me diante de todos os desafios enfrentados durante este período. Ao meu esposo Luiz, por compreender a minha ausência em muitos momentos para que eu pudesse me dedicar à conclusão do curso e pelo apoio recebido. Ao meu filho Lucas, com quem pude estar mais presente fisicamente durante esse tempo, e que sempre me animou com carinho e palavras de incentivo.

Agradeço à minha saudosa mãe, que sempre me motivou a ir além. Sua memória e seu apoio incondicional continuam vivos em meu coração. Ao meu querido pai, que sempre se dedicou à nossa família e cujo suporte foi fundamental. Vocês foram a minha base para chegar até aqui.

Agradeço imensamente à Marinha do Brasil pela escolha e pela oportunidade de dedicação para a conclusão desse curso. Aos professores do IME, pelo valioso conhecimento transmitido e pelo tratamento sempre cordial com o qual fui recebida por todos. Foram fundamentais para o meu crescimento acadêmico. Um agradecimento especial aos meus orientadores, Cel Júlio Cesar Duarte e CF(T) Kelli de Faria Cordeiro, pelo suporte inestimável, sabedoria e orientação que foram cruciais para a conclusão do meu trabalho de mestrado. Sou grata pela disponibilidade, paciência e dedicação em me auxiliar na construção do meu conhecimento.

Agradeço também à professora Maria Claudia Reis Cavalcanti “Yoko“ e ao professor João Luis Rebelo Moreira, membros da banca examinadora, pelos comentários, observações e sugestões que contribuíram e enriqueceram este trabalho.

Por fim, agradeço a todos os amigos e colegas que, de alguma forma, contribuíram para minha jornada de mestrado. Meu sincero agradecimento, em especial ao amigo CC(T) Leonardo Ferreira, que me acompanhou todas as fases do mestrado, sempre atencioso e solícito, alertando-me sobre prazos e datas e com palavras de coragem e incentivo.

"Busquem, pois, em primeiro lugar o Reino de Deus e a sua justiça, e todas essas coisas lhes serão acrescentadas.

(Bíblia Sagrada, Mateus 6:33)

RESUMO

As soluções de Inteligência Artificial, especialmente aquelas relacionadas ao Aprendizado de Máquina (AM), têm alcançado níveis notáveis de desempenho devido ao contínuo avanço da capacidade computacional, à disponibilidade abundante de dados e à evolução dos métodos de aprendizado. Em consequência, os modelos de AM se tornaram cada cada vez mais complexos e sofisticados, comprometendo a compreensão humana sobre os resultados alcançados. A fim de aumentar a interpretabilidade dos modelos de AM surgiu a IA Explicável, do inglês *Explainable AI (XAI)*. A XAI é de fundamental importância para aumentar a confiança nas predições de AM, e tornou-se de uso crucial para interpretação, principalmente nos modelos preditivos em áreas críticas. Para proporcionar um melhor entendimento sobre os dados, a proveniência dos dados oferece uma explicação sobre sua origem e sua derivação. Algumas pesquisas já exploram a utilização de informações sobre a proveniência dos dados em diversas fases do ciclo de AM para contribuir com a explicabilidade, no entanto, existe ainda uma lacuna na relação entre os dados de proveniência e a explicabilidade do modelo fornecida pelas técnicas de Inteligência Artificial Explicável (XAI). Com o intuito de solucionar essa questão, este estudo propõe a *Explainable Machine Learning Model supported by Pre-processing Provenance* (xMML-PPP), uma abordagem para capturar os dados de proveniência, especialmente durante a fase de pré-processamento, e relacioná-los com os resultados das técnicas de explicabilidade. Para isso, também foi proposto um modelo de dados relacional que serve como base para o nosso conceito de explicabilidade de dados. O principal objetivo é aumentar a explicabilidade dessas técnicas, complementando-as com informações provenientes da fase de pré-processamento. Para aplicação da abordagem, foi desenvolvida uma ferramenta *xMML-PPP Tool*, onde diversas informações do ciclo são capturadas, inclusive da fase de pré-processamento, e armazenadas no *xMML-PPP Prov*, repositório utilizado pela ferramenta para armazenamento dos dados capturados, onde, por meio de consultas aos dados armazenados no *xMML-PPP Prov*, as informações são recuperadas. A abordagem foi avaliada por meio de dois estudos de caso, nos quais foram realizados dois experimentos com configurações distintas para cada um dos estudos de caso. Isso viabilizou a análise do comportamento da explicabilidade em diferentes cenários. Os modelos foram treinados utilizando a *xMML-PPP Tool* com o algoritmo *Random Forest*, e o método de explicabilidade SHAP foi aplicado. Os resultados dos experimentos apresentaram que a melhoria na explicabilidade dos modelos de AM foi alcançada principalmente por meio da compreensão da derivação dos atributos que constituíram o modelo, enriquecida pela explicabilidade de dados.

Palavras-chave: Proveniência; PROV-DM; IA Explicável; Pré-Processamento; IA.

ABSTRACT

Artificial Intelligence solutions, especially those related to Machine Learning (ML), have achieved remarkable levels of performance due to the continuous advancement in computational capacity, the abundant availability of data, and the evolution of learning methods, which have become increasingly complex and sophisticated. To enhance the interpretability of ML models, Explainable AI (XAI) has emerged. XAI is of fundamental importance in increasing confidence in ML predictions and has become crucial for interpretation, especially in predictive models in critical domains. To provide a better understanding of the data, data provenance offers an explanation of its origin and derivation. Some studies have already explored the use of data provenance information in various stages of the ML lifecycle to contribute to explainability. However, there is a gap in the relationship between data provenance and the model's explainability provided by Explainable Artificial Intelligence (XAI) techniques. In order to address this issue, this study proposes *Explainable Machine Learning Model supported by Pre-processing Provenance* (xMML-PPP), an approach to capture provenance data, especially during the pre-processing phase, and relate it to the results of explainability techniques. To achieve this, a relational data model has also been proposed, which serves as the foundation for our data explainability concept. The main objective is to enhance the explainability of these techniques by complementing them with information derived from the pre-processing phase. For the application of this approach, a tool called *xMML-PPP Tool* has been developed, where various cycle information, including that from the pre-processing phase, is captured and stored in *xMML-PPP Prov*, the repository used by the tool to store the captured data, where, through queries to the data stored in *xMML-PPP Prov*, the information is retrieved. The approach was evaluated through two case studies, in which two experiments with different configurations were conducted for each of the case studies. This enabled the analysis of the interpretability behavior in different scenarios. The models were trained using the *xMML-PPP Tool* with the Random Forest algorithm, and the SHAP interpretability method was applied. The results of the experiments presented that the improvement in the explainability of ML models was mainly achieved through understanding the derivation of the attributes that constituted the model, enriched by data explainability.

Keywords: Provenance; PROV-DM; Explainable AI; Preprocessing; AI.

LISTA DE ILUSTRAÇÕES

Figura 1 – Proveniência Prospectiva em comparação a Retrospectiva 23
Figura 2 – PROV-DM — Principais estruturas 25
Figura 3 – Termos das principais tarefas de pré-processamento 27
Figura 4 – Categorias de Aprendizado de máquina 29
Figura 5 – Conceito XAI. 32
Figura 6 – Exemplo de método XAI local — técnica SHAP. 35
Figura 7 – Arquitetura do DPDS 41
Figura 8 – Arquitetura *Assistant* PP 41
Figura 9 – Arquitetura VAMSA 42
Figura 10 – Arquitetura USRPRUNG 43
Figura 11 – Visão do design do ExplAIner 44
Figura 12 – Ciclo de VIDA CSE. 45
Figura 13 – Arquitetura PROV-IO. 46
Figura 14 – Design Ascendente da API em camadas EdnaML. 47
Figura 15 – Platataforma ProML. 48
Figura 16 – Modelo conceitual da Ontologia proposta 49
Figura 17 – Arquitetura da xMML-PPP 53
Figura 18 – Macroprocesso xMML-PPP 54
Figura 19 – Processo Pré-processamento de dados 56
Figura 20 – Processo - Treinamento e Avaliação de dados 57
Figura 21 – Processo - Explicação do Modelo 57
Figura 22 – Processo - Explicação do Modelo e dos Dados 58
Figura 23 – Processo - Captura de Proveniência de Dados 59
Figura 24 – Ciclo de vida de aprendizado de máquina com fase de explicabilidade de modelo e de dados 59
Figura 25 – Diagrama conceitual abordagem xMML-PPP 61
Figura 26 – Elementos conceituais xMML-PPP (Workflow, Operadores e Dataset) e as Principais estruturas Prov - W3C 62
Figura 27 – Elementos conceituais xMML-PPP (Workflow, XAI e Experimento) e as Principais estruturas Prov - W3C 62
Figura 28 – Componentes arquiteturais da xMML-PPP 63
Figura 29 – Tela inicial da ferramenta “xMML-PPP Tool” 67
Figura 30 – Modelo de dados abordagem xMML-PPP 68
Figura 31 – Gráfico SHAP do experimento 1 - Titanic 76
Figura 32 – Gráfico SHAP do experimento 2 - Titanic 77
Figura 33 – Informações de origem do atributo Groupsize 77

Figura 34 – Informações de pré-processamento do atributo Groupsize 78
Figura 35 – Informações de valor de contribuição do atributo Groupsize 78
Figura 36 – Exemplo de visualização gráfica da proveniência do atributo *Groupsize* através do gráfico SHAP 79
Figura 37 – Exemplo de visualização gráfica da proveniência do atributo Family_Size através do gráfico SHAP 79
Figura 38 – Gráfico SHAP do experimento 1 - Covid19 - México 80
Figura 39 – Gráfico SHAP do experimento 2 - Covid19 - México 81
Figura 40 – Consulta Q1 - Atributos que derivaram o conjunto de treinamento . . . 82
Figura 41 – Consulta Q2 - Atributos construídos 82
Figura 42 – Consulta Q2A - Derivação de atributos construídos 83
Figura 43 – Consulta Q3 - Atributos com maior contribuição no modelo 84
Figura 44 – Consulta Q4A - Medidas de avaliação do Modelo 84
Figura 45 – Consulta Q4B- Parâmetros utilizados no Modelo 85
Figura 46 – Consulta Q5 - Descrição de atributos específicos 85
Figura 47 – Consulta Q6 - Operações de pré-processamento dos atributos de maior contribuição 86
Figura 48 – Gráfico SHAP - Primeiro experimento 105
Figura 49 – Gráfico SHAP - Terceiro experimento 106
Figura 50 – Gráfico SHAP - Quarto Experimento 108
Figura 51 – Gráfico SHAP - Sexto experimento 109
Figura 52 – Informações de origem do atributo GroupSurvived - o atributo alvo (Survived) é utilizado na sua construção 110

LISTA DE TABELAS

Tabela 1 – Tabela de Métodos XAI 36
Tabela 2 – Operações em pipelines de AM de preparação de dados do Orange e do Scikit-Learn 40
Tabela 3 – Sumário de Trabalhos Relacionados 50
Tabela 4 – Funcionalidades da ferramenta 67
Tabela 5 – Resultados de Experimentos - Base do Titanic 72
Tabela 6 – Relatório de Classificação - Base do Titanic - Experimento 1 73
Tabela 7 – Relatório de Classificação - Base do Titanic - Experimento 2 73
Tabela 8 – Resultados de Experimentos - Base COVID-19 México 74
Tabela 9 – Relatório de Classificação - Base COVID-19 México - Experimento 1 . 75
Tabela 10 – Relatório de Classificação - Base COVID-19 México - Experimento 2 . 75
Tabela 11 – Questões de consultas de proveniência 81
Tabela 12 – Titanic Dataset 101
Tabela 13 – Descrição dos atributos COVID-19 - México 101
Tabela 14 – Relatório de Classificação - Base do Titanic - Primeiro experimento . . 104
Tabela 15 – Relatório de Classificação - Base do Titanic - Terceiro experimento . . 106
Tabela 16 – Relatório de Classificação - Base do Titanic - Quarto experimento . . . 107
Tabela 17 – Relatório de Classificação - Base do Titanic - Sexto experimento 109

LISTA DE ABREVIATURAS E SIGLAS

AM Aprendizado de Máquina

BPMN Business Process Model and Notation

IA Inteligência Artificial

ICE Individual Conditional Expectation

KDD Knowledge Discovery and Data Mining Process

LIME Local Interpretable Model-Agnostic Explanations

PDP Partial Dependence Plot

PROV-DM Provenance Data Model

RF Random Forest

SGBDR Sistema Gerenciador de Banco de Dados Relacional

SHAP Shapley Additive exPlanations

SVM Máquina de Vetor de Suporte

SQL Structured Query Language

W3C World Wide Web Consortium

XAI IA Explicável

SUMÁRIO

1 INTRODUÇÃO 15
1.1 MOTIVAÇÃO 16
1.2 CARACTERIZAÇÃO DO PROBLEMA 17
1.3 OBJETIVO 18
1.4 JUSTIFICATIVA 19
1.5 METODOLOGIA 19
1.6 ORGANIZAÇÃO DA DISSERTAÇÃO 20

2 FUNDAMENTAÇÃO TEÓRICA 21
2.1 PROVENIÊNCIA DE DADOS 21
2.1.1 FORMAS DE PROVENIÊNCIA 21
2.1.2 CAPTURA, MODELO E INFRAESTRUTURA DE PROVENIÊNCIA 23
2.1.2.1 *PROVENANCE DATA MODEL* 24
2.2 PRÉ-PROCESSAMENTO DE DADOS 26
2.3 APRENDIZADO DE MÁQUINA 28
2.3.1 ALGORITMOS DE APRENDIZADO DE MÁQUINA 29
2.4 IA EXPLICÁVEL - XAI 31
2.4.1 MÉTODOS XAI 33
2.4.2 CLASSIFICAÇÃO DOS MÉTODOS XAI 36

3 REVISÃO DA LITERATURA 38
3.1 *CAPTURING AND QUERYING FINE-GRAINED PROVENANCE OF PRE-PROCESSING* 39
3.2 *DATA PROVENANCE FOR DATA SCIENCE (DPDS)* 40
3.3 ASSISTANT PP 41
3.4 VAMSA 42
3.5 USRPRUNG 42
3.6 EXPLAINER 43
3.7 PROV-ML 44
3.8 PROV-IO 45
3.9 EDNAML 46
3.10 PROML 47
3.11 SEMANTIC DESCRIPTION OF EXPLAINABLE MACHINE LEARNING WORKFLOWS FOR IMPROVING TRUST 48
3.12 COMPARATIVOS DOS TRABALHOS 50

4 ABORDAGEM XMML-PPP 52
4.1 EXPLICABILIDADE DE DADOS E DE MODELO 52
4.2 METODOLOGIA DE CONCEPÇÃO 53
4.3 PROCESSO DA XMML-PPP 54
4.4 MODELO CONCEITUAL 58

5 IMPLEMENTAÇÃO DA XMML-PPP 63
5.1 FERRAMENTA "XMML-PPP TOOL" 66
5.1.1 FUNCIONALIDADES DA FERRAMENTA "XMML-PPP TOOL" 66
5.2 REPOSITÓRIO DE PROVENIÊNCIA "XMML-PPP PROV" 67

6 APLICAÇÃO DA XMML-PPP 70
6.1 EXPERIMENTOS COM A BASE DE DADOS TITANIC 71
6.2 EXPERIMENTOS - BASE DE DADOS COVID-19 MÉXICO 73
6.3 EXPLICABILIDADE DO RESULTADO DOS EXPERIMENTOS 75
6.3.1 EXPLICABILIDADE - BASE DE DADOS DO TITANIC 76
6.3.2 EXPLICABILIDADE - BASE DE DADOS DO COVID-19 80

7 CONCLUSÃO 88

REFERÊNCIAS 92

APÊNDICE A – DESCRIÇÃO DAS TABELAS 98

APÊNDICE B – DICIONÁRIO DE DADOS 101

APÊNDICE C – A INFLUÊNCIA DOS ATRIBUTOS NA EXPLICABILIDADE E NO RESULTADO DO MODELO 104

1 INTRODUÇÃO

As soluções de Inteligência Artificial (IA) permeiam amplamente a sociedade contemporânea, sendo implícitas em uma variedade de setores e atividades cotidianas. Essa forma avançada de automação e processamento de dados encontra-se entrelaçada em domínios tão diversos como publicidade, recomendações de produtos, avaliação de crédito, diagnósticos médicos, entre outros.

É notável como essas aplicações se inseriram na sociedade moderna, estabelecendo-se como instrumentos primordiais de influência e auxílio nos mais variados contextos. Na esfera publicitária, por exemplo, essas soluções são habilmente empregadas para otimizar a eficácia de campanhas, direcionando-as de forma personalizada e assertiva ao público-alvo (1). Também, são vistas no campo do comércio eletrônico, onde assumem um papel crucial na formulação de recomendações personalizadas, facilitando a experiência do consumidor e aumentando as chances de satisfação (2). Adicionalmente, a IA é amplamente utilizada no setor financeiro, especialmente na análise de crédito. Nesse contexto, algoritmos inteligentes têm a capacidade de processar e avaliar com precisão o perfil de cada pessoa, permitindo determinar sua capacidade de pagamento (3). Essa automação de processos proporciona maior agilidade e eficiência na tomada de decisões, reduzindo erros e minimizando riscos.

Não obstante, na área da saúde, a IA já desempenha um papel relevante nos diagnósticos médicos, atuando como uma poderosa ferramenta de apoio aos profissionais de saúde. Por meio de algoritmos avançados de Aprendizado de Máquina (AM), os sistemas de IA podem analisar grandes volumes de dados clínicos e imagens médicas, identificando padrões e fornecendo informações cruciais para auxiliar no diagnóstico e tratamento de doenças (4).

Indubitavelmente, o aumento dessas soluções de IA, propiciada pela capacidade dos algoritmos apresentarem resultados cada vez mais satisfatórios, agrega elevado valor na sociedade moderna. No entanto, torna-se imperativo destacar que, para aplicações críticas, como no exemplo de previsões na área de saúde, os resultados gerados por modelos de AM, que podem ter grande impacto na tomada de decisão, devem ser acompanhados por uma justificativa (5). Nesse contexto, é crucial enfatizar a importância de compreender os resultados das previsões e a confiabilidade dos mesmos.

Contudo, muitos modelos de AM existentes são insuficientes para fornecer explicações claras sobre como e por que os resultados foram obtidos. Isso tem gerado preocupações crescentes de que esses modelos possam ser injustos, opacos ou não intuitivos (6). Como resultado, a crescente opacidade dos sistemas de IA e o possível impacto nos usuários finais destacam a necessidade da explicabilidade nesses sistemas, resultando em um campo

de pesquisa atualmente conhecido como Explicabilidade, ou IA Explicável (XAI) (7). O principal objetivo dessa área de estudo é tornar as previsões de modelos complexos (também conhecidos como "caixa-preta") mais compreensíveis. Isso significa oferecer maior transparência nos resultados dos algoritmos. Ao alcançar essa transparência, a confiança nos sistemas que utilizam IA é fortalecida, pois os usuários conseguem entender melhor como as decisões são tomadas e quais fatores influenciam nelas. Nesse sentido, em um sistema de diagnóstico de doenças, poderiam ser fornecidos os fatores que foram mais relevantes para um sistema decidir em relação a um paciente ser diagnosticado com determinada patologia.

Adicionalmente, outro fator importante, que implica no resultado dos modelos de AM, são os dados utilizados nos seus treinamentos. Cumpre destacar que esses dados podem conter vieses, que podem levar a conclusões errôneas no modelo. Dessa forma, os dados de entrada em um modelo de AM precisam ser analisados. Além disso, devem ser tratados para que o resultado do modelo seja ainda mais satisfatório. Nesse contexto, a realização de operações de pré-processamento tem como objetivo contribuir para o tratamento e melhoria da qualidade dos dados analisados. O tratamento dos dados, pelas operações de pré-processamento, produz um impacto no desempenho e decisão dos modelos de IA (8). Nesse sentido, a captura das operações realizadas na fase de pré-processamento pode servir para compreender o tratamento dos dados e entender a contribuição de cada operação no resultado final do modelo.

Dessa forma, a XAI desempenha um papel crucial ao fornecer métodos e técnicas para melhor compreensão de como os modelos de AM tomam decisões, ao mesmo tempo em que é pautada a necessidade de olhar para os dados de proveniência como um componente que contribui para a explicabilidade. Dessa maneira, alguns autores têm se manifestado a favor da utilização da proveniência como meio de fornecer uma explicação adicional (8, 9).

A proveniência dos dados oferece uma explicação detalhada da origem dos dados e do processo de sua derivação. A coleta da proveniência dos dados tem o potencial de auxiliar na interpretação dos resultados, proporcionando um melhor entendimento sobre os dados utilizados, como os dados de entrada de um modelo de IA, assim como o conhecimento das etapas de pré-processamento que influenciaram o resultado final. Nesse sentido, a união das áreas do conhecimento de proveniência e XAI, exploradas em conjunto pode fornecer um diferencial na complementação da explicabilidade em AM.

1.1 Motivação

Com o crescimento da utilização de sistemas opacos de IA surgiu a necessidade de prover maior interpretabilidade para os resultados desses sistemas. No entanto, alguns autores contestam a capacidade das técnicas de XAI fornecerem explicabilidade por si

só. No estudo conduzido por Jaigirdar et al.(10), por exemplo, os autores ressaltam a importância da explicação em sistemas baseados em IA, uma vez que é necessário atender aos requisitos de qualidade e operar de maneira transparente, abrangente, compreensível e explicável. Destacam-se assim algumas questões em aberto nessa área, fornecendo uma visão mais abrangente da explicabilidade em sistemas de IA, com o intuito de facilitar a tomada de decisão. Além disso, o trabalho ilustra a potencial contribuição dos dados de proveniência, os quais são adequados para sistemas baseados em IA ao discutir o papel fundamental da implementação de gráficos de proveniência na explicação das propriedades da IA. Igualmente, no trabalho de Scherzinger, Seifert e Wiese(9), são abordados os impactos das transformações dos dados no modelo aprendido em um ambiente distribuído, destacando os potenciais desafios de se vincular as contribuições de pesquisas sobre proveniência de dados com os esforços de explicabilidade em AM.

A XAI é uma ferramenta essencial para compreender o processo pelo qual um modelo de AM chega a um resultado específico. Tal ferramenta utiliza uma variedade de técnicas com enfoques diferentes para identificar quais atributos foram relevantes para uma determinada decisão tomada pelo modelo. No entanto, é de suma importância destacar que, caso não sejam adequadamente registradas as operações realizadas nos dados antes da geração do modelo, as informações referentes a essas operações não serão conhecidas, tornando-se impossível obter conhecimento sobre o pré-processamento desses dados e compreender como essas operações podem ter influenciado o resultado do modelo.

Vários autores (7, 9, 10, 11) já levantaram a questão de que a proveniência pode ser usada para aumentar a transparência e a explicabilidade em sistemas baseados em IA. No entanto, não foram encontradas abordagens que direcionassem pesquisas no sentido de verificar a aplicação da proveniência para complementar a explicabilidade das técnicas XAI.

Atualmente, por ainda haver várias questões em aberto, a proveniência que visa contribuir com a explicabilidade do aprendizado em AM ainda carece de bastante atenção. Dessa forma, a motivação desse trabalho é contribuir com a explicabilidade dos modelos de AM, ao permitir visualizar a relação dos resultados das técnicas de explicabilidade com as operações de pré-processamento realizadas nos dados. Para isso, uma análise dos dados de proveniência da fase de pré-processamento que resultou no modelo será realizada.

1.2 Caracterização do Problema

O aumento da utilização da IA em sistemas críticos trouxe a necessidade de entendimento dos resultados dos modelos utilizados nesses sistemas. Assim, houve o crescimento do estudo da área de XAI nos últimos anos. As técnicas XAI tem por objetivo principal resolver o problema da transparência e da falta de compreensão dos modelos de

IA. Contudo, há pouca exploração ainda sobre como os dados que derivaram um modelo foram tratados, representando um problema em relação à transparência dos dados que derivaram o modelo.

Adicionalmente, a compreensão da proveniência das operações de pré-processamento nos dados que alimentam um modelo de IA é de grande relevância, uma vez que se sabe que tais etapas podem ter uma influência significativa nos resultados do modelo e, por consequência, na sua capacidade de explicação. Embora já seja reconhecida a importância de se entender a proveniência dos dados e o impacto do pré-processamento nos modelos de IA, ainda há uma lacuna no conhecimento atual de modo a comprovar que a proveniência, aliada a técnicas de explicabilidade, pode contribuir no problema da falta de transparência, desejável sobretudo para os sistemas de IA que tratam de assuntos críticos. Desse modo, são necessárias mais pesquisas e propostas visando aprofundar o entendimento dessa questão e entender como a proveniência pode contribuir para a explicabilidade, com vistas a atender essa lacuna existente.

1.3 Objetivo

O objetivo geral deste trabalho é desenvolver uma nova abordagem que contribua para a explicabilidade do AM, visando ampliar a abrangência da explicabilidade do modelo por meio da utilização dos dados de proveniência do fluxo de trabalho empregado na gestão de um modelo. Para atender a esse objetivo, a abordagem proposta combina o uso das técnicas de explicabilidade do modelo, fornecidas pela área de XAI, com a utilização da proveniência dos dados do AM, sobretudo na etapa de pré-processamento. Desse modo, esse objetivo geral pode ser decomposto nos seguintes objetivos específicos:

1. Identificar os dados de proveniência necessários que possam contribuir para complementar a explicabilidade Post-Hoc;

2. Capturar os metadados de informação da estrutura do fluxo de trabalho de uma operação de classificação de AM, bem como as operações de pré-processamento realizadas para transformação dos dados, adicionalmente, permitindo a captura das informações de treino e explicação do modelo com vistas a contribuir com as explicações; e

3. Realizar análises dos dados de proveniência capturados durante a execução do fluxo de trabalho, inclusive os dados relativos à explicabilidade do modelo, com vistas a alcançar o objetivo de agregar explicações das transformações realizadas nos dados.

1.4 Justificativa

Nos últimos anos, o interesse pela área de pesquisa de explicabilidade em IA cresceu bastante. O crescente interesse por essa área foi impulsionado por várias razões, incluindo a necessidade de se entender as decisões tomadas por seus algoritmos em diversos domínios críticos como, por exemplo, nas áreas militar e da saúde. À medida que a utilização dos benefícios proporcionados pela IA se tornou mais generalizada, a demanda por transparência e responsabilidade tem se intensificado significativamente. Nesse contexto, compreender de maneira aprofundada o funcionamento desses sistemas tornou-se um aspecto fundamental.

Conforme a Estratégia brasileira de Inteligência Artificial (EBIA), elaborada pelo Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC), a área de IA foi definida como prioridade, no que se refere a projetos de pesquisa, de desenvolvimento e inovações para o período de 2020 a 2023. Assim, de acordo com um dos seus objetivos estratégicos: "contribuir para a elaboração de princípios éticos para o desenvolvimento e uso de IA responsáveis", este trabalho visa a contribuir com pesquisas na área da IA transparente e ética.

Adicionalmente, o emprego da IA transparente pode ser de relevante emprego, com vistas a contribuir para a garantia da capacidade operacional das Forças Armadas, em alinhamento com a Estratégia Nacional de Defesa (END), e em atendimento aos Objetivos Estratégicos do Exército (OEE), sobretudo, os objetivos 7 e 9, "Aprimorar a gestão estratégica da informação" e "Aperfeiçoar o sistema de Ciência, tecnologia e Inovação", respectivamente.

Além disso, é possível estabelecer uma integração da tecnologia de IA em uma ampla gama de sistemas já adotados nas três Forças Armadas, com o objetivo primordial de aprimorar sua eficiência operacional. Adicionalmente, ao incorporar a proveniência dos dados e estabelecer mecanismos de rastreamento, é possível aumentar a confiança nos resultados obtidos, impulsionando a excelência e a transparência desses sistemas.

1.5 Metodologia

Após a delineação do escopo deste trabalho, foram realizadas pesquisas com vistas a encontrar trabalhos que tivessem abordado soluções de proveniência em aplicações de AM. Inicialmente, foram realizadas pesquisas nas principais bases de dados: Scopus, ACM, IEEE Xplore e ScienceDirect, até se chegar a um conjunto de palavras-chave que obtivesse os resultados mais adequados para responder às questões de pesquisa.

Consoante o resultado das pesquisas realizadas, dos trabalhos retornados pela busca, alguns utilizavam a proveniência para uma fase específica do ciclo de vida de

AM(12, 13, 14, 15), outros, para todo o ciclo (16). No entanto, não foi encontrado nenhum trabalho que capturasse a proveniência até a fase da explicabilidade do modelo. Dessa forma, foi estabelecida a necessidade de incluir a proveniência da fase de explicabilidade do modelo para poder complementar a explicabilidade do modelo com a proveniência dos dados de todo o ciclo de AM, sobretudo, a da fase de pré-processamento.

Com o intuito de suprir a necessidade identificada, deu-se início à concepção da abordagem. Primeiramente, foi estabelecida a compreensão do macroprocesso operacional necessário para alcançar o objetivo estipulado, e, em seguida, foi desenvolvido o modelo de dados para dar suporte à proveniência previamente definida. Posteriormente, foram criados componentes arquiteturais para a construção da ferramenta de captura de proveniência, conforme proposto.

Por fim, a abordagem proposta foi aplicada em dois conjuntos de dados, um com ênfase no melhor entendimento da capacidade da proposta e outro com uma aplicação contemporânea, visando satisfazer uma demanda real. A aplicação da abordagem foi conduzida com o intuito de verificar se a coleta dos dados do ciclo de vida de AM e as análises subsequentes conseguiriam fornecer explicações complementares, contribuindo assim com as explicações post-hoc geradas por técnicas XAI.

1.6 Organização da Dissertação

Esta dissertação está estruturada em 7 capítulos a fim de atingir os objetivos apresentados. Para isso, além desta introdução, o Capítulo 2 apresenta os fundamentos teóricos com conceitos básicos sobre os assuntos de Proveniência de Dados, Aprendizado de Máquina e IA Explicável.

No Capítulo 3 é realizada a Revisão da Literatura, onde são relacionados os principais trabalhos que utilizam a proveniência de dados para o contexto de AM. Já no Capítulo 4, a abordagem proposta neste trabalho é apresentada.

No Capítulo 5, os detalhes técnicos de implementação da abordagem, com a construção dos artefatos de apoio são abordados e, no Capítulo 6, são apresentados os casos de implementação dos experimentos para validação da abordagem.

Finalmente, o Capítulo 7 apresenta a conclusão e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os referenciais teóricos necessários para o entendimento da abordagem proposta. Além dos conceitos sobre Proveniência de dados (seção 2.1), serão apresentados também a teoria sobre aprendizado de máquina (seção 2.3) e Pré-Processamento de dados (seção 2.2), e por fim, conceitos sobre os métodos de explicabilidade (seção 2.4) dos Modelos de AM.

2.1 Proveniência de Dados

A proveniência de dados compreende todas as informações referentes à origem dos dados. Conforme definido em (17) é "um registro que identifica as pessoas, informações ou entidades e atividades envolvidas na produção, influência ou entrega de um dado ou coisa", ou seja, todas as informações sobre o processo de produção de determinado produto, seu histórico, desde o início de sua concepção.

Em essência, a proveniência pode ser vista como metadados que descrevem um processo de produção. Segundo Herschel, Diestelkämper e Lahmar(18), a captura e processamento da proveniência é importante para avaliar a qualidade, garantir a reprodutibilidade, e também para reforçar a confiança do artefato produzido. Assim, por meio da proveniência de dados é possível descobrir as possíveis falhas na concepção do objeto ou coisa, ou reproduzir a criação de novo objeto por meio do registro histórico capturado.

A captura e gerenciamento da proveniência para tarefas computacionais é relevante para um amplo domínio de aplicações. Em experimentos científicos, a proveniência auxilia na interpretação e na compreensão dos resultados, sendo possível examinar a sequência de etapas que contribuíram para um determinado resultado, verificar os dados de entrada e reproduzir o resultado (19).

2.1.1 Formas de Proveniência

Uma questão de relevante importância em proveniência é sobre a sua representação. Nos trabalhos de Davidson e Freire; Pérez, Rubio e Sáenz-Adán(20, 21), duas formas de proveniência podem ser definidas: a prospectiva e a retrospectiva. No entanto, Khan et al.(22) acrescentam um terceiro tipo de proveniência para fluxo de trabalho, a proveniência de evolução do fluxo de trabalho.

Proveniência Prospectiva

Davidson e Freire(20) definem a proveniência prospectiva como: "Corresponde aos passos que precisam ser seguidos (ou a uma receita) para gerar um produto de dados

ou uma classe de produtos de dados". Já no trabalho de Khan et al.(22), a proveniência prospectiva é definida como: " refere-se às receitas usadas para capturar um conjunto de tarefas computacionais e sua ordem, por exemplo, a especificação do fluxo de trabalho". Assim, com base nas definições dos trabalhos citados a proveniência prospectiva pode ser definida como o planejamento e a documentação antecipados dos procedimentos e etapas necessárias para criar um produto de dados ou uma categoria de produtos de dados.

Proveniência Retrospectiva

No trabalho de Davidson e Freire(20), a proveniência retrospectiva é definida como: "A proveniência retrospectiva captura os passos executados, bem como informações sobre o ambiente de execução utilizado para derivar um produto de dados específico". Já no trabalho de Khan et al.(22), tem-se a seguinte definição para a Proveniência retrospectiva:

> "refere-se ao registro detalhado da implementação de uma tarefa computacional, incluindo os detalhes de cada processo executado com informações abrangentes sobre o ambiente de execução usado para derivar um produto específico".

A proveniência retrospectiva também preserva informações sobre os recursos que são gerados durante a execução do fluxo de trabalho (18). Nesse sentido, com base nos trabalhos citados a proveniência retrospectiva pode ser definida como o registro detalhado das ações realizadas, juntamente com informações abrangentes sobre o contexto de execução empregado para gerar um produto de dados específico. Isso envolve a documentação completa da implementação de uma tarefa computacional, incluindo todos os processos executados, enquanto se mantém um registro completo das condições e do ambiente em que essas ações ocorreram.

Proveniência da Evolução do fluxo de trabalho

No trabalho de Khan et al.(22), a proveniência de evolução do fluxo de trabalho é definida por:

> "refere-se ao rastreamento de qualquer alteração no fluxo de trabalho existente, resultando em outra versão do fluxo de trabalho que pode produzir os mesmos ou diferentes artefatos de dados resultantes".

A proveniência de evolução facilita iterações rápidas em diferentes dados, parâmetros e modificações no fluxo de trabalho (18). Dessa forma, a proveniência da evolução do fluxo de trabalho permite entender como o processo evoluiu e como as decisões tomadas durante as mudanças afetaram os resultados dos dados gerados.

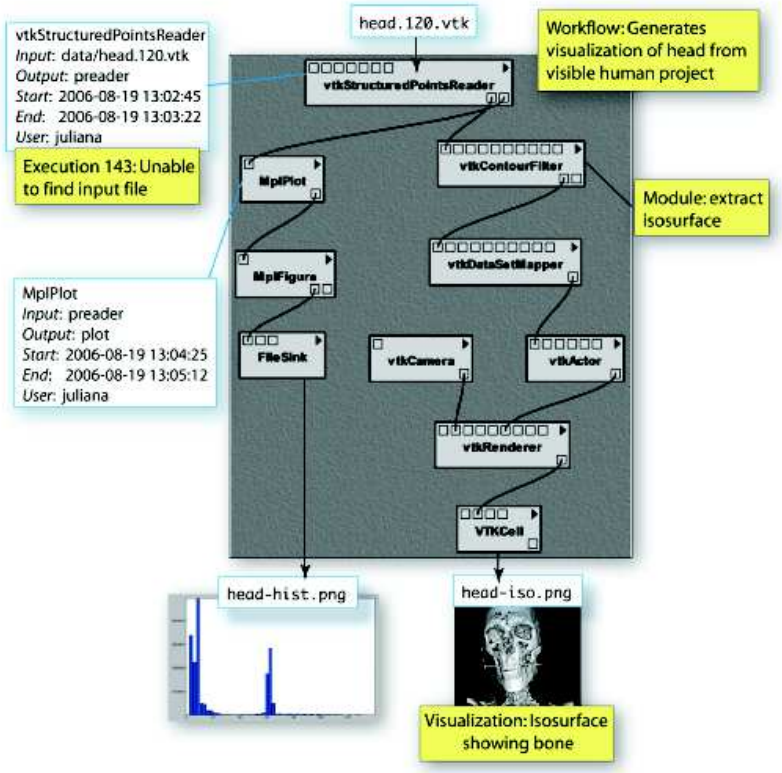

Figura 1 – Proveniência Prospectiva em comparação a Retrospectiva. Fonte: (20)

Na Figura 1, é apresentado um exemplo do fluxo de trabalho que gera dois produtos de dados: um histograma dos valores escalares de um conjunto de dados de grade estruturada; e uma visualização de uma isosuperfície do conjunto de dados. À esquerda, são exibidas algumas informações da proveniência retrospectiva coletadas durante uma execução deste fluxo de trabalho. A definição do fluxo de trabalho fornece a proveniência prospectiva que detalha a sequência (uma receita) para a obtenção desses dois tipos de produtos de dados.

2.1.2 Captura, Modelo e Infraestrutura de proveniência

De acordo com Freire et al.(19), uma solução de gerenciamento de proveniência é composta por três principais componentes: o mecanismo de captura, um modelo de representação e a infraestrutura para o acesso.

Mecanismo de Captura de Proveniência

O mecanismo de captura de proveniência é o componente responsável pela coleta de informações de proveniência do experimento, dividindo-se em três classes principais, baseadas no fluxo do trabalho, do processo e do sistema operacional (SO). Aqueles baseados em fluxo de trabalho são anexados ou integrados em um sistema de fluxo de trabalho; os mecanismos baseados em processos requerem que cada serviço ou processo envolvido em

uma tarefa computacional se documente; e os mecanismos baseados em sistema operacional não requerem modificações em *scripts* ou programas existentes; em vez disso, eles contam com a disponibilidade de funcionalidades específicas no nível do sistema operacional. Os sistemas de fluxo de trabalho podem capturar a proveniência prospectiva e retrospectiva, enquanto os mecanismos baseados no sistema operacional e no processo capturam apenas a proveniência retrospectiva.

Modelo de Representação de Proveniência

No modelo de representação de proveniência é definido como as informações de proveniência da abordagem são representadas ou modeladas. Embora os modelos sejam divergentes de várias maneiras, incluindo o uso de estruturas e estratégias de armazenamento, todos eles compartilham um tipo essencial de informação: processos e dependências de dados. Os modelos de proveniência tendem a variar conforme o domínio e as necessidades do usuário.

Infraestrutura de acesso

A infraestrutura de acesso define as questões diversas relacionadas a base da estrutura da abordagem, tais como estratégias de armazenamento, definição de formas de acesso, mecanismo de consultas aos dados de proveniência, dentre outras. Exemplos de estratégia de armazenamento são: arquivos XML ou tuplas armazenadas em tabelas de banco de dados relacionais.

2.1.2.1 *Provenance Data Model*

O Provenance Data Model (PROV-DM) é o modelo de representação de dados recomendado pela World Wide Web Consortium (W3C). O PROV-DM distingue as estruturas centrais, formando a essência das informações de proveniência, das estruturas estendidas que atendem a usos mais específicos de proveniência. O modelo de dados PROV-DM tem um design modular e está estruturado de acordo com seis componentes: (1) entidades e atividades, e a época em que foram criadas, usadas ou finalizadas; (2) derivações de entidades; (3) agentes responsáveis pelas entidades geradas e pelas atividades que aconteceram; (4) uma noção de pacote, ou seja, um mecanismo para apoiar a proveniência da proveniência; (5) propriedades para vincular entidades que se referem à mesma coisa; e (6) coleções formando uma estrutura lógica para os seus membros (17). Em sua essência, a proveniência descreve o uso e a produção de entidades por atividades, que podem ser influenciadas de várias maneiras pelos agentes. Esses tipos principais e seus relacionamentos são ilustrados pelo diagrama UML da figura 2.

Os conceitos do núcleo do PROV-DM (17) e os seus relacionamentos são resumidamente descritos a seguir.

- Entidade (*Entity*): é algo que se deseja descrever. Elas podem ser físicas, digitais ou

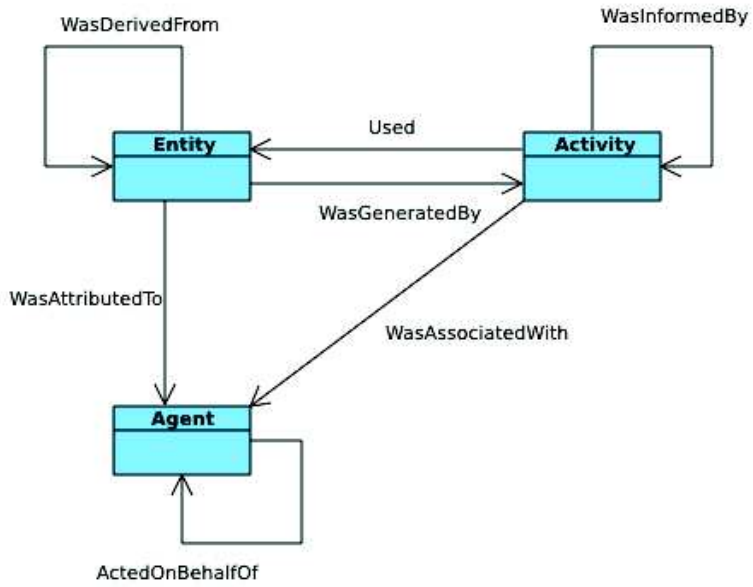

Figura 2 – PROV-DM — Principais estruturas. Fonte: (17)

conceituais. Podem ser reais ou imaginárias, como, por exemplo, registros, conjuntos de dados, papéis e modelos;

- Atividade (*Activity*): é algo que ocorre durante um período de tempo e age sobre ou com entidades, ou seja, as atividades utilizam as entidades e, também, podem produzir entidades. As atividades podem incluir o consumo, e processamento, e transformação, e modificação, e realocação, e uso ou a geração de entidades;
- Agente (*Agent*): é algo que tem alguma forma de responsabilidade por uma atividade que ocorre, pela existência de uma entidade ou pela atividade de outro agente. Um agente pode ser um tipo específico de entidade ou atividade;
- Geração (*WasGeneratedBy*): é a produção de uma nova entidade por uma atividade;
- Usado (*used*): é a utilização de uma entidade por uma atividade;
- Informação (*wasInformedBy*): é a geração de uma entidade por uma atividade e seu uso subsequente por outra atividade;
- Derivação (*wasDerivedFrom*): é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova. O foco da derivação é conectar uma entidade gerada a uma entidade usada;
- Atribuição (*wasAttributedTo*): é a atribuição de uma entidade a um agente;
- Associação (*wasAssociatedWith*): é uma atribuição de responsabilidade a um agente de uma atividade, indicando que o agente tem uma função na atividade; e

- Delegação (*actedOnBehalfOf*): é a atribuição de autoridade e responsabilidade de um agente, podendo ser para si ou para outro agente, com o intuito de realizar uma atividade específica como um delegado ou representante.

Em resumo, o PROV-DM é uma especificação da W3C que fornece um modelo padronizado para representar informações de proveniência de dados, possibilitando entender o histórico e as relações entre entidades, atividades e agentes envolvidos na criação e modificação dos dados. Assim, por ter sua recomendação desde 2013, tem sido utilizada por vários trabalhos como referência.

2.2 Pré-processamento de Dados

A fase de pré-processamento é iniciada tão logo os dados são coletados e organizados na forma de um conjunto de dados. Um dos objetivos da fase de pré-processamento é solucionar problemas nos dados e melhorar a qualidade destes. Operações para tratar dados corrompidos, dados ausentes, dados duplicados e remoção de atributos irrelevantes são exemplos de tarefas realizas nesta fase.

Os conjuntos de dados atuais, em virtude de seu grande volume, são usualmente formados de fontes heterogêneas, e altamente suscetíveis a ruídos, ausência de dados e inconsistências. Nesse sentido, as ações na fase de pré-processamento visam a preparar os dados para a fase posterior, a fase de extração do conhecimento. Dessa forma, seu objetivo é o de proporcionar um melhor desempenho ao algoritmo de AM, uma vez que a baixa qualidade dos dados poderá levar a predições de algoritmo de baixa qualidade.

As etapas do pré-processamento podem ser definidas como "as funções relacionadas com a captação, a organização, o tratamento e a preparação dos dados para a etapa da Mineração de Dados" (23). Ou seja, as técnicas de pré-processamento têm como objetivo primário a preparação dos dados para a subsequente etapa de descoberta de conhecimento.

Na Figura 3, disponibilizada no trabalho de Moura(24), são apresentados os principais termos das atividades de pré-processamento, conforme apresentados na literatura de Goldschmidt, Passos e Bezerra(23), nas literaturas de Gama et al.(25), García, Luengo e Herrera(26) e Han, Pei e Kamber(27), ficando evidenciado a diversidade de termos desse domínio.

A seguir apresentamos as operações de captação, organização e tratamento dos dados, segmentadas em oito atividades, conforme apresentado em Goldschmidt, Passos e Bezerra(23):

- **Seleção de Dados**: Nesta atividade é realizada a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante

Goldschmidt *et al.*, 2015	**Faceli** *et al.*, 2015	**Garcia** *et al.*, 2015	**Han** *et al.*, 2011
1- Seleção de dados	1- Integração de dados	1- Integração de dados	1- Integração de dados
2- Limpeza	2- Eliminação manual de atributos	2- Limpeza de dados	2- Limpeza de dados
3- Codificação	3- Amostragem de dados	3- Transformação de dados	3- Redução de dados
4- Enriquecimento de dados	4- Balanceamento de dados	4- Redução de dados	4- Transformação de dados
5- Normalização de dados	5- Limpeza de dados	5- Seleção de características	
6- Construção de atributos	6- Transformação de atributos		
7- Correção de prevalência	7- Redução da dimensionalidade		
8- Partição do conjunto de dados			

Figura 3 – Termos das principais tarefas de pré-processamento. Fonte: (24)

o processo de Knowledge Discovery and Data Mining Process (KDD). Conforme apontado no trabalho de Moura(24), outros autores Gama et al.(25), García, Luengo e Herrera(26), Han, Pei e Kamber(27) denominam essa atividade como integração de dados. Considerando que os dados estejam reunidos em uma mesma estrutura tabular, a função de seleção de dados pode ter dois enfoques distintos: a escolha de atributos (redução de dados vertical) ou a escolha de registros (redução de dados horizontal) a serem considerados no processo de KDD;

- **Limpeza**: Nesta atividade ocorre a correção de dados incompletos, ruidosos ou inconsistentes. Entende-se por dados incompletos quando há informações ausentes ou insuficientes para determinados atributos. Dados ruidosos são os dados atípicos, divergentes do padrão normal esperado (*outliers*). Os dados são considerados inconsistentes quando contêm algum tipo de discrepância semântica entre si;

- **Codificação:** Tem a finalidade de transformar os domínios de valores de determinados atributos do conjunto de dados. É realizada para atender às necessidades específicas dos algoritmos de Mineração de Dados. Esta codificação pode ser Numérica–Categórica, quando divide valores de atributos contínuos em codificados; ou Categórica-Numérica, quando representa valores de atributos categóricos por códigos numéricos;

- **Enriquecimento dos dados:** tem a finalidade de agregar mais informações a cada registro do conjunto de dados, para fornecerem mais elementos ao algoritmo de AM;

- **Normalização de dados:** consiste em ajustar a escala dos valores de cada atributo de forma que estes sejam mapeados para valores restritos a pequenos intervalos, tais como de –1 a 1, ou de 0 a 1. É realizada para ajustar a escala de valores dos atributos e evitar que os algoritmos sejam influenciados de maneira tendenciosa;

- **Construção de atributos:** consiste em gerar atributos derivados, ou seja, atributos gerados a partir de atributos existentes;

- **Correção de Prevalência:** A finalidade dessa atividade é corrigir um eventual desequilíbrio na distribuição de registros com determinadas características. Alguns

métodos que podem ser usados para a correção da prevalência são: amostragem estratificada, replicação aleatória de registros ou ainda, utilizar algoritmos de AM preparados para lidar com problemas de prevalência; e

- **Partição do conjunto de dados:** para a avaliação dos modelos de AM construídos após as operações de pré-processamento é necessário que o conjunto de dados sejam divididos em pelo menos dois conjuntos: um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento deve conter os registros a serem utilizados na construção do modelo de conhecimento. Já o conjunto de testes deve conter os registros a serem utilizados na avaliação do modelo de conhecimento gerado. Outras técnicas usadas para particionar o conjunto de dados são: *Holdout*, *K-folding*, *Leave-one-out* e *Bootstrap*.

2.3 Aprendizado de Máquina

O campo da IA é um dos mais recentes em ciências e engenharia, tendo seu início logo após o início da Segunda Guerra Mundial. Abrange uma grande variedade de subcampos, do aprendizado e percepção até tarefas mais específicas como demonstração de teoremas matemáticos, direção autônoma ou jogos de xadrez (28).

O crescimento da complexidade dos problemas, da velocidade e do volume de dados a serem computacionalmente tratados motivou o desenvolvimento de ferramentas computacionais mais sofisticadas, mais independentes da intervenção humana para a aquisição de conhecimento. Na maioria, essas ferramentas se baseiam em AM, uma subárea da IA que faz parte de várias das tecnologias atualmente utilizadas (25).

O aprendizado de máquina pode ser definido como o "Aprendizado de uma Experiência E, em uma tarefa T e uma medida de desempenho D, desde que o desempenho D na tarefa T é melhorado com a experiência E" (29). É enfatizado aqui que o aprendizado ocorre quando a experiência leva a um melhor desempenho na tarefa específica. Nesse sentido, a finalidade do AM é fazer o computador aprender uma atividade a partir da experiência adquirida.

Os algoritmos de AM têm sido amplamente utilizados em diversas tarefas, que podem ser divididas em Preditivas e Descritivas. A figura 4 ilustra uma possível hierarquia das categorias de aprendizado e algumas tarefas associadas a cada categoria.

Nas tarefas preditivas, os algoritmos são usados em um conjunto de dados de treinamento rotulados para induzir a predição para um novo objeto representado pelos valores de seus atributos preditivos, o valor de um atributo alvo. Esse tipo de aprendizado é chamado de Supervisionado, por ser simulada a presença de um supervisor externo, que conhece o valor do atributo alvo. Esses tipos de modelos podem ser usados, por exemplo,

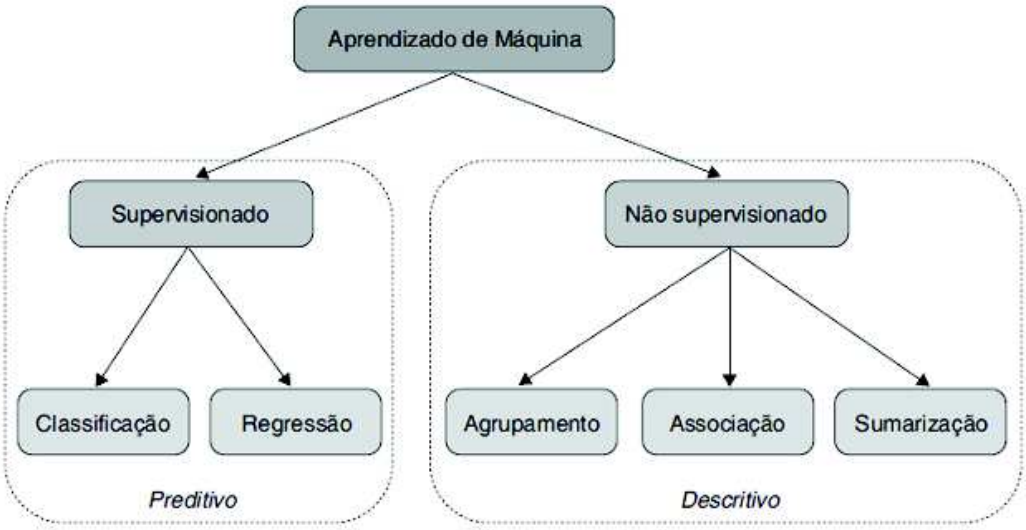

Figura 4 – Categorias de Aprendizado de máquina. Fonte: (25)

para prever o estado de saúde de um paciente a partir de seus sintomas (25).

Nas tarefas descritivas, os algoritmos de AM extraem padrões dos valores dos atributos preditivos de um conjunto de dados, não havendo a figura de um supervisor externo. Esses algoritmos são categorizados como de aprendizado não supervisionado. Uma das principais tarefas deste grupo é realizar o agrupamento de dados procurando grupos de objetos similares entre si no conjunto de dados (25).

Vale destacar que um requisito importante para algoritmos de AM é que possam lidar com dados imperfeitos, ou seja, dados que apresentam algum tipo de problema como presença de ruídos, dados redundantes ou ausentes. As técnicas de pré-processamento, entretanto, permitem identificar e reduzir, ou até eliminar esses problemas (25).

2.3.1 Algoritmos de aprendizado de máquina

Por meio dos algoritmos de AM é realizado o aprendizado com dados históricos. Existem diversos tipos de algoritmos, para todas as categorias de AM, ou seja, aprendizado supervisionado e não supervisionado. A seguir serão explicados alguns destes algoritmos.

Exemplos de algoritmos de AM:

Algoritmo Floresta Aleatória, do inglês Random Forest (RF) (30). É um método de AM que combina múltiplas árvores de decisão para realizar tarefas de classificação e regressão. Ele pertence à categoria de algoritmos de "*ensemble learning*" (aprendizado por comitê), onde vários modelos são combinados para obter um resultado mais robusto e preciso. No *Random Forest*, cada nó é dividido usando a melhor divisão entre um subconjunto aleatório de preditores selecionados para aquele nó específico. Essa estratégia, embora possa parecer contraintuitiva, mostra um desempenho ótimo em comparação com muitos outros classificadores, como máquinas de vetores de suporte e

redes neurais (30). Nos tópicos abaixo, é apresentada uma descrição do funcionamento do *Random Forest*, conforme Liaw e Wiener(31):

- Amostragem *bootstrap*: São criadas "n" amostras *bootstrap* a partir dos dados originais.
- Construção das árvores: Para cada amostra *bootstrap*, é construída uma árvore de classificação ou regressão não podada. Em cada nó da árvore, em vez de selecionar a melhor divisão entre todos os atributos, é feita uma amostragem aleatória de " m " atributos e a melhor divisão é escolhida entre esses atributos.
- Predição de novos dados: Para prever novos dados, as previsões das "n" árvores são agregadas. No caso de classificação, é usada a votação majoritária, e no caso de regressão, é calculada a média das previsões.
- Estimativa da taxa de erro: Uma estimativa da taxa de erro pode ser obtida com base nos dados de treinamento usando a seguinte abordagem. Para cada iteração de amostragem *bootstrap*, as previsões são feitas nos dados que não estão na amostra *bootstrap* (conhecidos como "*out-of-bag*" ou OOB data) utilizando a árvore construída com essa amostra *bootstrap*. As previsões OOB são agregadas e calcula-se a taxa de erro. Essa taxa de erro é chamada de estimativa OOB da taxa de erro.
- Uma vez que todas as árvores são construídas, o RF realiza a classificação de um novo exemplo mediante um processo de votação. Cada árvore emite uma classificação e, no caso de classificação, a classe mais frequente é selecionada como a predição final. No caso de regressão, a média das predições de todas as árvores é tomada como resultado final.

Em geral, a estratégia de divisão de nós usando um subconjunto aleatório de preditores no RF apresenta um desempenho significativamente bom em comparação com outros classificadores e oferece uma vantagem na robustez contra o *overfitting* (30). Vale destacar que a implementação deste algoritmo não permite o seu treinamento com atributos categóricos ou com dados ausentes na base.

Algoritmo Máquina de Vetor de Suporte (SVM), do inglês *Support Vector Machines*. É um método de aprendizado de máquina supervisionado usado tanto para classificação quanto para regressão. Os fundamentos do SVM são provenientes da Teoria de Aprendizagem Estatística de Vapnik (32). A ideia central do SVM é encontrar um hiperplano no espaço de características que melhor separe as diferentes classes de dados. No caso da classificação binária, o SVM busca um hiperplano que maximize a margem entre as classes, definida como a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte. Esses vetores de

suporte desempenham um papel importante na determinação do hiperplano de decisão. O SVM também permite lidar com dados não linearmente separáveis, por intermédio de uma técnica chamada "*kernel trick*". Essa técnica mapeia os dados de entrada para um espaço de características de dimensão superior, onde eles podem se tornar linearmente separáveis. Dessa forma, o SVM pode construir um hiperplano de decisão não linear no espaço de características transformado.

Ao treinar um modelo SVM, o objetivo é encontrar o hiperplano de decisão que maximize a margem e minimize a probabilidade de erro de classificação nos dados de treinamento. Para fazer isso, é utilizado um processo de otimização que envolve a minimização de uma função de custo regularizada, que combina a maximização da margem com a penalização por erros de classificação.

2.4 IA Explicável - XAI

A XAI surgiu com o propósito de tornar os resultados dos modelos de AM mais compreensíveis para os humanos. Conforme Adadi e Berrada(33), o termo XAI foi concebido em 2004 por Lent, Fisher e Mancuso(34), para descrever a capacidade de um sistema de explicar o comportamento da IA em aplicativos de jogos de simulação. No entanto, o problema de explicabilidade existe desde que os pesquisadores estudaram explicações para sistemas especialistas em meados da década de 1970 (33).

Conforme a Agência de Projetos de Pesquisa Avançada de Defesa, do inglês *Defense Advanced Research Projects Agency* (DARPA) (35), "XAI é definido como um conjunto de técnicas de aprendizagem que permitem que os usuários humanos entendam, confiem adequadamente e gerenciem com eficácia a geração de parceiros artificialmente inteligentes".

A Figura 5 evidencia o objetivo da XAI em combinar técnicas de interface homem-computador para traduzir modelos em diálogos de explicação compreensíveis e úteis para o usuário final.

Embora os termos explicabilidade e interpretabilidade sejam utilizados de forma intercambiável pela comunidade de AM, existem pequenas diferenças entre eles (36). Conforme Barredo Arrieta et al.(37), interpretabilidade se refere a uma característica passiva de um modelo, referindo-se ao nível no qual um determinado modelo faz sentido para um observador humano. O objetivo da interpretabilidade é tornar as decisões tomadas pelo sistema transparentes e fáceis de compreender para os usuários humanos (38). Por outro lado, explicabilidade é definida como a capacidade de explicar ou fornecer o significado em termos compreensíveis para um ser humano.

Assim, o termo explicabilidade está mais relacionado aos mecanismos internos de funcionamento dos modelos caixa-preta (36). Em (39), é discutido como a explicabilidade

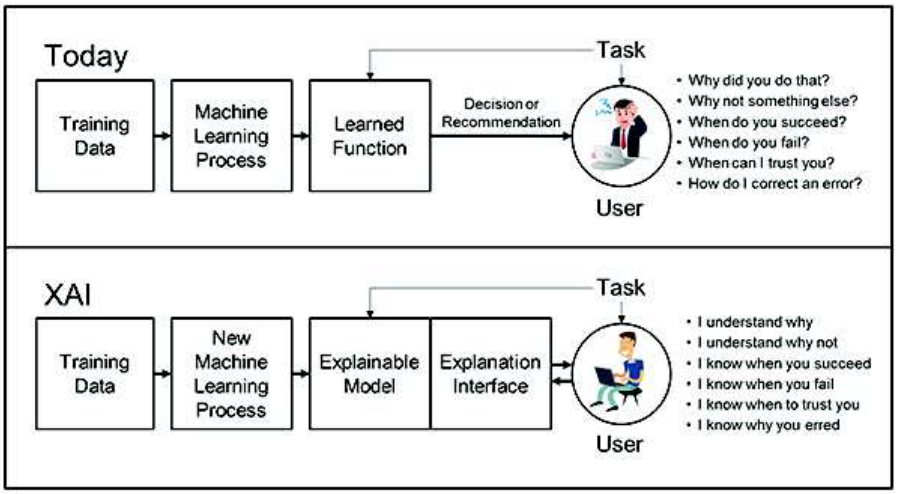

Figura 5 – Conceito XAI. Fonte: (35)

pode ser usada para garantir que os algoritmos estejam funcionando conforme o esperado. De acordo com esse artigo, a explicabilidade é uma ferramenta importante para garantir que os algoritmos estejam funcionando de maneira justa e confiável.

A necessidade de explicabilidade, conforme Srihari(40), pode ser motivada por:

- Justificativa: A capacidade de explicar a decisão de uma pessoa a outras pessoas. Entender a lógica por trás das previsões do modelo auxiliaria os usuários a decidirem quando confiar ou não em suas previsões;
- Controle: Refere-se ao cumprimento da legislação. Por exemplo, na União Europeia, o Regulamento geral de proteção de dados, do inglês *General Data Protection Regulation* (GDPR) prevê o direito a explicação para decisões automatizadas de alto risco;
- Melhoria: Para melhorar o desempenho de um sistema, primeiramente é necessário o entendimento dos seus pontos fracos, como detectar tendências no sistema; e
- Descoberta: Os sistemas de IA aprendem com milhões de exemplos para identificar padrões únicos nos dados, o que pode ser extremamente útil para as pessoas.

Existem modelos considerados intrinsecamente interpretáveis, os quais podemos citar, a regressão linear, a regressão logística e árvores de decisão. Estes modelos são conhecidos como caixa-branca, pois seus algoritmos são de fácil compreensão. No entanto, geralmente, essa explicabilidade natural vem com um custo no desempenho (41).

Uma caixa preta ou *Black Box* é uma função muito complicada para ser compreendida por qualquer ser humano, uma função proprietária ou um modelo difícil de solucionar. Os Modelos de Aprendizado Profundo, por exemplo, são modelos de caixa preta porque são recursivos, não intuitivos e difíceis de serem entendidos pelas pessoas (40).

2.4.1 Métodos XAI

Existem diversas perspectivas a considerar ao classificar um método de XAI. As principais taxonomias dos métodos são: escopo, tipo do modelo e o momento em que as explicações são geradas.

O escopo refere-se ao nível de interpretação. Quanto ao escopo, um método pode ser classificado como **global/Local**. Um método XAI chamado de nível **global** se concentra na explicabilidade do funcionamento de todo o modelo, bem como nos mecanismos de tomada de decisão, facilitando a compreensão de toda a lógica do modelo. Por outro lado, um método de nível **local** se concentra em explicar as decisões de um modelo para uma única instância.

Quanto ao tipo do modelo, os métodos podem ser classificados como **independentes do modelo**, também chamados de agnósticos do modelo ou **dependentes do modelo**, também chamado de *Specific-Model*. Os agnósticos do modelo são os métodos que separam a explicação de um modelo de aprendizado de máquina, permitindo que o método de explicação seja compatível com uma variedade de modelos. Por outro lado, métodos **dependentes do modelo** são aqueles nos quais as técnicas de XAI são específicas do modelo, assim, buscam interpretar baseando-se em alguma estrutura interna do modelo para conseguir extrair informações, dessa forma, só podem ser adotadas para explicar apenas um tipo específico de algoritmo.

O momento em que as explicações são geradas definem se a explicação é gerada durante a construção do modelo (**Ante-Hoc**), também conhecida como *Instrinsic Interpretable Model* ou após um modelo ser treinado (**Post-Hoc**). Métodos Ante-Hoc realizam a explicação desde o início da construção do Modelo. O objetivo da IA Ante-hoc é produzir modelos mais explicáveis, enquanto mantém um alto nível de desempenho, ou seja, produzir uma precisão de previsão. A abordagem Ante-Hoc são providas pelos métodos conhecidos como caixa-branca, onde a explicação está contida na estrutura interna do método. No entanto, as explicações Post-Hoc estão associadas às técnicas aplicadas para gerar explicações ao modelo após seu treinamento ter sido concluído, em resumo, métodos Post-Hoc, explicam a saída do modelo *Black Box*.

Além dessas perspectivas a considerar ao classificar um método XAI, conforme Arrieta et al.(42), existem diferentes abordagens de explicação utilizadas nas técnicas agnósticas ao modelo. Essas abordagens de explicação podem ser categorizadas em técnicas de simplificação do modelo, técnicas de relevância de atributos e técnicas de visualização. As técnicas de simplificação do modelo referem-se a métodos que aproximam um modelo opaco por meio da construção de um modelo mais simples e interpretável. Essa abordagem envolve a criação de um novo modelo com base no modelo original treinado, mas com uma estrutura mais compreensível. Essas técnicas são úteis para facilitar a interpretação de

modelos complexos.

As técnicas de visualização consistem em gerar representações visuais que destacam padrões, características ou regiões de interesse nos dados de entrada relevantes para as decisões do modelo. Essas visualizações podem assumir diferentes formas, como mapas de calor, sobreposições de ativação, saliência de atributos ou representações gráficas. Um exemplo de técnica nessa categoria é o Individual Conditional Expectation (ICE), que se baseia na geração de visualizações explicativas.

Por sua vez, as técnicas de relevância de atributos visam calcular a influência de cada atributo nos resultados do modelo. Elas fornecem *insights* sobre o funcionamento interno do modelo de aprendizado de máquina, atribuindo uma pontuação de relevância para cada variável utilizada. Tais técnicas são úteis para entender quais atributos são mais importantes para o modelo e como eles afetam as decisões tomadas por ele. Exemplos de técnicas nessa categoria incluem o Shapley Additive exPlanations (SHAP) e o Local Interpretable Model-Agnostic Explanations (LIME). A seguir, alguns métodos XAI Post-Hoc que utilizam a abordagem de relevância de atributos, abordagem utilizada neste trabalho, são descritos:

LIME

O LIME(43) é uma técnica utilizada para explicar as decisões tomadas por modelos de aprendizado de máquina complexos, como redes neurais ou algoritmos de *Random Forest*. Trata-se de um método local e agnóstico do modelo. O LIME treina um modelo interpretável, por exemplo, uma árvore de decisão, em um conjunto de dados feito a partir da permutação de amostras e a previsão correspondente da caixa preta. Ou seja, ele fornece explicações localmente sobre como uma determinada previsão foi feita aproximando qualquer modelo complexo por meio de modelos substitutos.

O processo geral do LIME envolve os seguintes passos:

- Seleção de instâncias: um conjunto de instâncias é selecionado a partir do conjunto de dados original.
- Perturbação das instâncias: as instâncias selecionadas são perturbadas aleatoriamente, criando instâncias modificadas.
- Predições do modelo complexo: o modelo de AM complexo faz previsões para as instâncias modificadas.
- Construção do modelo interpretável: um modelo interpretável é criado usando as instâncias modificadas e suas previsões correspondentes.
- As características mais importantes para a predição de cada instância são identificadas usando o modelo interpretável.

Ao fornecer uma explicação local para cada instância, o LIME permite entender quais características foram mais influentes na tomada de decisão do modelo complexo em uma base individual. Embora o modelo aprendido possa ter uma boa aproximação do comportamento local, ele não tem uma boa aproximação do global.

Shapley Values (SHAP)

O SHAP(44) é um método de explicação local baseado na utilização de valores Shapley. É baseado na teoria dos jogos de coalizão que melhora a explicabilidade calculando os valores de importância para cada atributo utilizando previsões individuais. O SHAP atribui uma pontuação a cada atributo em uma predição, indicando a contribuição desse recurso para a predição em comparação com o valor de referência. Essas pontuações de importância são baseadas no valor Shapley, um conceito da teoria dos jogos que mede a contribuição marginal de cada jogador para a função de pagamento.

Supõem-se aqui que cada valor de característica da instância é um jogador em um jogo, e a previsão é o pagamento geral distribuído entre os jogadores, ou seja, os atributos(recursos). O valor Shapley é a contribuição média na previsão de todas as coalizões possíveis de recursos, o que o torna computacionalmente caro quando há muitos recursos. Por exemplo, para k número de recursos, haverá 2k número de coalizões.

Na Figura 6, pode-se ver um exemplo de uma explicação utilizando a técnica local SHAP. A figura apresenta uma saída do método SHAP, aplicado em um modelo caixa preta para uma determinada instância, cujos atributos possuem os valores especificados. Após a aplicação do método, a explicação do modelo pode ser verificada pelos atributos que mais contribuíram para o resultado do modelo. Para a instância selecionada, o atributo *Age* foi o que mais contribuiu positivamente para o resultado.

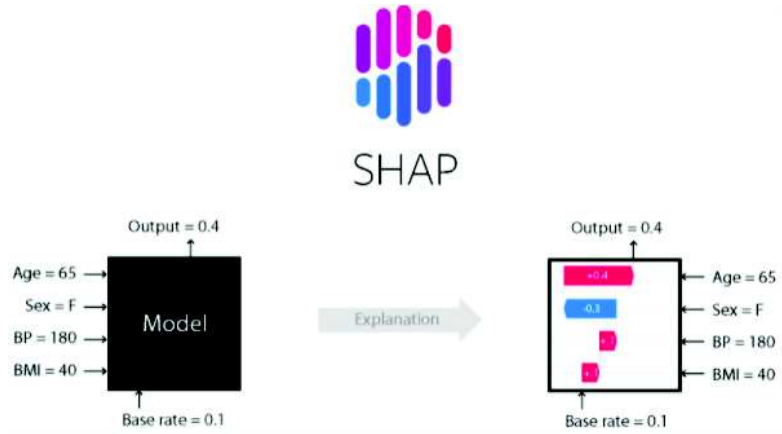

Figura 6 – Exemplo de método XAI local — técnica SHAP. Fonte: (45).

Feature Importance

É uma abordagem usada para entender e quantificar a importância relativa das diferentes características (ou atributos independentes) em um modelo. Para medir a

importância do recurso, calcula-se o aumento do erro de previsão do modelo após permutar o recurso. Essa técnica ajuda a identificar quais características têm maior influência nas previsões do modelo. Assim, ao permutar recursos importantes aumenta-se o erro do modelo, e ao permutar recursos sem importância, mantém-se o erro do modelo inalterado.

Existem várias maneiras de calcular a importância das características, e cada método pode ser aplicável a diferentes tipos de modelos. Dois importantes exemplos são a importância baseada em árvores de decisão e a importância baseada em permutação.

Quando se utiliza modelos baseados em árvores de decisão, como o *Random Forest* (30) ou o *Gradient Boosting* (46), é possível calcular a importância das características com base em sua contribuição para a redução da impureza (Gini ou entropia) nos nós da árvore.

2.4.2 Classificação dos métodos XAI

A Tabela 1 classifica os diferentes métodos de explicabilidade a partir de um conjunto de perspectivas: Post-hoc ou Ante-hoc; agnóstico do modelo ou específico do modelo e global ou local.

Método	Ante-hoc/ Post-hoc	Agnóstico/ Específico	Global/ Local
Linear/Logistic Regression	Ante-hoc	Específico	Ambos
Decision Trees	Ante-hoc	Específico	Ambos
Decision Rules	Ante-hoc	Específico	Ambos
K-Nearest Neighbors	Ante-hoc	Específico	Ambos
Partial Dependence Plot (PDP)(47)	Post-hoc	Agnóstico	Global
Individual Conditional Expectation (ICE)(48)	Post-hoc	Agnóstico	Ambos
Acumulated Local Efects (ALE) Plot	Post-hoc	Agnóstico	Global
Feature Interaction	Ambos	Agnóstico	Global
Feature Importance	Ambos	Agnóstico	Global
Global Surrogate	Post-hoc	Agnóstico	Global
Local Surrogate (LIME)	Post-hoc	Agnóstico	Local
Shapley Values (SHAP)	Post-hoc	Agnóstico	Local
Conterfactual explanations(49)	Post-hoc	Agnóstico	Local
Adversarial examples	Post-hoc	Agnóstico	Local
PrototypesKim, Rudin e Shah(50)	Post-hoc	Agnóstico	Local
Influential Instances	Post-hoc	Agnóstico	Local

Tabela 1 – Tabela de Métodos XAI. Adaptado de Islam et al.(41)

Ao analisar a tabela de Métodos XAI, adaptada de Islam et al.(41), percebe-se que dentre os métodos analisados, os Agnósticos do modelo, os Post-Hoc e os locais são os mais presentes. Além disso, conforme também pode-se observar na tabela, os métodos *Logistic Regression* e *Decision Rules* são exemplos de métodos que representam a categoria Ante-Hoc, pois se tratam de métodos interpretáveis. Os métodos LIME e SHAP são

exemplos de métodos Post-hoc, locais e agnósticos do modelo. Igualmente, os métodos globais podem ser exemplificados pelos métodos Partial Dependence Plot (PDP) e *Feature Importance*.

3 REVISÃO DA LITERATURA

Atualmente, dentre outros objetivos, as necessidades práticas têm impulsionado a pesquisa em IA, e alguns trabalhos de proveniência visando diferentes fases do ciclo de vida de AM já têm sido propostos nesta área do conhecimento. A fim de buscar trabalhos que utilizam proveniência no contexto do AM foram conduzidas pesquisas nas principais bases de dados: Scopus, ACM, IEEE e ScienceDirect, com o intuito de responder as seguintes questões:

- O objetivo do trabalho envolve soluções de proveniência para o contexto do AM?
- A captura da proveniência é realizada na fase de pré-processamento do ciclo de vida de AM?
- Quais são as técnicas, métodos, algoritmos e ferramentas utilizadas?
- A proveniência de dados é combinada com a explicabilidade em AM ou com alguma técnica de XAI?

Todos os trabalhos selecionados foram estudados e, para responder às questões de pesquisa, foram extraídas principalmente as seguintes informações: objetivo do trabalho; proposta de solução; se houve a captura e recuperação de proveniência para ciclo de vida AM, além de como, era capturada essa informação. Além disso, também foram verificados trabalhos cujo objetivo era contribuir com a explicabilidade em AM, ainda que a proposta de solução não fosse através da proveniência. Outrossim, foi verificado o alinhamento com o modelo PROV-DM da W3C. Como foram encontrados poucos trabalhos que abrangessem a fase de pré-processamento, também foram considerados outras fases do ciclo de vida.

Os trabalhos selecionados correspondem às publicações entre 2017 e 2022. Dessa forma, a busca foi realizada por intermédio do título, resumo e palavras-chave em cada biblioteca, utilizando a seguinte *string*: ("*provenance*" OR "*data provenance*") AND ("*explainable ai*" OR "*explainability*") AND *preprocessing*. Todos os trabalhos que utilizavam proveniência, recomendavam a utilização da proveniência para AM ou cujo objetivo do trabalho fosse contribuir com a explicabilidade em AM foram considerados. Posteriormente, outros trabalhos foram encontrados e agregados à lista dos selecionados. No entanto, só consideramos os trabalhos que incluíam a fase de pré-processamento ou todo o ciclo de vida de AM.

É importante destacar que os trabalhos que utilizam a proveniência somente para a fase de treino (15, 51), por exemplo, não foram considerados, uma vez que o objetivo desses

trabalhos é, em geral, a captura da configuração do modelo e seu respectivo desempenho, visando a melhoria da configuração desses modelos.

A obtenção de uma quantidade reduzida de trabalhos que atenderam aos critérios de busca demonstra a natureza disruptiva e atual da pesquisa em andamento. No entanto, mesmo diante desse cenário, foi possível adquirir os fundamentos necessários para conduzir o trabalho de forma adequada e situá-lo no contexto atual do estado da arte na área.

3.1 *Capturing and Querying Fine-grained Provenance of Preprocessing*

Neste trabalho, Chapman et al.(8) propõem a formalização e categorização de um conjunto básico de operadores de pré-processamento de dados projetados para limpar, transformar e alterar os dados na preparação de modelos preditivos. O objetivo é possibilitar a explicação sobre o efeito de cada transformação em um *pipeline* de pré-processamento nos dados alimentados em um modelo.

Primeiramente, foi proposta uma formalização e categorização de um conjunto básico de operadores para redução, aumento e transformação de dados, onde é mostrado como *pipelines* de pré-processamento de dados comuns podem ser expressos como uma composição desses operadores. Dessa forma, para cada um desses operadores, é associado um padrão de proveniência, de baixa granularidade para *pipelines* de AM, que descreve o efeito do operador nos dados no nível de detalhe apropriado. As seguintes operações de pré-processamento foram propostas no trabalho:

- Redução de dados;
- Transformações de dados; e
- Aumento de dados.

Como contribuição deste trabalho foi implementada uma biblioteca em Python para captura de proveniência onde são feitas anotações associadas a operadores da álgebra relacional para descrever os seus efeitos nos elementos individuais dos dados.

Na Tabela 2, adaptada desse trabalho, os autores relacionam, para as bibliotecas Orange e Scikit-learn, as respectivas operações de pré-processamento, categoria e função relacionada.

Orange	**ScikitLearn**	**Category**	**Operator**
Feature Statistics	Feature selection	Data Reduction	Feature selection
Select Data by index	Dataframe op.		Instance selection
Select Columns	Feature Selection		Drop Columns
SelectRows	Dataframe op.		Drop Rows
Data Sampler	Imbalaced learn		Undersampling
Impute	SimpleImputer	Data Transformation	Imputation
ApplyDomain	FunctionTransformer		ValueTransformation
Edit Domain	BinarIzer		Binarization
Preprocess	Normalizer		Normalization
Discretize	kBinDiscretizer		Discretization
Feature Constructor	FunctionTransformer	Data augmentation	Space Transformation
Create Class	FunctionTransformer		Instance Generation
Data Sampler	Imbalanced-learn		Oversampling
Corpus	Label Encoder		String Indexer
Preprocess	OneHotEncoder		One-Hot Encoder

Tabela 2 – Operações em pipelines de AM de preparação de dados do Orange (47) e do Scikit-Learn (52). Adaptado de (8)

3.2 *Data Provenance for Data Science (DPDS)*

Nesse artigo, Chapman et al.(12) abordam a importância da justificativa e explicação de cada etapa de um pipeline de ciência de dados que envolve a limpeza, transformação e alteração de dados para preparação de um modelo de dados. É apresentada a ferramenta Data Provenance for Data Science (DPDS), que auxilia os especialistas em dados a coletar, armazenar e investigar a proveniência de cada elemento individual em um conjunto de dados. A ideia é permitir ao cientista de dados que utiliza a ferramenta poder compreender como os passos de pré-processamento alteram o perfil de dados de qualquer conjunto de dados utilizado no processo de ciência de dados, como, por exemplo, no conjunto de treino. Tal modelo é implementado como um grafo dirigido acíclico, do inglês *Directed Acyclic Graph* (DAG), que representa o histórico completo de proveniência de dados.

A plataforma DPDS fornece uma interface de usuário para visualizar o histórico completo de proveniência de dados, desde a sua origem até o uso final. A interface permite que os usuários naveguem no DAG de proveniência, visualizem as transformações aplicadas aos dados em cada etapa da análise e acessem informações detalhadas sobre os dados em cada etapa.

A Figura 7 mostra a arquitetura da plataforma DPDS. O componente *Prov-generator* é o responsável por produzir a proveniência de cada operador no pipeline, analisando o efeito desse operador no conjunto de dados subjacente. Neste trabalho os autores adotam o padrão W3C como modelo de proveniência.

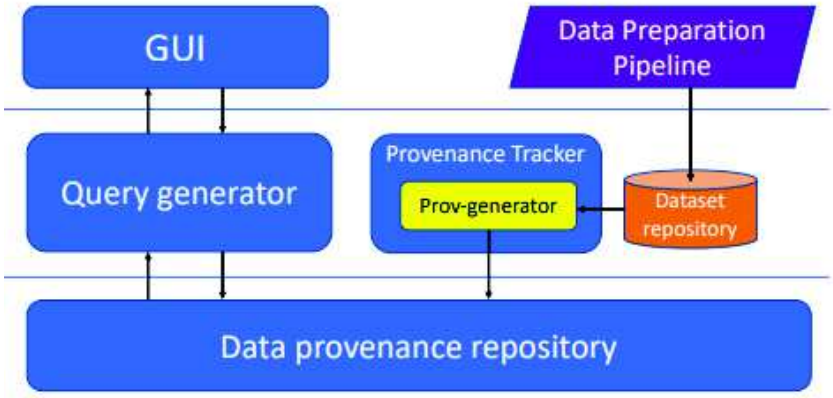

Figura 7 – Arquitetura do DPDS. Fonte: (12)

3.3 Assistant PP

Nesse trabalho, Moura et al.(14) apresentam um assistente para auxiliar o usuário não-especialista na seleção de operadores de pré-processamento de dados. A ferramenta, além de guiar os usuários não-especialistas no preparo do conjunto de dados para treino de um modelo, captura os operadores de pré-processamento utilizados durante o processo.

Além disso, também foi realizado neste trabalho um levantamento dos termos de pré-processamento de acordo com livros, artigos científicos e normas da área. Os termos levantados foram uniformizados em um glossário de termos usando o critério da maior ocorrência do termo e da categorização observados nas literaturas referenciadas.

A Figura 8 ilustra a arquitetura da ferramenta *Assistant* PP, que foi construída conforme a Ontologia PPO-O, apresentada também nesse trabalho.

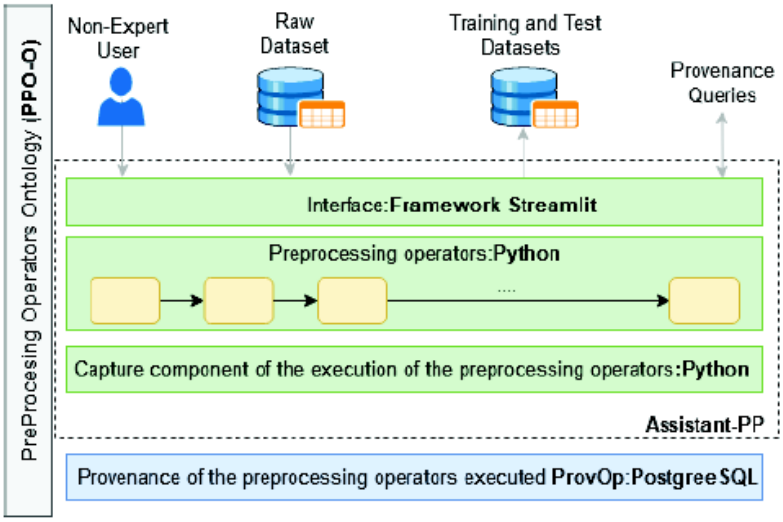

Figura 8 – Arquitetura *Assistant* PP. Fonte: (14)

3.4 VAMSA

Nesse artigo, Namaki et al.(13) apresentam o problema de rastreamento de proveniência em AM. O objetivo nesse trabalho é rastrear automaticamente quais colunas em um conjunto de dados foram usadas para derivar os recursos de um modelo de AM específico. Para o rastreamento da proveniência em *scripts* de forma automática é apresentado o sistema VAMSA, um sistema modular que extrai a proveniência de scripts Python sem exigir nenhuma alteração no código.

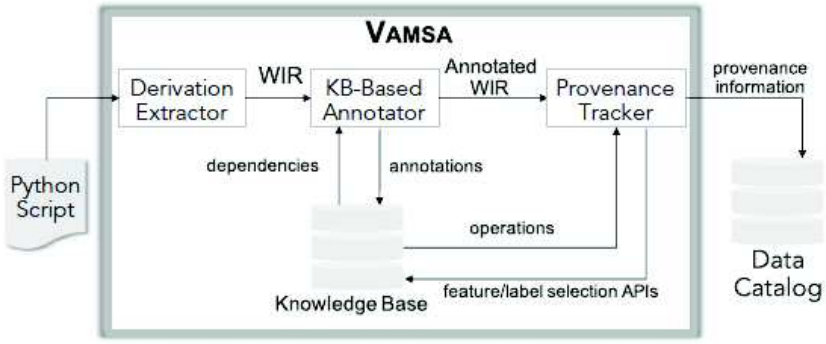

Figura 9 – Arquitetura VAMSA. Fonte: (13)

A Figura 9 apresenta a arquitetura VAMSA. Especificamente, o VAMSA processa scripts de ciência de dados com os seus três módulos principais: o *Derivation Extractor*, o *KB-Based Annotator* e o *Provenance Tracker*.

O *Derivation Extractor* gera uma representação intermediária do fluxo de trabalho (WIR), extraindo do *script*, os principais elementos incluindo as bibliotecas importadas. Já o *KB-based Annotator* anota variáveis no WIR com base nas suas funções no *script*. Finalmente, o *Provenance Tracker* infere um conjunto de colunas que foram explicitamente incluídas ou excluídas das características, utilizando o anotador WIR.

Adicionalmente, nesse trabalho os autores relatam ter realizado uma pesquisa em algumas empresas que trabalham com *Big Data* para levantar a necessidade de rastreamento de proveniência em AM, e o seguinte resultado foi obtido: 88% dos entrevistados relatam que a proveniência pode ser útil em vários cenários, dentre eles, compartilhamento do modelo, conformidade e imparcialidade.

3.5 USRPRUNG

Neste artigo, Rupprecht et al.(53) apresentam o USRPRUNG, um sistema de coleta de proveniência transparente projetado para ambientes de ciência de dados. A filosofia do trabalho é capturar a proveniência e construir a linhagem integrando-se ao ambiente

de execução para rastrear automaticamente os parâmetros de configuração estáticos e de tempo de execução de *pipelines* de ciência de dados, sem exigir que os cientistas de dados façam alterações em seus códigos.

A arquitetura URSPRUNG consiste em três componentes principais: *Provenance Sources*, *Collection System* e *Provenance Store and GUI*. O componente *Provenance Sources* representa as fontes de proveniência, fornecendo a base do sistema e emitindo as proveniências relativas a eventos sendo capturadas e processadas. O *Collection system* está no centro de URSPRUNG e é responsável por consumir os eventos emitidos pelas fontes de proveniência. Pelo componente *Provenance Store and GUI*, os eventos de proveniência são persistidos no banco de dados de proveniência. Para o armazenamento da proveniência, o banco de dados usado é o relacional devido à sua escalabilidade e desempenho rápido para consultas. A Figura 10 apresenta a arquitetura USRPRUNG.

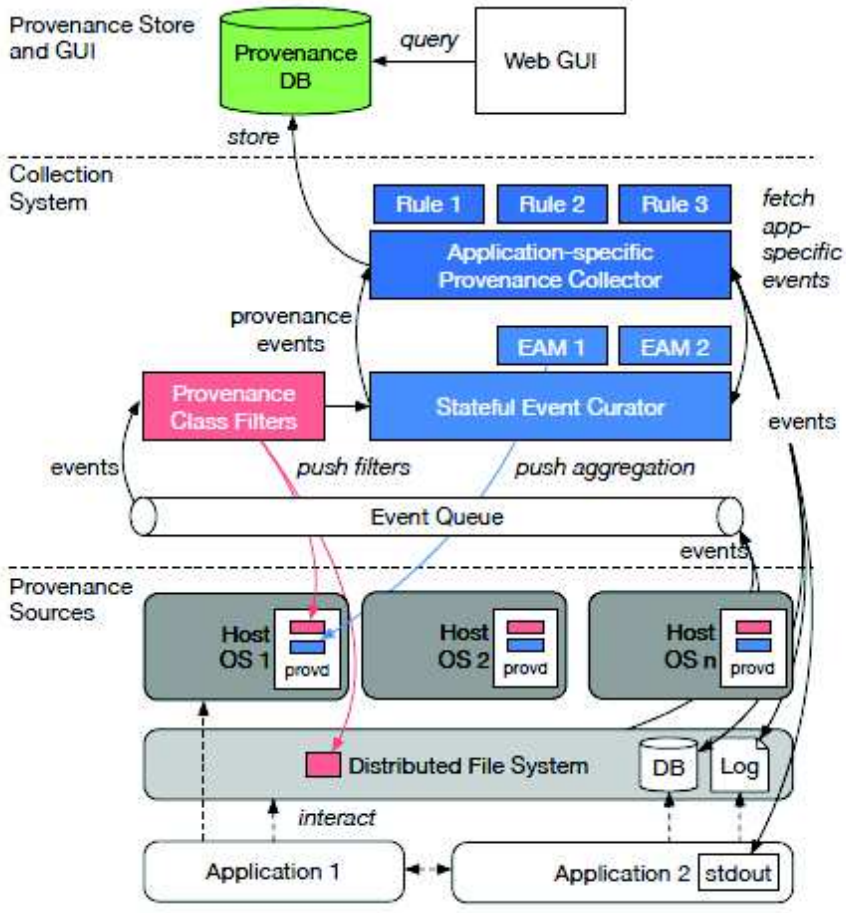

Figura 10 – Arquitetura USRPRUNG. Fonte: (53)

3.6 ExplAIner

O objetivo do trabalho de Spinner et al.(54) é propor e apresentar uma implementação do *framework* conceitual do explAIner, uma ferramenta de *visual analytics* para

interação e explicabilidade em modelos de AM. Além disso, neste trabalho também é feita uma avaliação da abordagem por meio de um estudo com usuários de diferentes níveis de experiência para avaliar a qualidade da abordagem e seu impacto no fluxo de trabalho. O *framework* é implementado como um *plug-in* do TensorBoard que é uma plataforma amplamente utilizada na comunidade de AM e fornece funcionalidades nativas para visualização e análise de modelos de AM. Esse *framework* é utilizado principalmente para visualizar gráficos de modelos, monitorar métricas de desempenho, analisar distribuições de tensores, entre outras tarefas relacionadas à análise de modelos de AM.

Na Figura 11 é mostrada uma representação visual do design e do fluxo de trabalho do explAIner.

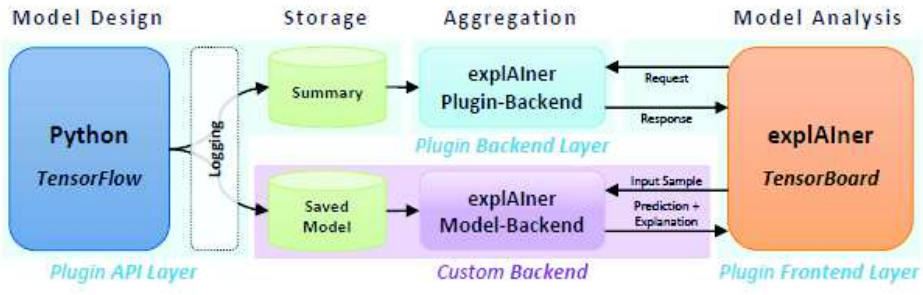

Figura 11 – Visão do design do ExplAIner. Fonte: (54)

A ferramenta explAIner também realiza a captura de proveniência, através da implementação de uma barra de proveniência no TensorBoard, que permite ao usuário anotar e salvar descobertas interessantes durante o processo de exploração e explicação. A barra de proveniência atua como um quadro negro digital persistente, abrangendo partes dos mecanismos de monitoramento e controle global, como rastreamento e relatórios de proveniência. Além disso, a barra de proveniência desempenha um papel importante na fase de relatórios, permitindo que os usuários documentem seus pensamentos, estruturem o processo de forma narrativa e forneçam justificativas e evidências para suas ações.

3.7 PROV-ML

No trabalho de Souza et al.(16) é proposta uma solução ponta-a-ponta para rastrear as transformações de dados que ocorrem no ciclo de vida do AM, desde a curadoria de dados brutos até a geração de modelos treinados, fornecendo captura de proveniência e análise de dados por meio de consultas de proveniência com um vocabulário padrão. É apresentada uma nova representação de dados de proveniência chamada de PROV-ML, construída sobre o W3C PROV (17) e o ML Schema (MLS) (55).

Neste trabalho são considerados os Projetos de Ciência de Computação e Engenharia, do inglês *Computational Science and Engineering* (CSE) em grande escala, os quais são

muitas vezes multidisciplinares, com colaboração de usuários com diferentes habilidades nos dados de domínio, por exemplo, matemática, estatística, métodos computacionais, cientistas de dados, dentre outros. Dessa forma, é considerada a necessidade de dados de proveniência desses usuários, que geralmente, realizam tipos distintos de análise, com diferentes requisitos de proveniência. Neste trabalho o ciclo de vida do AM no CSE é dividido em três fases principais: Curadoria de Dados, Preparação de Dados para o Aprendizado e Aprendizado, conforme figura 12.

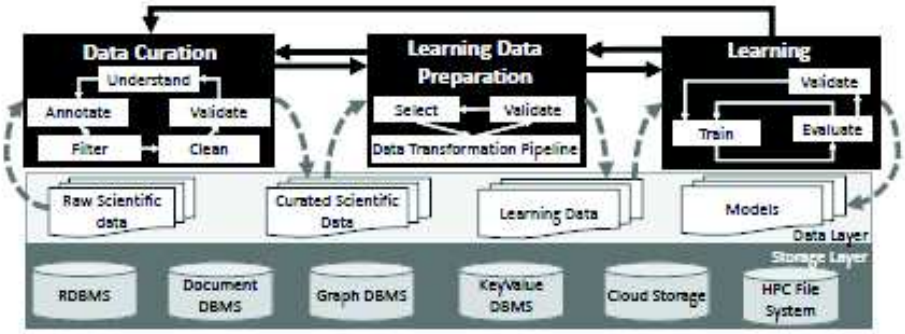

Figura 12 – Ciclo de VIDA CSE. Fonte: (16)

A PROV-ML é uma representação de proveniência de dados para fluxos de trabalho no ciclo de vida de AM no CSE. É compatível com W3C PROV e MLS e estende a representação dos dados de proveniência do fluxo de trabalho do ProvLake (56), o qual é uma extensão do PROV, fornecendo suporte para a preparação de dados de aprendizado do ciclo de vida. Ele herda os benefícios do ProvLake, permitindo a integração da proveniência de dados específicos do domínio processados por fluxos de trabalho na fase de curadoria.

3.8 PROV-IO

Neste trabalho de Han et al.(57), os autores propõem um framework de proveniência centrado em I/O para dados científicos em sistemas de computação de alto desempenho, do inglês High Performance Computing (HPC), que aborda os desafios das soluções existentes e atende às necessidades específicas dos cientistas de domínio. O framework PROV-IO é composto por três principais componentes:

- Provenance Tracking (rastreamento de proveniência): captura as operações de entrada e saída (I/O) de múltiplas interfaces de I/O em um workflow científico;
- Provenance Store (armazenamento de proveniência): persiste a proveniência capturada em Triplo RDF; e
- User Engine (motor do usuário): permite que os usuários consultem e visualizem informações de proveniência.

Além desses três componentes principais, o framework também inclui o modelo PROV-IO, derivado do padrão W3C e das características típicas de fluxos de trabalho e necessidades de proveniência de cientistas de domínio. A Figura 13 ilustra o framework PROV-IO.

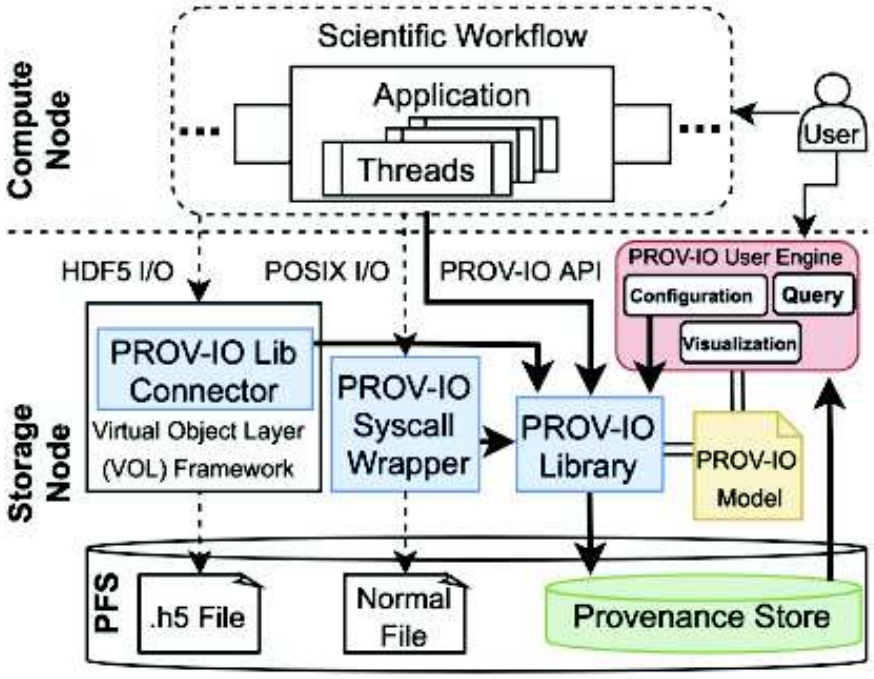

Figura 13 – Arquitetura PROV-IO. Fonte: (57)

O framework PROV-IO utiliza um modelo de proveniência, PROV-IO Model, para descrever a proveniência de dados científicos, derivado do padrão W3C PROV (Provenance Data Model), mas adaptado para atender às necessidades específicas de cientistas de domínio e fluxos de trabalho científicos. Nesse sentido, introduz novos conceitos para lidar com as necessidades específicas de cientistas de domínio, como a classe extensível (Extensible Class), que permite a definição de novas classes de proveniência, e a relação de dependência (Dependency Relation).

3.9 EdnaML

No trabalho de Suprem et al.(58), os autores apresentam uma API e framework declarativos para aprendizado profundo reprodutível. É apresentado como o framework pode ser utilizado para gerenciar pipelines de AM em alto nível de abstração, enquanto ainda oferece flexibilidade para modificar qualquer parte do pipeline por meio de blocos de construção. O EdnaML é projetado com uma estrutura *bottom-up*, que consiste em blocos de construção básicos para um pipeline de AM, como abstrações para dados, modelo, treinamento e implantação. Por ser um sistema de gerenciamento de pipeline de AM, de ponta a ponta de código aberto, é altamente extensível e flexível. Isso permite que

os usuários adicionem novas funcionalidades e personalizem o EdnaML para atender às suas necessidades específicas de aprendizado de máquina. A Figura 14 ilustra a estrutura ascendente da API em camadas do EdnaML.

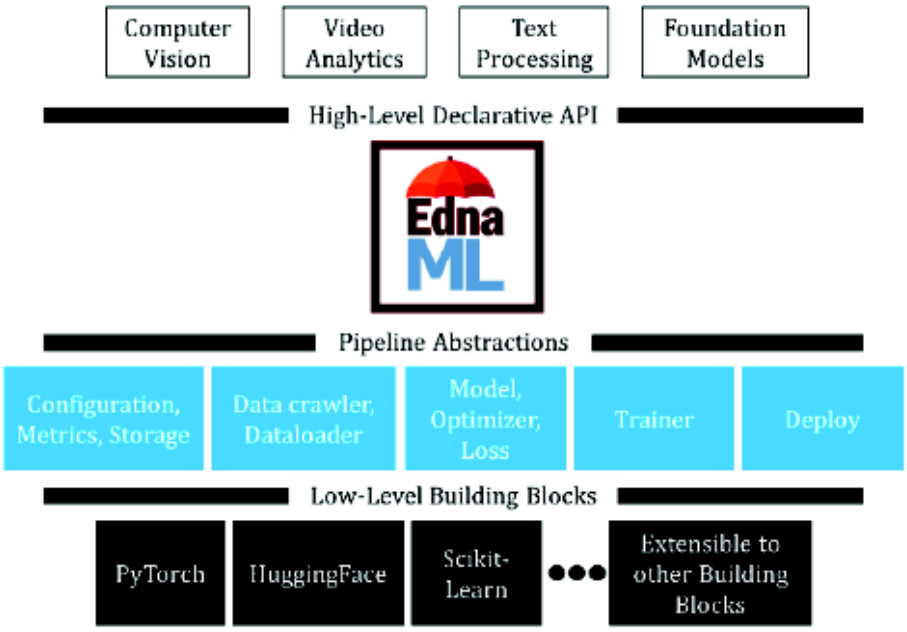

Figura 14 – Design Ascendente da API em camadas EdnaML. Fonte: (58)

A arquitetura geral do EdnaML, consiste em três principais componentes: o componente de configuração, o componente de pipeline e o componente de execução. O componente de configuração é responsável por definir a estrutura do pipeline, enquanto o componente de pipeline é responsável por definir as etapas. O componente de execução é responsável por executar o pipeline e gerenciar a proveniência dos dados e modelos em todas as fases do pipeline.

3.10 ProML

O trabalho de Kennedy et al.(59) apresenta a plataforma ProML, que utiliza blockchain e contratos inteligentes para gerenciar a proveniência de ativos de AM em equipes distribuídas. A solução descentralizada promete garantir a segurança, a privacidade e a justiça dos ativos de AM, sem depender de terceiros vulneráveis. Além disso, o artigo apresenta a abordagem *Artefact-as-a-State-Machine* (ASM) como uma nova arquitetura para capturar a proveniência de forma controlada pelos usuários. A abordagem ASM permite que os participantes controlem o quê e como as informações de proveniência são capturadas, sem expô-los às complexidades subjacentes de um blockchain e dos contratos inteligentes. A Figura 15 ilustra a plataforma ProML.

A plataforma é composta por vários nós ProML, cada um representando um participante em um fluxo de trabalho de AM. Os nós se conectam para formar um

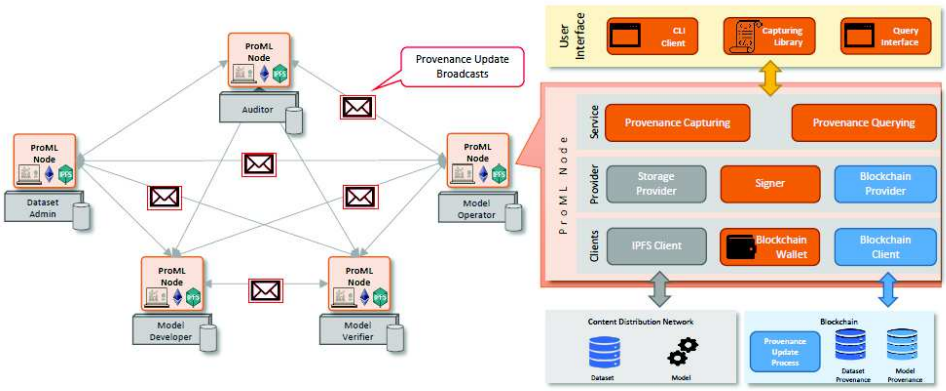

Figura 15 – Platataforma ProML. Fonte: (59)

blockchain e outras infraestruturas descentralizadas, como uma rede privada de distribuição de conteúdo baseada no *Interplanetary File System* (IPFS). A plataforma implanta contratos inteligentes no blockchain para capturar e atualizar informações de proveniência.

3.11 Semantic Description of Explainable Machine Learning Workflows for Improving Trust

O artigo de Nakagawa et al.(60) trata sobre a criação de uma ontologia para descrever os principais componentes do processo de AM e explicação Post-hoc. O objetivo principal é fornecer uma visão abrangente de como um modelo de AM chega a resultados específicos, facilitando a compreensão do usuário.

A ontologia foi desenvolvida com base no alinhamento dos conceitos do ML-Schema (MLS), que é a principal ontologia específica do domínio que serviu de inspiração para o desenvolvimento. O MLS foi expandido de modo a preencher algumas lacunas, como por exemplo, as execuções em experimentos. A ontologia proposta foi desenvolvida com base em uma ontologia fundamental, tornando-a compatível com outras ontologias já existentes que seguem o padrão da Unified Fundation Ontology (UFO) (61). Isso significa que ela pode se integrar facilmente com outras ontologias que compartilham a mesma estrutura definida pela UFO. Além disso, a ontologia proposta foi estruturada em três módulos distintos: um módulo geral que representa aspectos gerais de AM; um específico para classificação supervisionada e um terceiro módulo para a explicação, que representa o processo de explicação Post-Hoc.

A Figura 16 ilustra o modelo conceitual da ontologia do artigo. A área em cinza representa o módulo geral de AM. Neste módulo foram acrescentados conceitos para organizar as execuções em Experimentos. A área em amarelo representa o módulo específico para a classificação supervisionada. Nesse módulo, cada tarefa é representada como uma

subclasse da classe de operação no módulo específico de AM. Por último, a área em verde representa o módulo de explicação. Esse módulo representa o processo de explicação Post-hoc, adicionando as operações de Explicar e Avaliar a explicação aos participantes correspondentes.

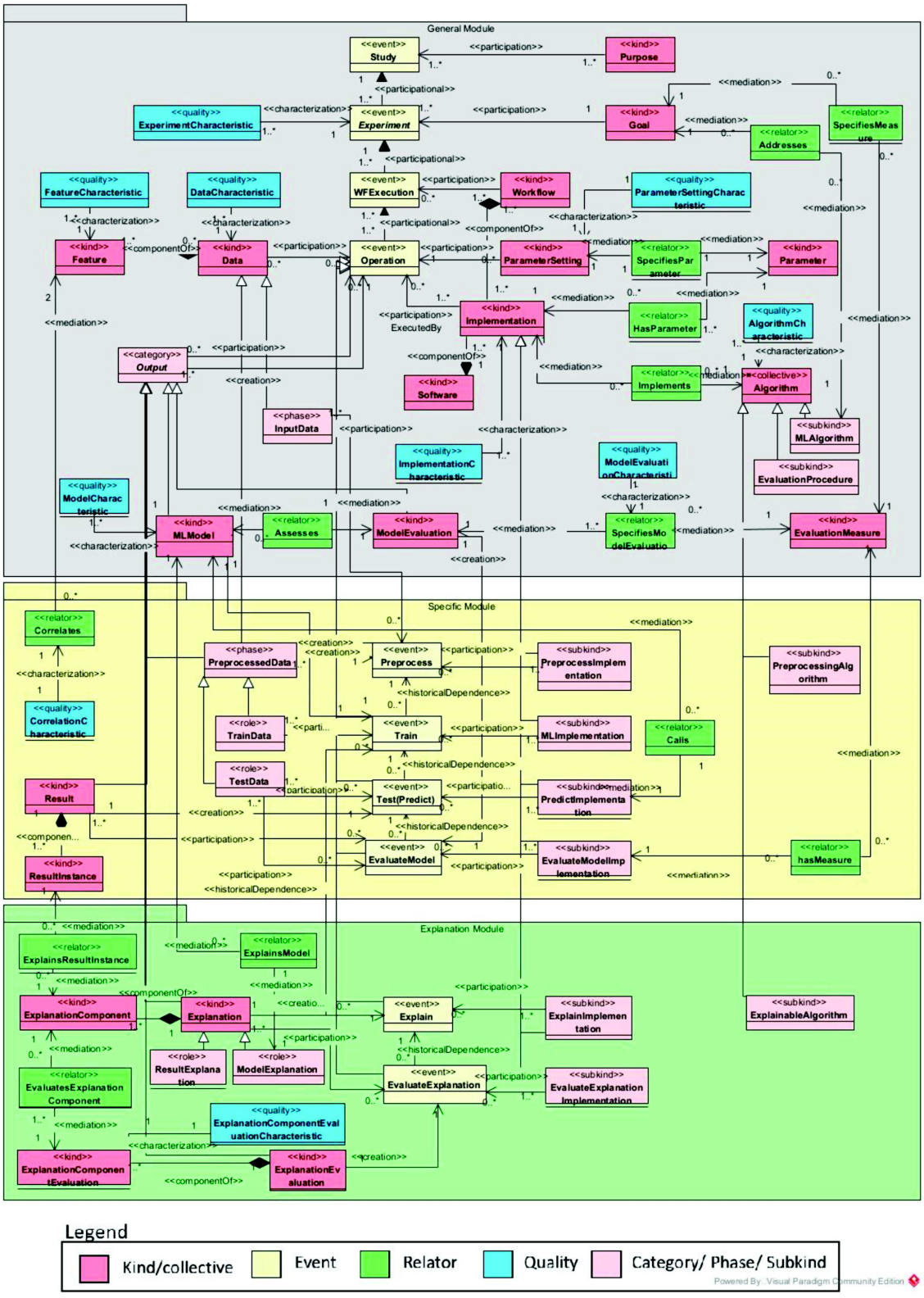

Figura 16 – Modelo conceitual da Ontologia proposta. Fonte: (60)

3.12 Comparativos dos trabalhos

A Tabela 3 apresenta uma comparação dos trabalhos relacionados, classificados conforme os seguintes critérios que indicam a captura de proveniência para a tarefa de AM, com ênfase no pré-processamento:

- **C1:** qual tipo de proveniência o trabalho captura. R-Retrospectiva, P-Prospectiva ou A-Ambas;
- **C2:** se a proveniência abrange a captura de operadores de Pré-processamento do ciclo de vida;
- **C3:** se o trabalho captura a Explicabilidade do Modelo (Técnica XAI);
- **C4:** se a captura de proveniência se baseia no modelo do W3C PROV;
- **C5:** se a abordagem captura outras fases do ciclo de vida de AM; e
- **C6:** se o escopo da proveniência é referente a fluxo de trabalho simples (S) ou complexo (C). Um fluxo de trabalho simples é aquele em que é considerado um único processo para execução, enquanto um fluxo de trabalho complexo refere-se a sistemas em que vários fluxos de trabalho podem se interligar.

Tabela 3 – Sumário de Trabalhos Relacionados

Trabalho	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Chapman et al.(8)	R	✓		✓		S
Chapman et al.(12)	R	✓		✓		S
Moura et al.(14)	R	✓				S
Namaki et al.(13)	R	✓				S
Rupprecht et al.(53)	R	✓			✓	S
Spinner et al.(54)	R		✓		✓	S
Souza et al.(16)	A	✓		✓	✓	C
Han et al.(57)	R			✓	✓	S
Suprem et al.(58)	A	✓			✓	S
Kennedy et al.(59)	A	✓			✓	S
Nakagawa et al.(60)	–		✓			S
Este trabalho	R	✓	✓	✓	✓	S

Nos trabalhos de Chapman et al.(8) e Chapman et al.(12), a captura da proveniência inclui a fase de pré-processamento, sendo utilizada uma metodologia compatível com a W3C, porém não há objetivos de capturar outras fases do ciclo de AM, nem tampouco de estender explicações com técnicas XAI.

Nos trabalhos de Moura et al.(14) e Namaki et al.(13) também há captura da fase de pré-processamento, no entanto, nesses trabalhos não é utilizado o padrão W3C, bem como não englobam outras fases do ciclo de vida de AM. Vale observar que em Namaki et al.(13) embora não sejam capturados os operadores de processamento, são extraídos dos scripts quais colunas foram excluídas e incluídas para compor o modelo.

No trabalho de Rupprecht et al.(53) não há compatibilidade com a padronização do W3C, além de também não estar explícito neste trabalho se há captura de dados de pré-processamento. Da mesma forma, nos trabalhos (58) e (59), também não foi utilizado modelo compatível com o W3C.

Diferentemente, em Souza et al.(16), é utilizado o padrão do W3C e há a captura em todas as fases do ciclo de vida, embora não especificamente dos operadores de pré-processamento. Da mesma forma, no trabalho de Han et al.(57) é utilizado o padrão W3C e o foco é em ambientes distribuídos, sendo que este trabalho se concentra em workflow científico de modo geral, enquanto o de Souza et al.(16) é específico para workflow de AM. Em ambos os trabalhos, porém, não há a utilização de técnicas XAI. Vale destacar que em Souza et al.(16) o foco é atender a captura de proveniência para fluxos de trabalhos complexos, realizados por equipes multidisciplinares em um mesmo projeto.

No trabalho de Spinner et al.(54), que teve um foco diferente dos demais, uma vez que contemplou explicações do modelo de AM, tendo como objetivo principal alcançar uma compreensão clara dos modelos por meio do uso do *Visual Analytics*. Nesse contexto, a proveniência refere-se às anotações feitas pelos usuários sobre sua própria compreensão das explicações fornecidas por esses modelos. É importante destacar que, na proposta desse trabalho, a proveniência não possui uma relação direta com uma fase específica do ciclo de vida.

Igualmente, no trabalho de Nakagawa et al.(60) o foco também foi contribuir com explicações em AM, no entanto, neste trabalho não há captura de proveniência. Neste estudo, também se buscou contribuir com a explicabilidade Post-Hoc, conforme o presente trabalho. Para alcançar esse objetivo, no entanto, foi proposta uma ontologia com a finalidade de enriquecer semanticamente os resultados XAI.

Em nenhum desses trabalhos foi possível encontrar captura de proveniência da explicação do modelo. Desse modo, torna-se evidente que o presente trabalho se distingue dos demais ao priorizar, principalmente, as etapas de pré-processamento e explicação do modelo, que ocorre após o treinamento do mesmo. Até onde foi possível investigar, essa proposta se diferencia dos trabalhos relacionados, uma vez que as obras citadas não visam a integração da proveniência dos dados com a explicabilidade do modelo.

4 ABORDAGEM XMML-PPP

Este capítulo apresenta a abordagem xMML-PPP - *Explainable Machine Learning Model supported by Pre-processing Provenance*, proposta nesta dissertação, destacando tanto a concepção da arquitetura quanto o processo envolvido. A xMML-PPP tem por objetivo melhorar a compreensão da explicabilidade dos modelos de AM, acrescentando, para isso, a proveniência dos dados, com foco na fase de pré-processamento. As operações de pré-processamento realizadas nos dados permitem reconhecer o tratamento que os dados recebem antes de serem utilizados no treinamento e o quanto eles contribuíram com o resultado do modelo. Dessa forma, a abordagem propõe obter maior compreensibilidade aos resultados do modelo, com a utilização de técnicas XAI cujo método de explicação utilize a importância dos atributos.

4.1 Explicabilidade de Dados e de Modelo

Neste trabalho a explicabilidade de AM é dividida em duas áreas distintas, porém relacionadas: (i) a explicabilidade de dados, fornecida pela proveniência dos dados, principalmente na fase de pré-processamento do ciclo de vida do modelo; e (ii) a explicabilidade do modelo, fornecida pelas técnicas XAI.

O termo explicabilidade de dados já foi usado no contexto de AM com o sentido de prover explicabilidade pela interpretação dos dados de treino (62, 63). Nesta dissertação esse conceito foi estendido para abranger a proveniência dos dados de treinamento, especificamente aos da fase de pré-processamento. Dessa forma, a "explicabilidade de dados" é definida aqui como uma técnica que captura a proveniência dos dados na fase de pré-processamento, melhorando simultaneamente a compreensão da explicabilidade do modelo por meio da criação de relações entre eles para aumentar a confiabilidade dos resultados do modelo de AM. Portanto, diferenciamos a explicabilidade de dados da proveniência dos dados, devido à sua relação interna com a explicabilidade do modelo de AM.

Adicionalmente, é de grande valor obter informações adicionais sobre os dados do fluxo de trabalho, pois isso pode contribuir significativamente para o entendimento do processo. Com esse objetivo, alguns desses dados são capturados ao longo do fluxo de trabalho.

Nesse contexto, a xMML-PPP também contempla o armazenamento dos dados referentes ao desempenho e a contribuição de cada atributo para o resultado do modelo. Esse armazenamento possibilita estabelecer uma relação entre a configuração e o desempenho do modelo com as etapas de pré-processamento realizadas, bem como compreender as

contribuições individuais dos respectivos atributos que influenciaram a construção do modelo.

Essa inclusão de dados do desempenho e das contribuições dos atributos permite uma análise mais abrangente e aprofundada do fluxo de trabalho. Ao considerar-se a configuração, o tratamento dos dados no pré-processamento e as contribuições específicas dos atributos, é possível obter-se *insights* valiosos sobre o processo de construção do modelo e entender quais fatores exercem maior impacto em seu desempenho.

4.2 Metodologia de concepção

Após a definição do objetivo da abordagem, iniciou-se a concepção da mesma. Primeiramente, elaborou-se o processo operacional da xMML-PPP, como ilustrado na Figura 18. Esse processo oferece uma visão geral da abordagem e procura principalmente facilitar a compreensão de cada etapa subsequente, desempenhando um papel fundamental na concepção das etapas posteriores. Seguindo o processo definido, foi possível conceber a arquitetura, que se encontra ilustrada na Figura 17.

A Figura 17, exibe os elementos arquiteturais utilizados pela xMML-PPP. Na arquitetura, destacam-se dois componentes principais: a ferramenta "xMML-PPP Tool" e o repositório de proveniência "xMML-PPP Prov". Ambos desempenham um papel fundamental, por permitirem alcançar o objetivo de gerenciar o ciclo de vida de aprendizado para classificação e capturar os dados de proveniência definidos. Enquanto a "xMML-PPP Tool" possibilita esse gerenciamento e captura, o "xMML-PPP Prov" armazena os dados e, posteriormente, permite a recuperação deles por meio de consultas. Por meio desses componentes da abordagem é possível, então, obter a explicabilidade de dados.

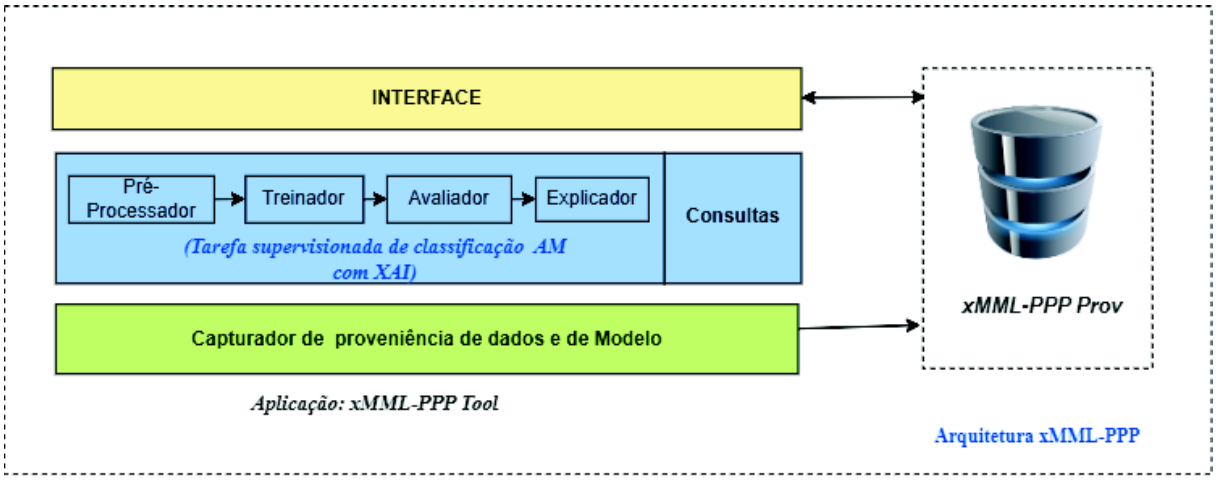

Figura 17 – Arquitetura da xMML-PPP

4.3 Processo da xMML-PPP

Na Figura 18, é possível visualizar uma visão macro das atividades abordadas no processo da xMML-PPP. Tal macroprocesso foi elaborado conforme o modelo de processo de negócio e notação Business Process Model and Notation (BPMN).[1]

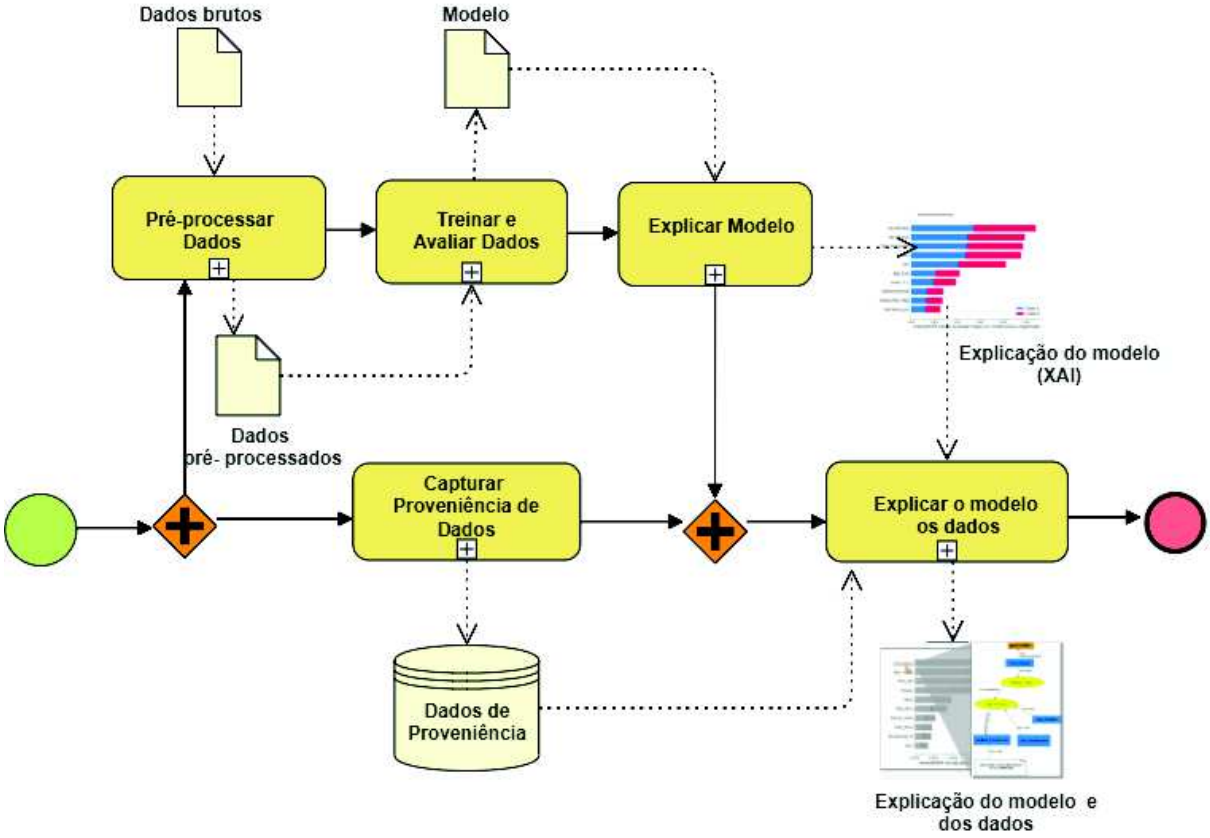

Figura 18 – Macroprocesso xMML-PPP

O processo *"Pré-processamento de dados"*, (Figura 19), começa com as atividades de carregamento e visualização do conjunto de dados, processadas ao longo do fluxo de trabalho. Nesta etapa, a informação do conjunto de dados tais como, nome, número de linhas, número de colunas, tamanho e localização são capturadas. Adicionalmente, destaca-se neste processo a captura de informação das operações de pré-processamento, para cada operação efetuada nesta fase, e o envio dessas informações para o repositório de dados de proveniência. O envio e a captura dessas informações são realizados por atividades integrantes ao processo *"Captura de proveniência de dados"* (Figura 23).

Da mesma forma, são realizadas atividades para relacionar e armazenar a descrição com a informação do atributo, como a etiqueta e o tipo de dados. A descrição do atributo refere-se ao significado de cada atributo. Utilizando o atributo "velocidade do *clock*" de um conjunto de dados que armazena informações sobre computadores de uma empresa, por exemplo, o valor da descrição do atributo seria a "quantidade de ciclos que

[1] https://www.omg.org/spec/BPMN/2.0/

o processador consegue realizar por segundo, determinando a velocidade do computador em Hertz(Hz)". Por outro lado, a informação relativa a este atributo é constituída pelos valores "velocidade do *clock*" e "*String*" para a informação relativa à etiqueta e ao tipo de dados, respectivamente. Este processo está relacionado com as fases 1 e 2 do ciclo de AM apresentado na Figura 24.

No processo de *"Treinamento e Avaliação de Dados"*,(Figura 20), os dados que foram previamente processados são utilizados para treinar o algoritmo selecionado com seus respectivos parâmetros. Durante esta etapa, os dados de treinamento utilizados, bem como os detalhes do algoritmo e seus parâmetros, juntamente com informações sobre o desempenho do modelo treinado, são armazenados no repositório de dados de proveniência pelas atividades paralelas do subprocesso de *"Captura de dados de proveniência"*,(Figura 23). Este subprocesso está diretamente relacionado às fases 3 e 4 do ciclo AM, conforme Figura 24.

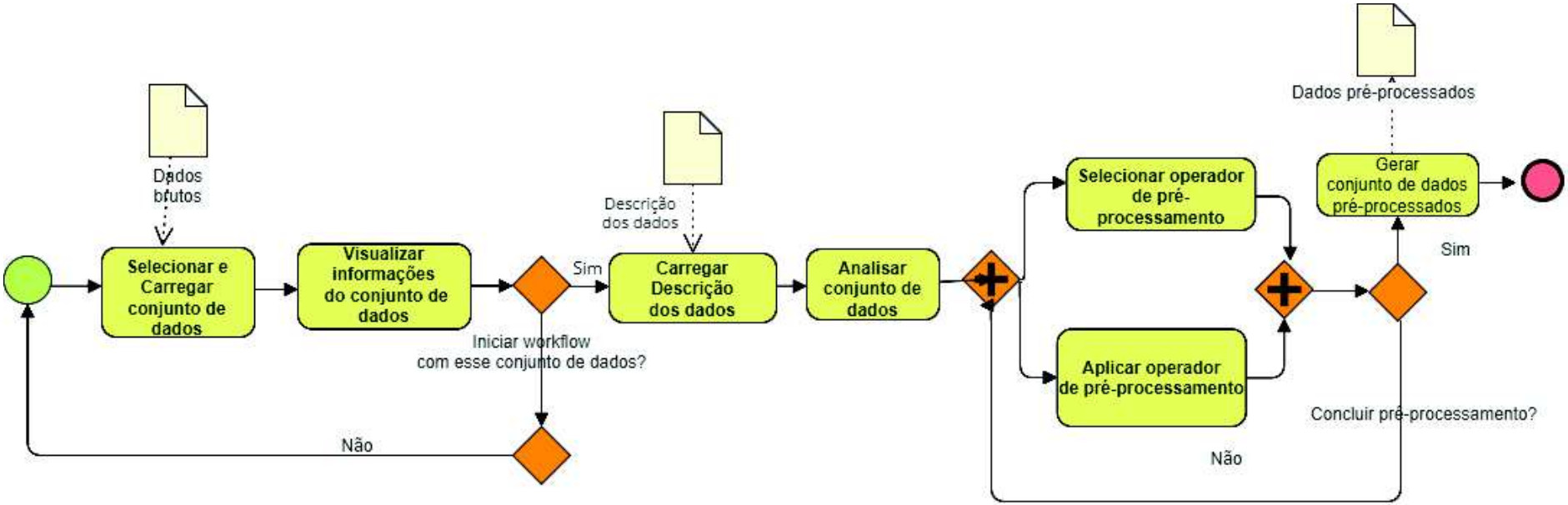

Figura 19 – Processo Pré-processamento de dados

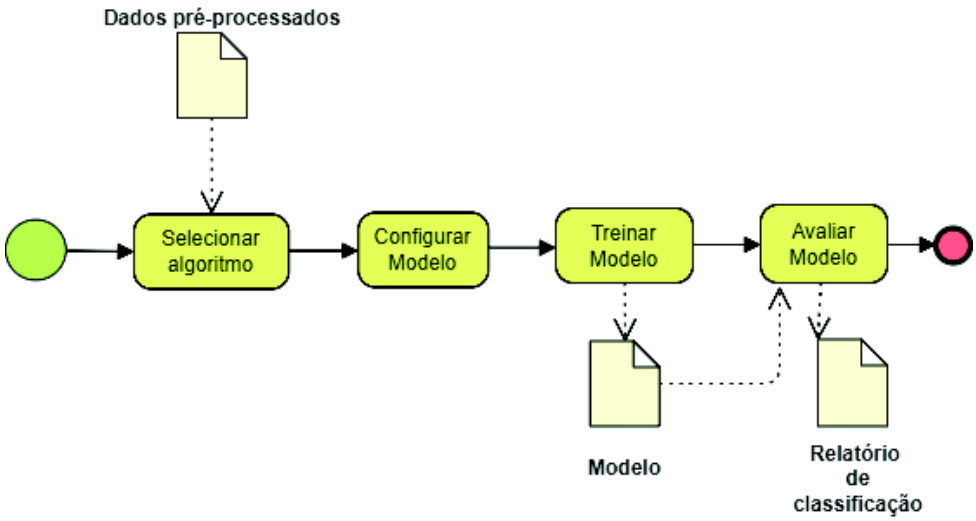

Figura 20 – Processo - Treinamento e Avaliação de dados

No processo *"Explicação do Modelo"* (Figura 21) são realizadas atividades para configurar e gerar a explicabilidade do modelo utilizando ferramentas XAI. Nesta fase, informações como o método XAI utilizado, o conjunto de dados (treinamento ou teste) e os valores de contribuição para cada atributo são enviados para o repositório de dados de proveniência pelas atividades paralelas do sub-processo *"Captura de proveniência de dados"*, conforme Figura 23. Este processo está relacionado com a fase 5 do ciclo AM apresentado na Figura 24.

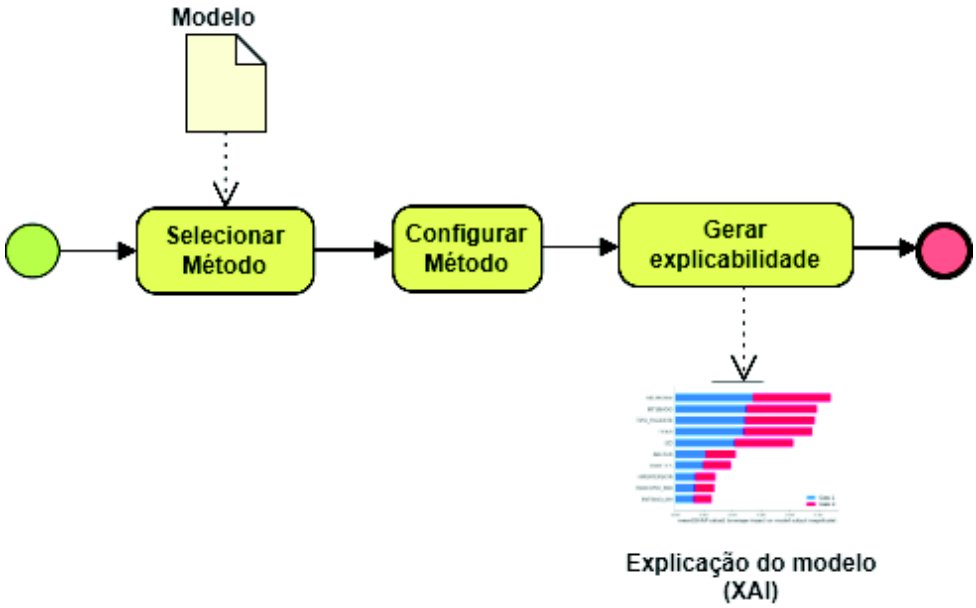

Figura 21 – Processo - Explicação do Modelo

Por fim, no processo denominado *"Explicação do Modelo e dos Dados"*, ilustrado na (Figura 22), e após a realização da explicação do modelo e a obtenção da contribuição de cada atributo, ocorre a recuperação dos dados previamente armazenados no repositório de proveniência. Por meio de consultas específicas, é possível obter informações relevantes

sobre o tratamento que cada atributo recebeu, a fim de se obter uma compreensão mais clara das operações que contribuíram para o resultado obtido. É importante destacar que este processo é fundamental para complementar a transparência e a confiabilidade do modelo de AM desenvolvido.

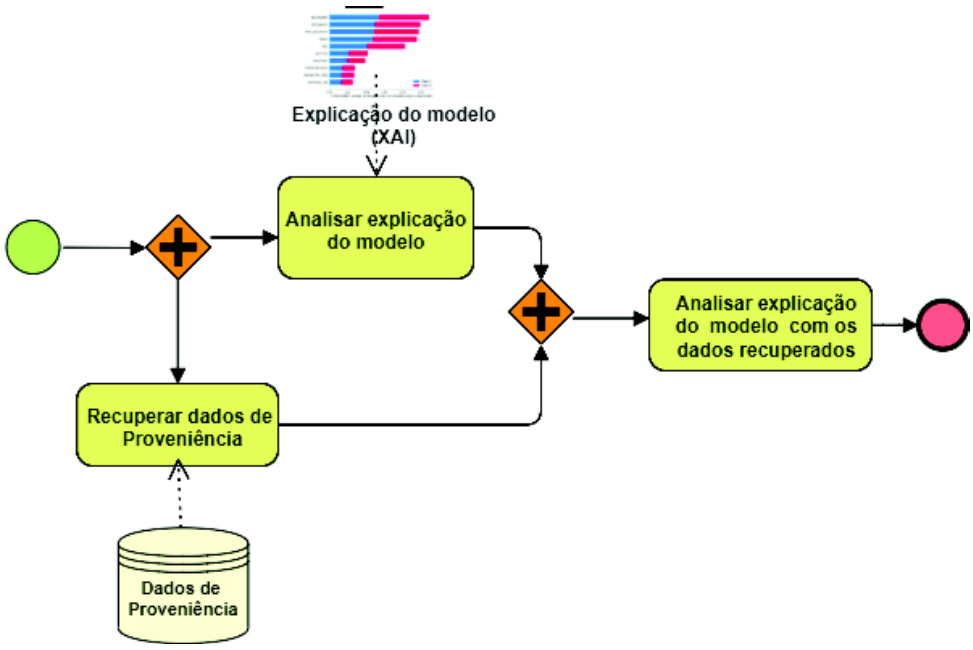

Figura 22 – Processo - Explicação do Modelo e dos Dados

A Figura 23 ilustra o processo de captura de proveniência de dados, conforme definida no processo da xMML-PPP. As atividades deste processo, *Capturar informações do conjunto de dados*, *Capturar informações das operações de pré-processamento*, *Capturar informações de treino* e *Capturar informações da explicabilidade*, são realizadas em momentos distintos, porém, da mesma forma, passam pelo processo de armazenamento para o repositório de proveniência.

Na Figura 24 é ilustrado o ciclo de vida básico de AM, complementado pela fase de explicabilidade do modelo fornecida pela técnica XAI e também pela explicabilidade de dados. É importante ressaltar que a explicabilidade do modelo ocorre após a conclusão do treinamento e avaliação do modelo, visando compreender seus resultados. Por outro lado, a explicabilidade de dados (destacada em cinza) ocorre após a conclusão de todo o ciclo, permitindo obter-se conhecimento tanto dos dados quanto da explicabilidade fornecida pelo modelo.

4.4 Modelo Conceitual

Esta seção apresenta o conceito das entidades utilizadas na abordagem.

- **Workflow:** O fluxo de trabalho (*workflow*) é a base de todo o processo. Ele representa

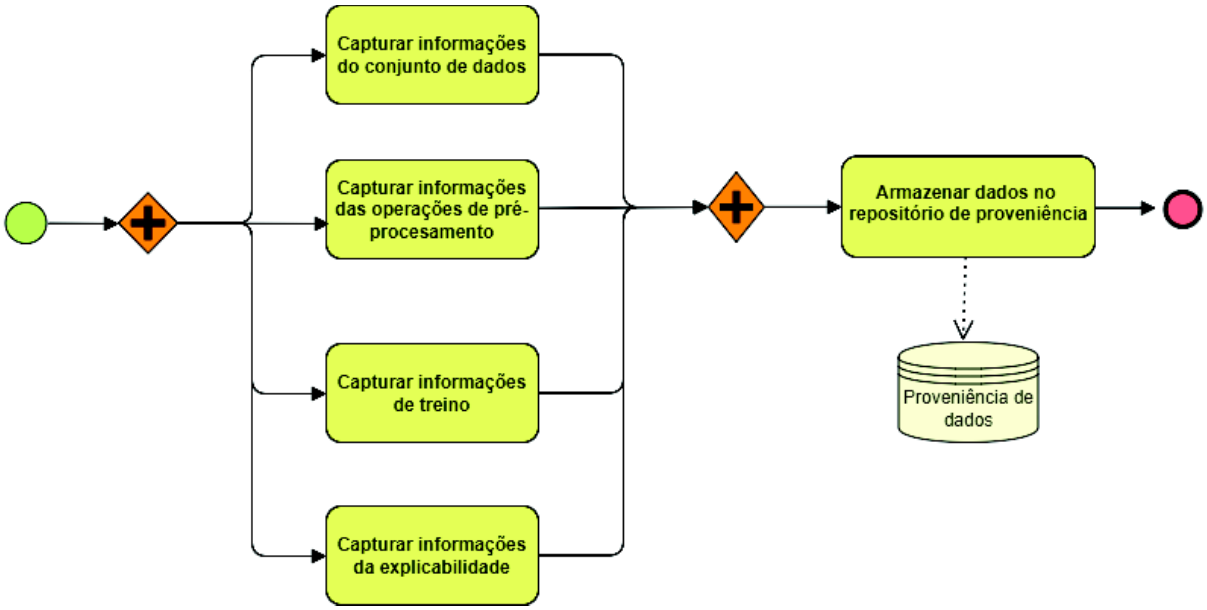

Figura 23 – Processo - Captura de Proveniência de Dados

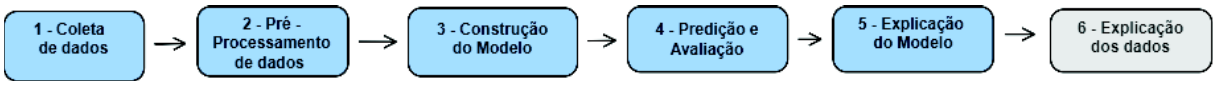

Figura 24 – Ciclo de vida de aprendizado de máquina com fase de explicabilidade de modelo e de dados

um processo que será realizado seguindo uma sequência específica de etapas. Cada *workflow* é identificado por um rótulo (ou nome) e é iniciado em um horário e data determinados. Cada *workflow* lida com um único conjunto de dados, o qual é carregado e processado ao longo de todo o fluxo de trabalho. As informações do conjunto de dados original são extraídas e, tem-se, dessa forma, informações de origem do conjunto de dados antes deste ser pré-processado.

- **Dataset:** Trata-se do conjunto de dados que será processado durante o fluxo de trabalho. Ele contém informações sobre onde os dados estão armazenados, a dimensão dos dados e outras características relevantes. O conjunto de dados é a matéria-prima para os experimentos, sendo processado conforme as etapas definidas no *workflow*.

- **Experimento:** Um Experimento é realizado em um workflow. Ele contém informações detalhadas sobre as configurações específicas desse experimento. Isso inclui detalhes como o algoritmo utilizado e os parâmetros específicos para o algoritmo, bem como os respectivos valores associados. Os experimentos podem gerar resultados ou *insights* que contribuem para o entendimento dos dados.

- **Parâmetros:** Referem-se as configurações específicas do experimento, segundo o método (algoritmo) escolhido. Dependendo da configuração dos parâmetros, o resultado (desempenho) do experimento pode ser influenciado.

- **Experimento_Atributos:** Refere-se aos atributos derivados para o treinamento de AM. Os atributos são essenciais para o modelo poder aprender padrões nos dados e generalizar esse conhecimento para realizar tarefas em novos dados. Os atributos do experimento são resultantes das operações de pré-processamento que os dados do conjunto de dados sofreram. No entanto, é válido destacar que as manipulações aplicadas aos dados do segundo experimento, por exemplo, dentro do mesmo fluxo de trabalho se acumulam junto às manipulações realizadas nos dados do primeiro experimento, já que os dados passam por transformações contínuas ao longo de todo o processo (workflow).

- **XAI:** A entidade XAI refere-se à descrição detalhada do experimento realizado dentro do *workflow*. Ela inclui informações que ajudam a compreender os resultados e *insights* obtidos a partir dos experimentos. As configurações usadas para a realização da explicação também são incluídas aqui, permitindo que outros usuários compreendam o processo e os resultados do experimento.

- **Operações de Pré-processamento:** As Operações de Pré-processamento são as transformações realizadas nos atributos do conjunto de dados, visando preparar os dados para a realização dos experimentos. Isso envolve o tratamento dos dados para torná-los adequados para treinamento.

- **Operadores** São as técnicas utilizadas visando efetuar as operações nos dados. Esses operadores podem incluir a transformação nos dados (como normalização, padronização, codificação de dados), limpeza de dados ausentes, além de criação ou exclusão de atributos. As operações de pré-processamento utilizam os operadores para realizar a transformação nos dados.

- **Atributos do conjunto de dados:** Os Atributos do conjunto de dados são as características individuais que o compõem. Eles têm informações associadas, como rótulo e tipo. Esses atributos são usados como entradas para os experimentos, sendo manipulados e processados conforme necessário. Estes atributos podem ser numéricos (contínuos), categóricos (Nominais ou ordinais), podem ser também derivados (calculados a partir de atributos originais do conjunto de dados). À medida que os atributos são pré-processados, as características do conjunto de dados, constituído por esses atributos, vão sendo modificadas.

Ao realizar uma associação com os elementos das principais estruturas do modelo conceitual da xMML-PPP com o modelo de representação de dados do W3C (Figura 2), visto na seção 2.1, do Capítulo 2, temos:

- **Agente:** Na xMML-PPP o agente é o workflow. Trata-se, neste caso, de um tipo específico de entidade, responsável pelas atividades de transformação dos dados do

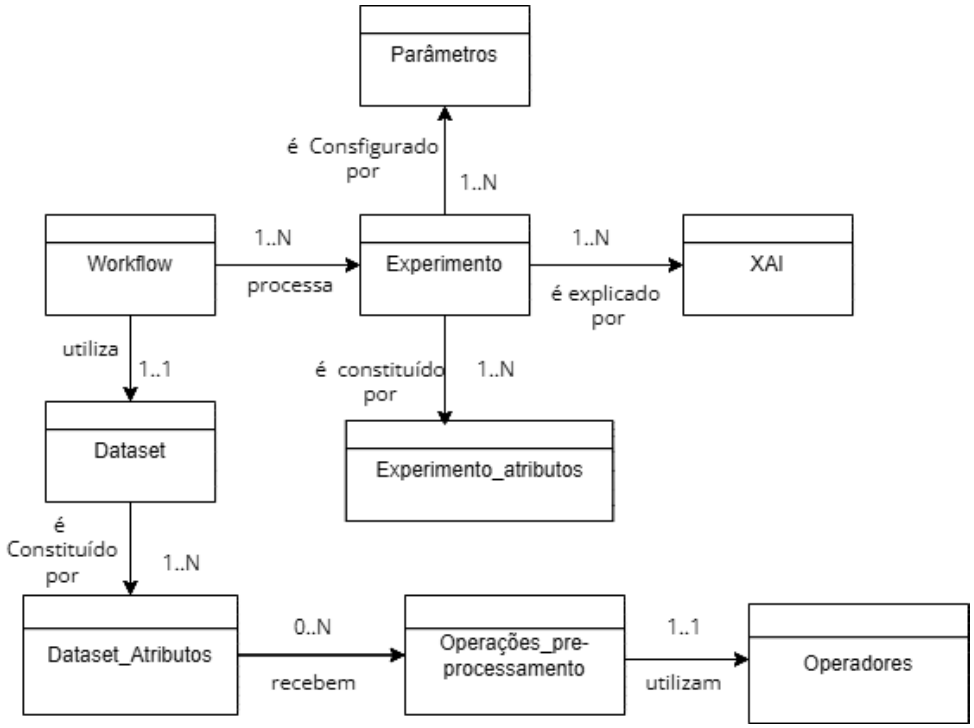

Figura 25 – Diagrama conceitual abordagem xMML-PPP

conjunto de dados, bem como pela execução dos experimentos e pela explicação destes.

- **Entidade:** O conjunto de dados é uma entidade do tipo *Collection*, a qual é constituída por outras entidades, os atributos. Os experimentos e os parâmetros também são tipos de entidades.
- **Atividade**: As operações de pré-processamento são atividades que agem sobre os dados os transformando, podendo, também, criar outros atributos derivados. Além das operações de pré-processamento, o XAI também é uma atividade, a qual usa os dados dos experimentos, gerando uma entidade de Explicação deste.

As Figuras 26 e 27 apresentam uma representação de parte dos elementos conceituais da xMML-PPP, juntamente com sua correspondência em relação aos elementos principais da estrutura do W3C *(Agent, Entity e Activity).*

O *workflow* corresponde ao *Agent*, que é do tipo denominado *SofwareAgent*. O *dataset*, corresponde à *Entity*, que se configura como um tipo de *Collection*, composta por atributos distintos. Os operadores correspondem às atividades (*Activity*), as quais são executadas no *dataset* pelo *workflow*.

O XAI, corresponde à Atividade (*Activity*), responsável por realizar a explicação do experimento. Por sua vez, o experimento é uma Entidade (*Entity*), a qual possui as características referentes aos experimentos executados.

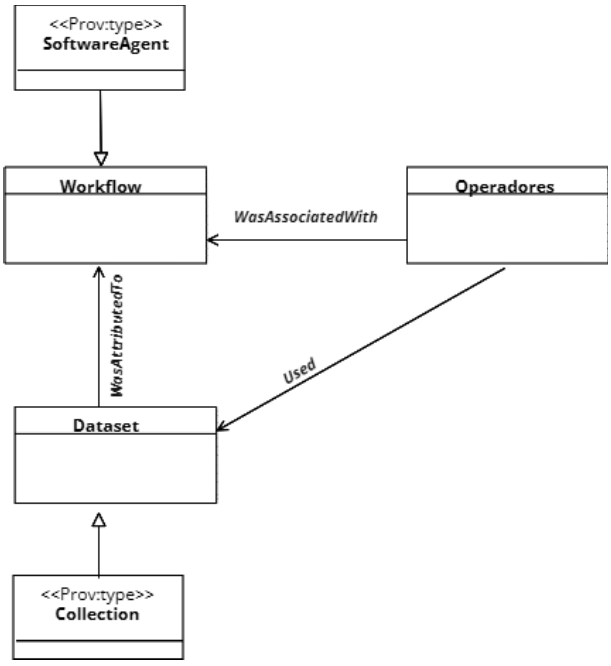

Figura 26 – Elementos Conceituais xMML-PPP (Workflow, Operadores e Dataset) e as principais estruturas Prov - W3C.

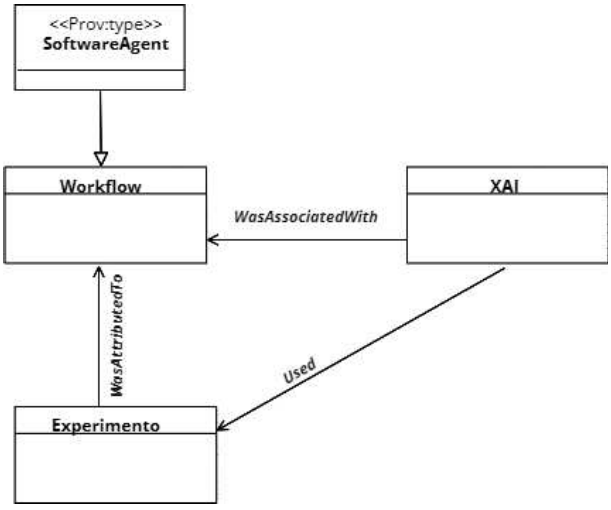

Figura 27 – Elementos Conceituais xMML-PPP (Workflow, XAI e Experimento) e as Principais estruturas Prov - W3C.

5 IMPLEMENTAÇÃO DA XMML-PPP

Neste capítulo são apresentados os artefatos gerados para a implementação da arquitetura da xMML-PPP. A Figura 28, a qual é uma extensão da Figura 17, acrescenta na arquitetura as respectivas ferramentas utilizadas em cada componente. Os dois componentes da arquitetura da abordagem xMML-PPP são: a ferramenta "xMML-PPP Tool" e o repositório de proveniência, "xMML-PPP Prov".

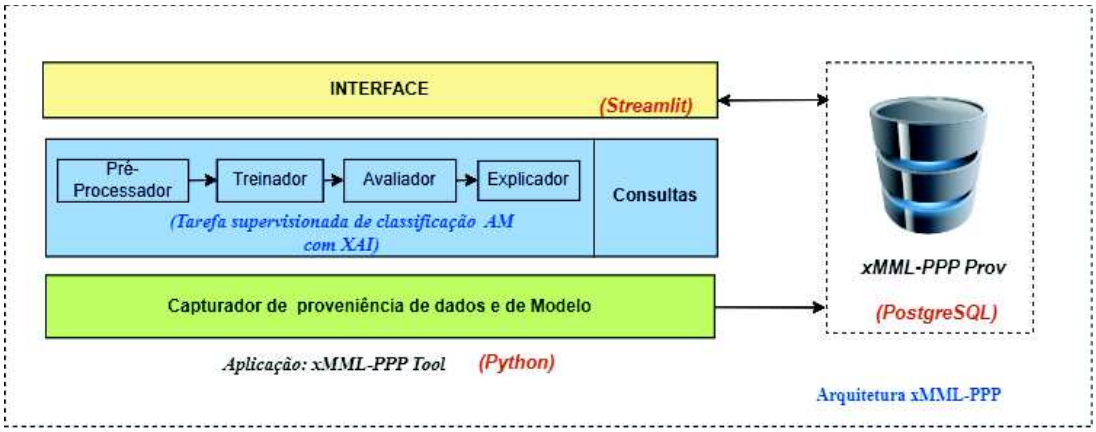

Figura 28 – Componentes arquiteturais da xMML-PPP

Na sequência, são apresentados os detalhes de implementação de cada componente da arquitetura da abordagem. Dessa forma, na seção 5.1 é apresentada a implementação da ferramenta e na seção 5.2 é apresentada a implementação do repositório.

O processo de desenvolvimento da ferramenta iniciou-se com a elaboração do documento de escopo. Esse documento desempenha a função de descrever as metas e objetivos a serem atingidos. Uma vez que o documento de escopo foi elaborado e os requisitos foram definidos, a codificação da ferramenta foi iniciada, considerando esse objetivo e funcionalidades previamente estabelecidos.

Dessa forma, estão definidos no documento de escopo:

- **Objetivo:** implementar as funcionalidades de transformação de dados brutos em dados pré-processados e treinados, utilizando algoritmos de tarefa supervisionada de classificação em aprendizado de máquina, com suporte à explicabilidade do modelo treinado e à explicabilidade dos dados utilizados para o treino, com vistas a fornecer o suporte para o atendimento ao objetivo da XMML-PPP.
- **Escopo:** lista com as seguintes funcionalidades, com vistas a atender ao objetivo da abordagem, conforme detalhamento do processo especificado na seção 4.3:
  - Carregar e explorar dados carregados:

    - Exibir informações do conjunto de dados, tais como, dimensão, tipos e nomes das colunas;
    - Calcular e exibir informações estatísticas de colunas quantitativas, tais como, média, desvio padrão, valores mínimo e máximo e medidas interquartis;
    - Apresentar informações detalhadas do conjunto de dados como tipo de coluna, número de dados faltantes por coluna, percentual de dados faltantes por coluna, e quantidade de informações únicas por coluna; e
    - Plotar gráfico de histograma para visualizar a distribuição dos valores e identificar desbalanceamento na coluna alvo, visualizar gráficos do tipo *boxplot* para identificação da presença de *outliers*, e também de matriz de correlação para análise de correlação entre as colunas quantitativas.
- Preparar o conjunto de dados para treino:
    - Realizar a limpeza dos dados: imputar valores ausentes. Para colunas qualitativas imputar como desconhecido, e para colunas quantitativas permitir imputar a média, a mediana, a moda, dentre outros valores pré-definidos;
    - Realizar a redução de dados: permitir selecionar quais colunas a serem excluídas;
    - Realizar a transformação de dados: para colunas do tipo quantitativo, permitir a normalização dos dados ou a padronização dos dados, e codificar dados, para colunas do tipo qualitativo;
    - Realizar a construção de atributos: permitir construir novos atributos (atributos derivados) a partir de atributos já existentes no conjunto de dados;
    - Realizar o particionamento dos dados: permitir realizar a divisão do conjunto de dados em treino e teste; e
    - Realizar a correção da prevalência dos dados: permitir a correção na distribuição dos registros. Para a realização da correção das amostras, permitir a utilização de técnicas como *oversampling* e *undersampling*.
- Treinar o conjunto de dados pré-processado:
    - Permitir a escolha do algoritmo a ser usado; e
    - Permitir a configuração dos parâmetros de treino.
- Acompanhar o resultado de desempenho do modelo:
    - Permitir a visualização a matriz de confusão para exibição da distribuição dos registros em termos de suas classes atuais e de suas classes previstas; e
    - Calcular e exibir as medidas de desempenho de acurácia, precisão, revocação e medida F1 da base de teste.

- Gerar o gráfico de explicabilidade do modelo treinado:
  - Permitir a escolha do método de explicabilidade de atributos a ser usada; e
  - Exibir o gráfico de explicabilidade para o modelo treinado.
- Suportar a proveniência retrospectiva:
  - Permitir a captura de dados e armazenamento no repositório de dados, os dados relativos à estrutura dos dados do fluxo de trabalho;
  - Permitir a captura de dados e armazenamento no repositório de dados, os dados relativos a cada operador de pré-processamento executado; e
  - Permitir a captura de dados e armazenamento no repositório de dados, os dados relativos ao resultado do modelo e a explicação do resultado do modelo provida por técnica XAI.
- Realizar consultas ao repositório de dados
  - Permitir realizar consultas pré-determinadas no repositório de dados para recuperação das informações necessárias a análise dos dados de proveniência; e
  - Permitir incluir comandos Structured Query Language (SQL) para realização de consultas próprias no repositório de dados para recuperação das informações necessárias a análise dos dados de proveniência.

A concepção do repositório de proveniência ocorreu simultaneamente ao desenvolvimento da ferramenta, visando atender às demandas de armazenamento dos dados de proveniência. O repositório desempenha a função de armazenar os dados de proveniência, tanto prospectivos quanto retrospectivos, capturados pela ferramenta da abordagem. Esses dados são essenciais para atender ao objetivo de rastreamento e armazenamento dos dados de origem e o histórico dos dados ao longo do processo.

A xMML-PPP considera a captura de dados relativos a cada fase do ciclo de vida de AM, no entanto, destaca-se a captura dos registros relativos à captura das operações de pré-processamento, que contribuem para a explicabilidade dos dados.

A concepção do modelo lógico foi realizada com vistas a atender ao objetivo e funcionalidades definidos no Documento de escopo, tendo como saída a representação gráfica das propriedades das entidades registradas pelo componente de captura de proveniência. Dessa forma, com base no modelo lógico, foi possível prosseguir com a implementação do esquema de dados por meio do Sistema Gerenciador de Banco de Dados Relacional (SGBDR) definido.

5.1 Ferramenta "xMML-PPP Tool"

A "xMML-PPP Tool" foi desenvolvida conforme os objetivos e especificidades da xMML-PPP para a sua validação e, assim, implementa as atividades descritas no documento de escopo. A ferramenta foi desenvolvida na linguagem Python (64), com o framework *"Streamlit"*.[1] O Streamlit é uma biblioteca Python para criar aplicações web sem a necessidade de codificar o seu *frontend*. Dessa forma, a atividade de implementação da ferramenta baseia-se nas ferramentas expostas e seu código-fonte está disponível na plataforma de hospedagem GitHub.[2]

A escolha da linguagem Python considera a amplitude de funções desta linguagem de programação, o que permite atender a todas as funcionalidades necessárias para a concepção da ferramenta da abordagem. Adicionalmente, também é considerada a sua vasta utilização na área de Ciência de Dados. O repositório de dados utilizado pela ferramenta foi implementado utilizando o SGBD de código aberto PostgreSQL (65) e codificado com a biblioteca de mapeamento objeto-relacional SQLAlchemy.[3]

A escolha do Banco de Dados relacional é justificada pelo fato de que essa tecnologia já faz parte do nosso domínio de conhecimento, o que resultou em uma implementação mais rápida e eficiente. Esta decisão permitiu que evitássemos a necessidade de dedicar tempo significativo ao estudo de uma nova tecnologia, o que, por sua vez, nos possibilitou concentrar nossos esforços em aprender e aplicar tecnologias novas, como as ferramentas de XAI, que são essenciais para o desenvolvimento da nossa abordagem.

A Figura 29 apresenta a tela inicial da "xMML-PPP Tool". Por intermédio do menu lateral esquerdo dá-se a entrada do arquivo referente à base de dados, onde é iniciado o fluxo de trabalho. Além disso, as opções de funcionalidades da ferramenta também são disponibilizadas para dar sequência ao trabalho, tais como exploração de dados, pré-processamento, treino e consultas aos dados.

5.1.1 Funcionalidades da ferramenta "xMML-PPP Tool"

A Tabela 4 condensa as principais funcionalidades implementadas na ferramenta "xMML-PPP Tool", de modo a permitir o atendimento dos requisitos das explicabilidades de modelo e dados, objetivo da xMML-PPP.

[1] https://https://streamlit.io/
[2] Código da Ferramenta: <http://github.com/RosanaLeandro/ppm-ml>
[3] https://www.sqlalchemy.org/

Figura 29 – Tela inicial da ferramenta "xMML-PPP Tool"

Tabela 4 – Funcionalidades da ferramenta

Funcionalidade	**Descrição**	**Exemplo**
Explorar dados	Visualizar gráficos e tabelas para entender os dados	Criar um gráfico de barras para mostrar a distribuição das classes
Pré-processar dados	Limpar, reduzir, transformar e construir atributos dos dados	Remover colunas com valores ausentes e criar um novo atributo derivado de outros atributos
Treinar conjunto de dados	Treinar um modelo de aprendizado de máquina nos dados	Treinar um modelo de AM para realizar previsões
Calcular métricas de desempenho	Avaliar se o modelo treinado obteve bom desempenho	Calcular a acurácia e a medida F1 do modelo treinado
Gerar explicabilidade do modelo	Entender como o modelo está tomando suas decisões	Usar o método XAI para mostrar a importância de cada atributo no modelo
Capturar proveniência	Registrar informações sobre o pré-processamento, treinamento e explicabilidade do modelo	Armazenar o nome do conjunto de dados usado, o algoritmo de treinamento usado e as métricas de desempenho calculadas
Consultar proveniência	Acessar informações sobre a origem dos dados e como eles foram processados	Verificar de onde vieram os dados usados para treinar o modelo e como eles foram pré-processados

5.2 Repositório de Proveniência "xMML-PPP Prov"

O repositório de proveniência da ferramenta é o componente da arquitetura xMML-PPP referente ao armazenamento dos dados estruturados capturados pela ferramenta desenvolvida para validação da abordagem. Seu foco é armazenar a proveniência retrospectiva dos dados trabalhados na abordagem, os quais são dados relativos: à especificação da estrutura do fluxo de dados, a cada operador de pré-processamento executado, à execução dos experimentos, e a explicabilidade dos modelos do experimento, a fim de que

a abordagem alcance ao seu objetivo.

Dessa forma, para o desenvolvimento da camada de captura de dados de proveniência torna-se necessário o projeto do seu modelo lógico (Figura 30). Este modelo especifica a representação de dados relacionados com a estrutura do fluxo de dados, e o fluxo de execução das operações de pré-processamento, que corresponde à proveniência retrospectiva.

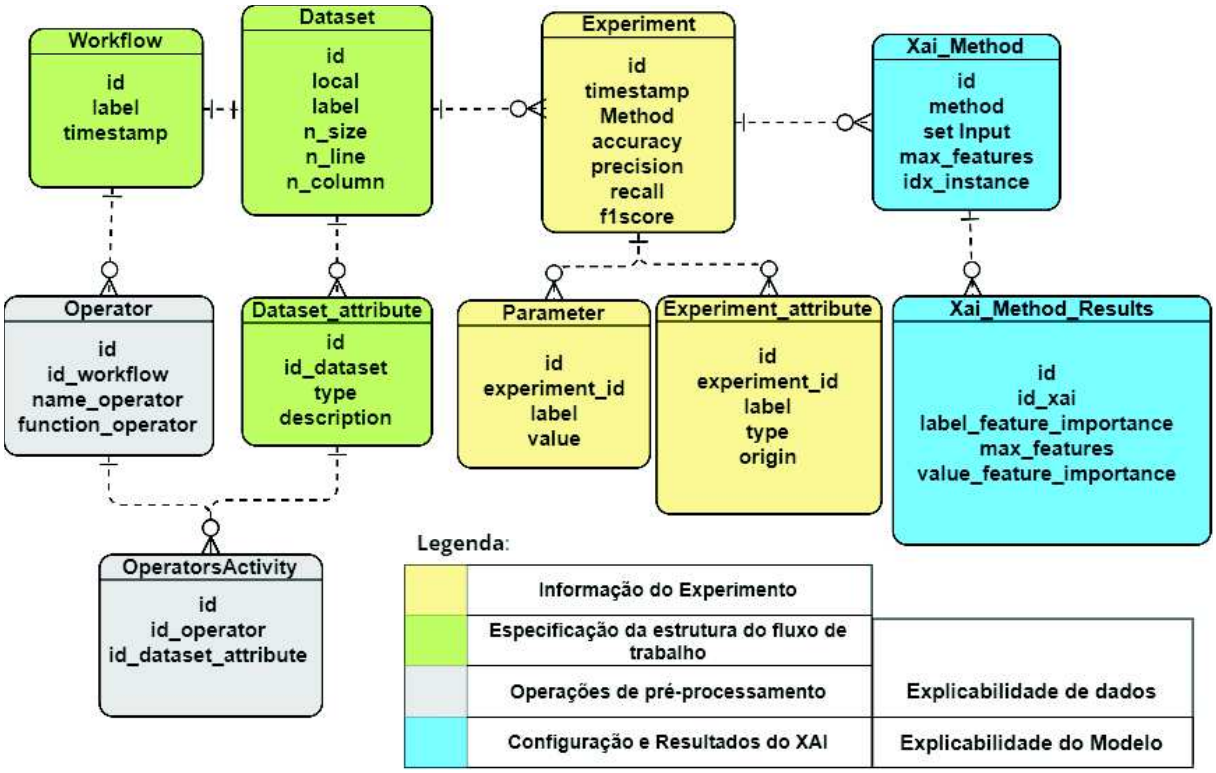

Figura 30 – Modelo de dados abordagem xMML-PPP

As entidades destacadas em verde (*Workflow*, *Dataset*, *Dataset_attribute*) têm informações básicas relacionadas com o fluxo de trabalho, o conjunto de dados e as suas informações de atributos originais. Já as entidades destacadas em cinza (*Operator*, *OperatorsActivity*) referem-se às atividades das operações de pré-processamento realizadas no conjunto de dados. As entidades destacadas em amarelo (*Experiment*, *Parameter*, *Experiment_attribute*) referem-se à informação dos experimentos realizados no fluxo de trabalho, para armazenar a proveniência de cada configuração do experimento e os respectivos resultados. Por último, as entidades destacadas em azul (*Xai_Method, Xai_Method_Results*) referem-se à configuração XAI e às suas informações de resultados. A descrição das entidades e seus atributos encontram-se no Apêndice A.

A atividade de implementação do modelo lógico do repositório de proveniência, conforme apresentado na Figura 30, resulta na criação dos objetos de dados no repositório de proveniência. Com isso, é possível criar o esquema físico e a coleção de objetos no banco de dados PostgreSQL.

É importante ressaltar que, para a implementação física dos objetos resultantes do

modelo lógico, houve a consolidação das entidades referentes às operações e operadores do pré-processamento. Em vez de armazenar os registros de proveniência das entidades representadas *Operator* e *OperatorsActivity* em tabelas separadas, optou-se por consolidá-los em uma única tabela. Essa medida foi adotada para otimizar o armazenamento das informações sobre a proveniência dos operadores de pré-processamento utilizados, além de simplificar as consultas e reduzir a complexidade. Assim, os registros dos dados de proveniência das entidades representadas no modelo lógico são consolidados nas tabelas *Workflow, Dataset, Experiment, Experiment_attribute, Dataset_attribute, Parameter, OperatorsActivity, Xai_Method e Xai_Method_results.*

6 APLICAÇÃO DA XMML-PPP

Este capítulo apresenta os resultados obtidos utilizando a xMML-PPP para complementar a explicabilidade em AM por intermédio da utilização da proveniência de dados. A adoção da xMML-PPP para analisar o tratamento que os dados recebem antes do treino constitui em um histórico que permite o enriquecimento da explicabilidade do modelo, com as análises do comportamento do modelo diante das alterações nos dados que foram utilizados para derivar o modelo.

Dessa forma, para avaliação da abordagem, foram efetuados experimentos com dois conjuntos de dados distintos. O primeiro é a base tradicional do Titanic (66), um navio que se afundou em 1912 na sua primeira viagem. Esta base de dados foi utilizada como exemplo didático para ilustrar como as operações de pré-processamento podem afetar o desempenho do modelo.

Já o segundo conjunto de dados é derivado de um problema atual, o que permite demonstrar uma aplicação real dessa abordagem. Devido à recente epidemia de SARS-CoV-2 (COVID-19), foram realizados muitos estudos visando aplicar técnicas de AM à resolução de problemas relacionados. Alguns deles visavam compreender os fatores de risco da doença, prever uma possível contaminação ou mesmo uma possível taxa de mortalidade devido ao vírus. Os dados demográficos e as condições de saúde pré-existentes permitem personalizar modelos preditivos para prever se um paciente será hospitalizado, virá a óbito ou necessitará de cuidados de saúde intensivos (67).

O conjunto de dados original utilizado no nosso segundo estudo foi produzido e disponibilizado pelo Governo Mexicano (68). No entanto, no presente trabalho, foi empregada uma versão modificada do mesmo conjunto de dados, também utilizada no trabalho de Muhammad et al.(69), que possui o seu dicionário já traduzido para o inglês e refere-se ao período de observação de 12 de abril de 2020 a 3 de junho de 2020. Este conjunto de dados está disponibilizado em (70).

Assim, a segunda base de dados consiste em informações epidemiológicas de pacientes do México suspeitos de contaminação pelo vírus SARS-CoV-2, os quais foram submetidos a um teste PCR-RT[1]. Esta base foi utilizada para um caso de dados da área de saúde, com dados recentes, que reflete melhor a necessidade da utilização da explicabilidade.

Em ambos os conjuntos de dados, são efetuados dois experimentos derivados. No primeiro experimento, é efetuado apenas o mínimo de pré-processamento necessário

[1] É um diagnóstico laboratorial, feito por biologia molecular, que permite identificar a presença do material genético (RNA) do vírus Sars-Cov-2 em amostras de secreção respiratória (71).

para aplicação dos algoritmos. No segundo experimento, por outro lado, as operações de pré-processamento são expandidas com a intenção de melhorar os resultados dos modelos.

Em todos os experimentos foi utilizado o algoritmo RF implementado utilizando-se a biblioteca scikit-learn (52). O RF foi selecionado, por já ser utilizado em outros estudos sobre COVID-19 (72, 73) e ter obtido bons resultados, além de ter apresentado bom desempenho em estudos anteriores com a base de dados do Titanic.

Para todos os experimentos, foram usados os seguintes parâmetros do RF: *_estimators*, número de árvores da floresta; *max_features*, número de atributos a serem consideradas ao procurar cada melhor divisão; *min_samples_split*, número mínimo de amostras necessárias para dividir um nó interno; *min_samples_leaf*, número de amostras necessárias para compor um nó folha; e *max_depth*, profundidade máxima da árvore.

A configuração dos valores dos parâmetros diferiu em cada estudo de caso, mas foi mantida constante em ambos os experimentos de cada estudo de caso para garantir que a alteração de desempenho ocorresse somente devido à mudança nas operações de pré-processamento.

Para comparar os resultados dos modelos, foram utilizadas as métricas de acurácia, precisão, revocação e medida-F1. A precisão avalia a proporção de observações positivas que são verdadeiramente positivas. Já a revocação mede a proporção de verdadeiros positivos em relação a todas as observações positivas no conjunto de dados. A medida-F1 é calculada como a média harmônica entre a precisão e a revocação. Em todas as métricas, foi utilizada a média "ponderada" para as classes. Esse cálculo é feito para cada classe e, em seguida, é calculada a média ponderada pelo suporte (ou seja, o número de instâncias verdadeiras para cada etiqueta). A média ponderada considera o desequilíbrio dos rótulos, diferentemente do método de média aritmética simples (também conhecido como "macro"). Como resultado, a medida-F1, quando calculada com essa média, pode não se situar exatamente entre os valores de precisão e revocação.

6.1 Experimentos com a Base de Dados Titanic

O conjunto de dados do Titanic contém 12 colunas e 891 instâncias, e disponibiliza os seguintes atributos: PassengerId, Pclass, *Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabine, Survived* e *Embarked*. O dicionário de dados referente a este conjunto de dados está disponível no Apêndice B. O objetivo nesse estudo é criar um modelo preditivo que responda à pergunta: "Quais tipos de pessoas têm mais probabilidade de sobreviver?".

Para responder essa pergunta, foram usados os dados disponíveis dos passageiros como, por exemplo, idade, sexo, classe econômica, etc. O atributo alvo para este caso de estudo é o *Survived*. Para o primeiro experimento nesta base foram realizadas as

seguintes operações básicas de pré-processamento, de forma que o RF pudesse ser executado: preenchimento de valores nulos, transformação e exclusão de atributos categóricos. Assim, os atributos nominais *Sex* e *Embarked* foram codificados usando o operador *OneHotEnconder*. Para o preenchimento dos valores nulos para o atributo *Embarked* foi utilizada a sua moda e para o atributo *Age* foi utilizada a sua média. Os atributos *PassengerId*, *Ticket* e *Cabine* foram removidos, pois representavam apenas possíveis identificações dos passageiros.

No segundo experimento, foram realizadas as mesmas operações do primeiro experimento, que consistiram no preenchimento de valores nulos e na transformação de atributos categóricos. Adicionalmente, conduziu-se uma análise exploratória prévia dos dados com o objetivo de obter uma compreensão inicial do conjunto de dados e identificar possíveis padrões. Durante a análise, foram identificadas algumas características prevalentes nos dados, tais como: pessoas do sexo feminino sobreviveram mais do que as do sexo masculino, pessoas da classe 1 sobreviveram mais do que das classes 2 e 3, dentre outras.

Assim, para este experimento, as seguintes operações de pré-processamento foram realizadas. O atributo *Title* foi criado extraindo-se o título junto ao nome do passageiro. Para este atributo foram mantidos apenas os valores *Master*, *Miss*, *Mr.*, *Mrs.*, que apresentavam uma maior frequência. As instâncias, cujos valores diferiam destes, tiveram o valor deste atributo substituído por "*Others*".

Também foi criado o atributo *Lastname*, que extrai o sobrenome dos nomes dos passageiros e contribuiu para a engenharia de outro atributo criado *Groupsize*, que se refere ao número de mulheres ou crianças com o mesmo sobrenome e, provavelmente, da mesma família. Outro atributo criado foi o *Family_Size*, que representa a soma do quantitativo de toda a família a bordo do Titanic, derivado dos atributos *Parch* e *SibSp*, de cada instância.

Além disso, os atributos categóricos (*Embarked*, *Title* e *Sex*) foram codificados com o operador *OnehotEncoder*. Os atributos *Name*, *Ticket* e *Lastname* foram excluídos e os valores dos atributos foram normalizados com o uso do operador *MinMaxScaler*. Os valores dos parâmetros usados na configuração do RF foram os seguintes: *max_features*: sqrt[2]; *min_samples_split*: 2; *min_samples_leaf*: 2; *max_depth*: 4; e *num_estimators*: 100.

A Tabela 5 mostra o desempenho do RF nos dois experimentos realizados com a base de Dados do Titanic.

Tabela 5 – Resultados de Experimentos - Base do Titanic

Id	**Acurácia**	**Precisão**	**Revocação**	**Medida F1**
1	81,72	82,48	81,72	81,17
2	84,33	84,32	84,33	84,20

Pode-se observar que, mesmo com poucas operações de pré-processamento realizadas no primeiro experimento, o RF conseguiu um bom desempenho na tarefa. Já no

[2] Indica que é utilizada a raiz quadrada da quantidade de atributos em cada árvore da floresta

segundo experimento, com mais operações de pré-processamento, incluindo a construção de atributos, percebe-se uma melhora ainda maior no seu desempenho. Na Tabela 6 é apresentado o relatório de classificação do experimento 1 da base de dados do Titanic e na Tabela 7 o relatório de classificação do experimento 2 desta mesma base.

Tabela 6 – Relatório de Classificação - Base do Titanic - Experimento 1

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0	79,00	93,00	86,00	157
1	87,00	66,00	75,00	111
Acurácia			82,00	268
Média Macro	83,00	79,00	80,00	268
Média Ponderada	82,00	82,00	81,00	268

Tabela 7 – Relatório de Classificação - Base do Titanic - Experimento 2

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0	84,00	90,00	87,00	157
1	84,00	77,00	80,00	111
Acurácia			84,00	268
Média Macro	84,00	83,00	84,00	268
Média Ponderada	84,00	84,00	84,00	268

Conforme evidenciado nos relatórios de classificação, o experimento 2 demonstra uma acurácia superior, como já apresentado na (Tabela 5). Adicionalmente, pode realizar uma classificação mais eficiente para ambas as classes em geral.

6.2 Experimentos - Base de dados Covid-19 México

Esse conjunto de dados possui 41 colunas e 263.007 instâncias contendo dados demográficos como idade, gênero, nacionalidade, se o paciente é imigrante, além de dados clínicos, de doenças pré-existentes do paciente como diabetes, asma, hipertensão, obesidade entre outras doenças, informações de gravidez, tabagismo e datas, como a data de início de sintomas, data de admissão na unidade de saúde e possível data do óbito, além do resultado do teste PCR-RT para a doença. O objetivo dos experimentos neste estudo é classificar a mortalidade de casos confirmados de COVID-19, por intermédio do resultado do teste PCR-RT. O dicionário de dados referente a este conjunto de dados está disponível no Apêndice C.

Para esta base foi necessário realizar algumas operações para preparar o conjunto de dados antes deste ser inserido na ferramenta "xMML-PPP Tool". Foi criado o atributo alvo, *DEAD*, o qual foi extraído do atributo original *FECHA_DEF*, que indica a data de

óbito. Esse novo atributo recebeu o valor 0 para indicar a classe negativa, não veio a óbito e, 1 para indicar a classe positiva, veio a óbito. Na sequência, foram removidos os atributos: id do paciente, *id_registro* e *FECHA_DEF*. As instâncias com informações desconhecidas das doenças pré-existentes, que eram indicadas no conjunto de dados pelos identificadores *98* e *99* também foram removidas. Cabe mencionar que, para este estudo de caso, foram selecionadas apenas as instâncias cujo resultado do exame PCR-RT foi positivo, uma vez que o objetivo do estudo é classificar se os pacientes que testaram positivo para COVID-19 vieram a óbito. Após essa seleção, o atributo que indicava o resultado do exame também foi removido e, finalmente, o conjunto de dados foi fracionado. O conjunto de dados final, utilizado nos experimentos, contém 38 colunas e 12874 exemplos.

Para o primeiro experimento foi realizado o mínimo possível de pré-processamento, ou seja, somente o necessário para que o conjunto de dados estivesse apto a ser treinado pelo algoritmo RF. Dessa forma, foram retirados os atributos categóricos existentes no conjunto de dados, como informações de datas e nomes de cidades.

Para o segundo experimento, foram realizadas, inicialmente, as mesmas operações de remoção de atributos categóricos existentes no conjunto de dados realizadas no primeiro experimento. Adicionalmente, também foram removidos alguns atributos com informações regionais. Também foi realizada uma análise exploratória prévia dos dados, sendo identificado que pacientes com mais de 40 anos tendo sido hospitalizados tiveram maior incidência de óbito. Assim, três atributos *(FAIXA_ETARIA, TOTAL_DISEASE e HAS_HIGHRISK)* foram criados.

O atributo criado *FAIXA_ETARIA*, recebe o valor 0 quando a idade do paciente for menor que 40 anos e 1 quando for igual ou maior que 40. O atributo *TOTAL_DISEASE* indica a quantidade de comorbidades apresentada pelo paciente infectado pelo vírus. Já o atributo *HAS_HIGHRISK* indica 1 caso o paciente possua uma idade maior ou igual a 40 anos, tendo sido hospitalizado. Dessa forma, a criação desse atributo é realizada como uma tentativa de auxiliar o modelo a identificar mais exemplos da classe minoritária. O modelo foi então treinado com 30 atributos.

Assim, os modelos de ambos os experimentos foram treinados com os seguintes parâmetros, utilizados na configuração do RF: *max_features*: sqrt; *min_samples_split*: 2; *min_samples_leaf*: 1; *max_depth*: 16; e *num_estimators*: 200.

A Tabela 8 mostra o resultado do desempenho para os dois experimentos realizados.

Tabela 8 – Resultados de Experimentos - Base COVID-19 México

Id	**Acurácia**	**Precisão**	**Revocação**	**Medida F1**
1	84,29	84,29	83,49	83,63
2	84,52	83,81	84,52	83,97

Tabela 9 – Relatório de Classificação - Base COVID-19 México - Experimento 1

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0	87,00	93,00	90,00	2953
1	71,00	56,00	63,00	910
Acurácia			84,00	3863
Média Macro	79,00	75,00	76,00	3863
Média Ponderada	83,00	84,00	84,00	3863

Tabela 10 – Relatório de Classificação - Base COVID-19 México - Experimento 2

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0	88,00	93,00	90,00	2953
1	71,00	58,00	64,00	910
Acurácia			85,00	3863
Média Macro	79,00	75,00	76,00	3863
Média Ponderada	84,00	85,00	84,00	3863

A Tabela 9 mostra o relatório de classificação relativo ao experimento 1 da base de dados da Covid-19 no México. Já a Tabela 10 mostra o relatório de classificação relativo ao experimento 2 da base de dados da Covid-19 no México.

Dessa forma, conforme pode-se observar os relatórios de classificação (Tabela 9) e (Tabela 10), ainda que para poucos exemplos, o modelo do experimento 2 consegue classificar melhor os pacientes da classe 1 do que o modelo do experimento 1.

6.3 Explicabilidade do resultado dos experimentos

Nesta seção, são exibidos os gráficos de explicabilidade criados a partir da técnica SHAP, que visam explicar a contribuição dos atributos em cada experimento. Além disso, consultas são realizadas ao repositório de dados para complementar a explicabilidade do modelo e auxiliar no entendimento dos dados que foram utilizados para construir o modelo.

As técnicas XAI são usadas para auxiliar no entendimento do resultado do modelo, porém as informações do tratamento realizadas nos dados que compõem o modelo, e que influenciaram no resultado, não são conhecidas com o XAI. A proveniência de dados ajuda a conhecer como esses dados foram tratados. Neste trabalho, o termo explicabilidade de dados é empregado para referenciar a proveniência das operações realizadas nos dados antes do treinamento. Na subseção 6.3.1 é apresentada a explicabilidade para os experimentos relacionados a base de dados do Titanic. Já na subseção 6.3.2 é apresentada a explicabilidade para os experimentos relacionados a base de dados da Covid-19 - México.

A explicabilidade do modelo é realizada pela aplicação de técnicas XAI. Existem

vários tipos de técnicas XAI, conforme subseção 2.4.1. Estas são usadas conforme o escopo da explicação, tipos de dados utilizado e abordagem da técnica. Para este trabalho é considerado o escopo global e a abordagem da técnica utilizada é a relevância dos atributos. Dessa forma, para a explicabilidade do modelo neste trabalho utilizamos em todos os experimentos o método SHAP. O SHAP foi escolhido dentre outras técnicas de explicabilidade existentes por ser um dos métodos mais abrangentes para explicação de métodos caixa preta, auxiliando na visualização das interações e na importância dos atributos (74).

6.3.1 Explicabilidade - Base de dados do Titanic

A Figura 31 refere-se ao SHAP correspondente ao experimento 1 da base Titanic e a Figura 32 refere-se ao SHAP correspondente ao experimento 2 desta base.

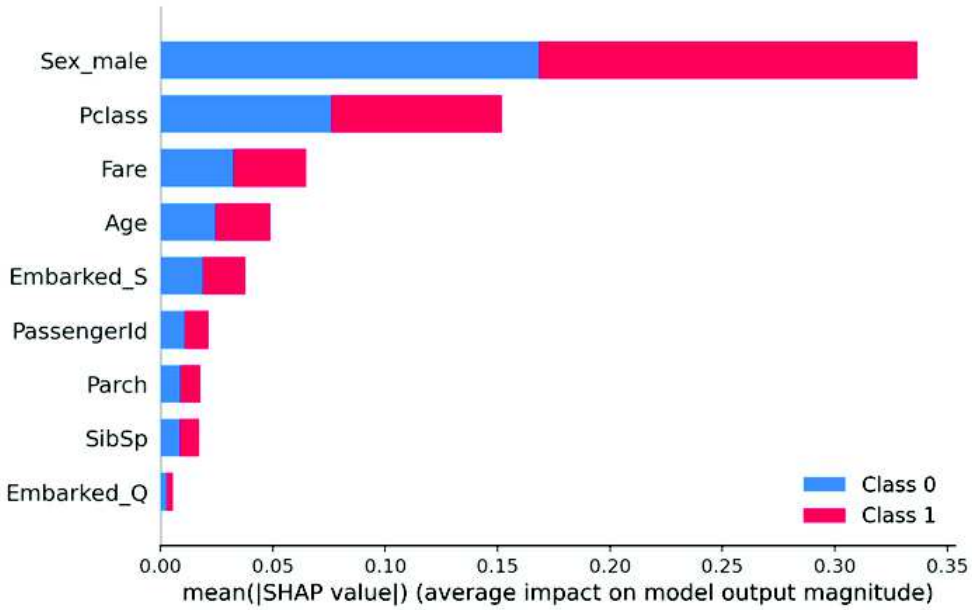

Figura 31 – Gráfico SHAP do experimento 1 - Titanic

Para o gráfico SHAP, apresentado na Figura 31, pode-se perceber que os atributos *Sex_Male, Pclass* e *Fare* foram os atributos que mais contribuíram para o resultado do modelo.

No entanto, ao analisar o gráfico SHAP apresentado na Figura 32, é possível perceber que o atributo construído *Groupsize*, seguido pelos atributos *Sex_Male* e *Title_Mr*, foram os que mais contribuíram para o resultado do modelo. Dessa forma, fica claro que o atributo construído *Groupsize* teve um impacto muito significativo no resultado do modelo, sendo o atributo mais relevante para a melhoria da precisão do modelo.

Utilizando-se consultas ao repositório de dados, pode-se obter mais informações referentes aos atributos utilizados no treinamento e exibidos no gráfico SHAP. Utilizando

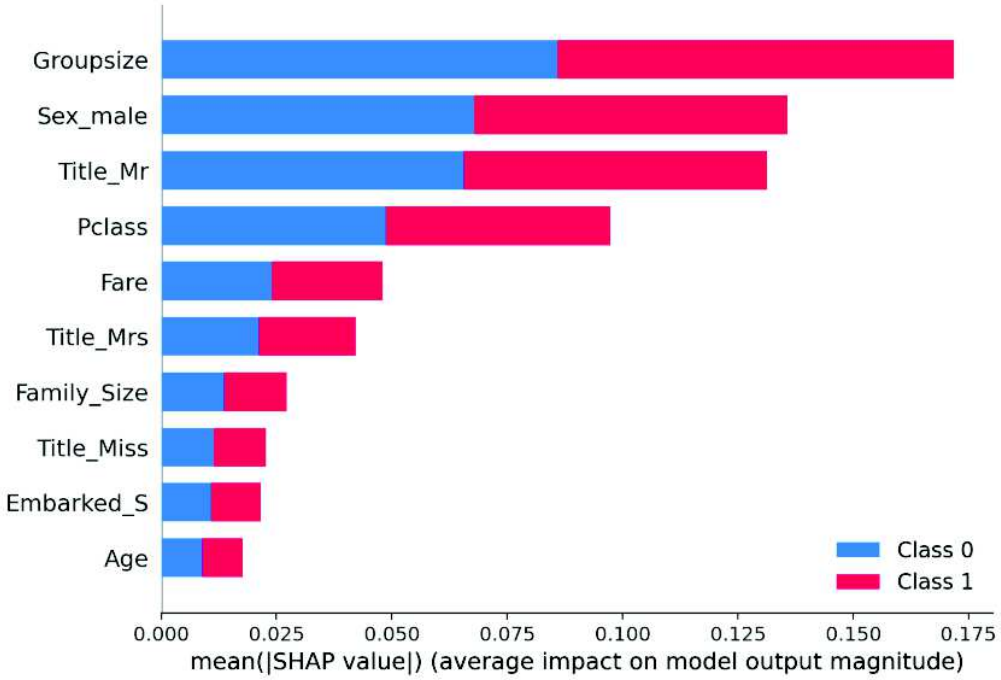

Figura 32 – Gráfico SHAP do experimento 2 - Titanic

como exemplo o atributo *Groupsize*, exibido como atributo de maior contribuição na Figura 32, foram realizadas as consultas para obter as seguintes informações: a origem de criação do atributo, as operações de pré-processamento efetuadas e o valor de contribuição do atributo no modelo.

```
SELECT Experiment_attribute.label as attribute,
experiment_attribute.origin as source
FROM  experiment_attribute,
experiment, dataset, workflow
WHERE experiment_attribute.experimentid=
experiment.id
AND  experiment.datasetid=dataset.id
AND dataset.workflowid = workflow.id
AND  workflow.id = 152
AND experiment.attribute.label
LIKE 'Groupsize';
```

	attribute	source
0	Groupsize	df['Groupsize'] = df['Lastname'].apply(lambda x: df.loc[(df['Sex']=='female') | (df['Title']=='Master')]. loc[df.loc[(df['Sex']=='female') | (df['Title']=='Master')] ['Last-name']==x]['Survived'].count())

Figura 33 – Informações de origem do atributo Groupsize

O resultado das informações dessas consultas pode ser visto nas Figuras 33, 34 e 35. O intuito dessas consultas é trazer explicabilidade dos dados que compõem o modelo,

ou seja, informações complementares às informações providas pelo gráfico XAI.

Já a Figura 34 mostra o comando SQL e o resultado obtido ao executar a pesquisa no repositório de proveniência referente às operações de pré-processamento executadas no atributo *Groupsize* no experimento 2 da Tabela 5.

```
SELECT "operatorsActivity".name,
"operatorsActivity".function,
"operatorsActivity".label_attribute
FROM  "operatorsActivity", workflow
WHERE "operatorsActivity".workflowid=
workflow.id AND  workflow.id = 152
AND "operatorsActivity".label_attribute
LIKE 'Groupsize';
```

	name	function	label
0	IncludeColumn	Atribute Construction	Groupsize
1	MinMaxScaler	Data Normalization	Groupsize
1	TrainTestSplit	Data Partition	Groupsize

Figura 34 – Informações de pré-processamento do atributo Groupsize

A Figura 35 mostra, então, o valor de contribuição do atributo *Groupsize*, que mais contribuiu para a predição do modelo.

```
SELECT "xai_Results".label_feature_importance,
"xai_Results".value_feature_importance,
FROM  "xai_Results",xai, experiment,
dataset, workflow
WHERE "xai_Results"."int"
AND xai.experimentid=experiment.id
AND experiment.datasetid=dataset.id
AND dataset.workdlowid=workflow.id
AND workflow.id = 152
AND "xai_Results".label_feature_importance
LIKE 'Groupsize';
```

	label_feature_importance	value_feature_importance
0	Groupsize	53.4932

Figura 35 – Informações de valor de contribuição do atributo Groupsize

As Figuras 36 e 37 referem-se a protótipos da ferramenta de explicabilidade SHAP com uma interface adaptada a fim de permitir a integração da visualização da informação de proveniência dos atributos derivados *Groupsize* e *Family_size* respectivamente, a partir da seleção destes atributos.

No Apêndice C foi conduzido um estudo visando destacar de maneira mais evidente o impacto que cada atributo construído exerce sobre os resultados do modelo. Isto é, visamos

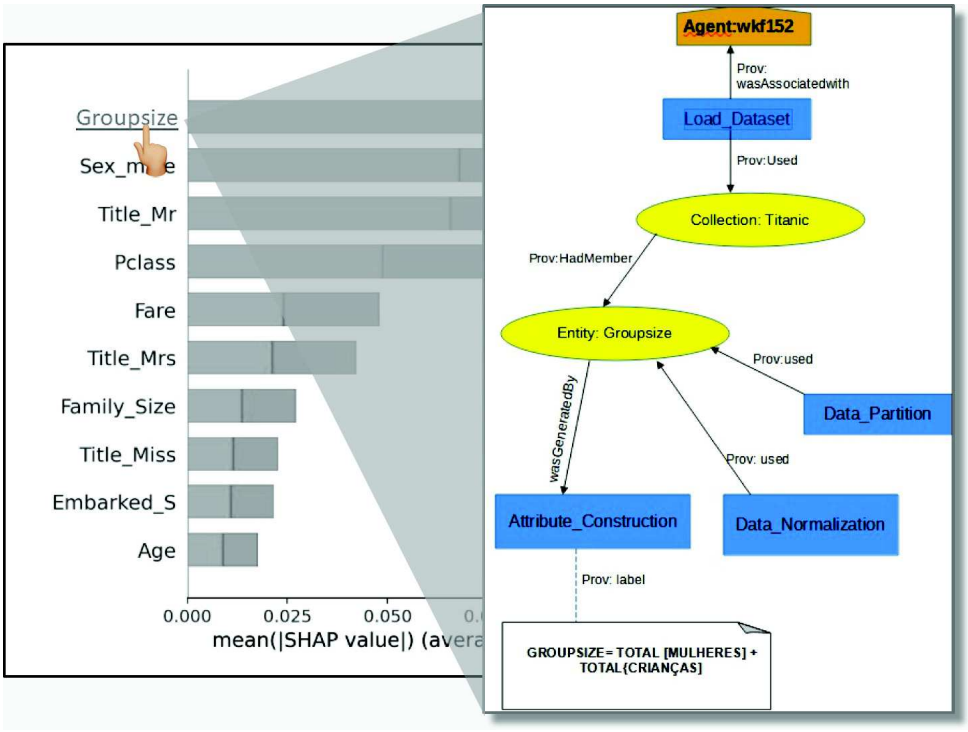

Figura 36 – Exemplo de visualização gráfica da proveniência do atributo *Groupsize* através do gráfico SHAP

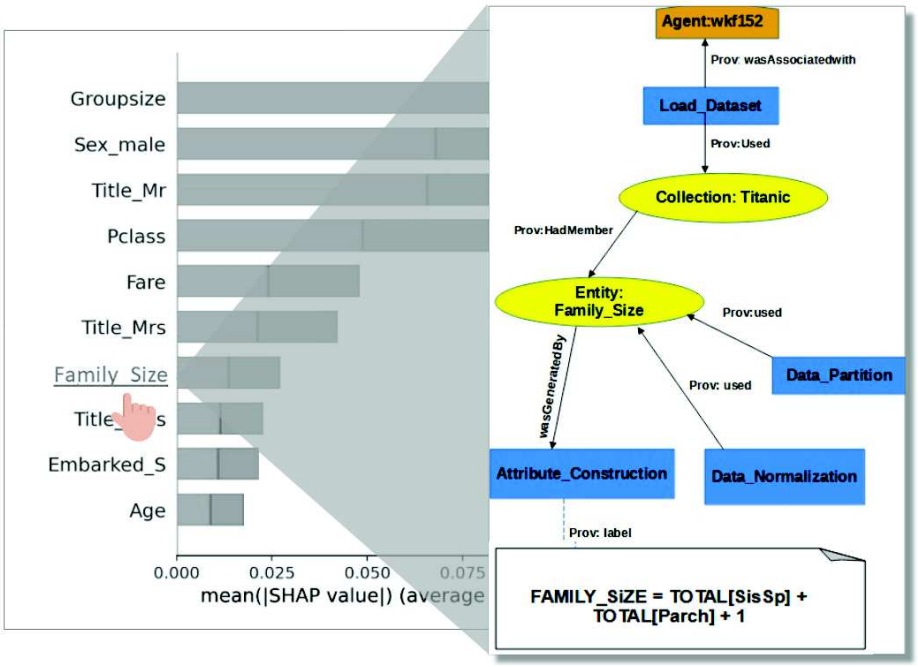

Figura 37 – Exemplo de visualização gráfica da proveniência do atributo Family_Size através do gráfico SHAP

compreender as etapas intermediárias que culminam no experimento final (experimento 2), ou seja, aquele que apresentou o melhor desempenho utilizando esta base de dados. Além desse aspecto, incorporamos um modelo que utilizou um atributo construído de maneira inadequada, resultando em uma modificação significativa no desempenho do

modelo como um todo. O propósito subjacente a esse estudo foi ressaltar a importância da proveniência dos dados, tanto para aprimorar a explicabilidade do processo quanto para elevar a confiabilidade do mesmo.

6.3.2 Explicabilidade - Base de dados do Covid-19

As Figuras 38 e 39 mostram os gráficos SHAP, gerados a partir dos modelos do experimento 1 e 2, respectivamente, referente a base de dados epidemiológicos do México.

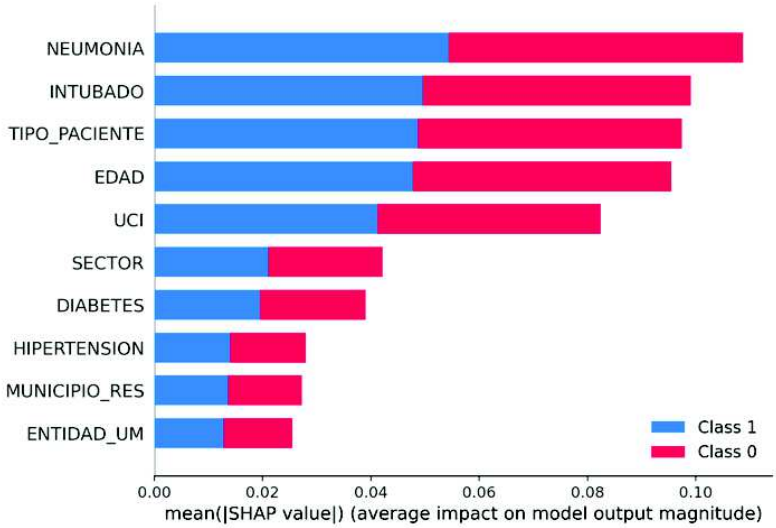

Figura 38 – Gráfico SHAP do experimento 1 - Covid19 - México

Conforme o gráfico SHAP apresentado na Figura 38, pode-se perceber que os atributos *NEUMONIA*, *INTUBADO*, *TIPO_PACIENTE* e *EDAD* foram os atributos que mais contribuíram com o resultado deste primeiro modelo.

No entanto, o gráfico SHAP apresentado na Figura 39 mostra que os atributos construídos *HAS_HIGHRISK* e *TOTAL_DISEASE*, que aparecem em sexta e sétima posições, respectivamente, em valores de contribuição, embora tenham tido uma contribuição modesta, certamente contribuíram para a classificação correta de alguns exemplos da base.

A ferramenta xMML-PPP já possui as principais consultas disponíveis para recuperar informações no repositório de dados, sem necessitar que sejam escritos códigos SQL. No entanto, por ser importante a construção de consultas livres que respondam melhor a questões específicas. Dessa forma, comandos SQL também podem ser aceitos. Visando obter uma ideia de questões de consultas importantes para provimento de explicabilidade de dados, algumas dessas questões são apresentadas na Tabela 11. Para responder a essas consultas, foi utilizado, como exemplo, o *fluxo de trabalho* referente ao experimento 2 do conjunto de dados de pacientes com COVID-19 do México.

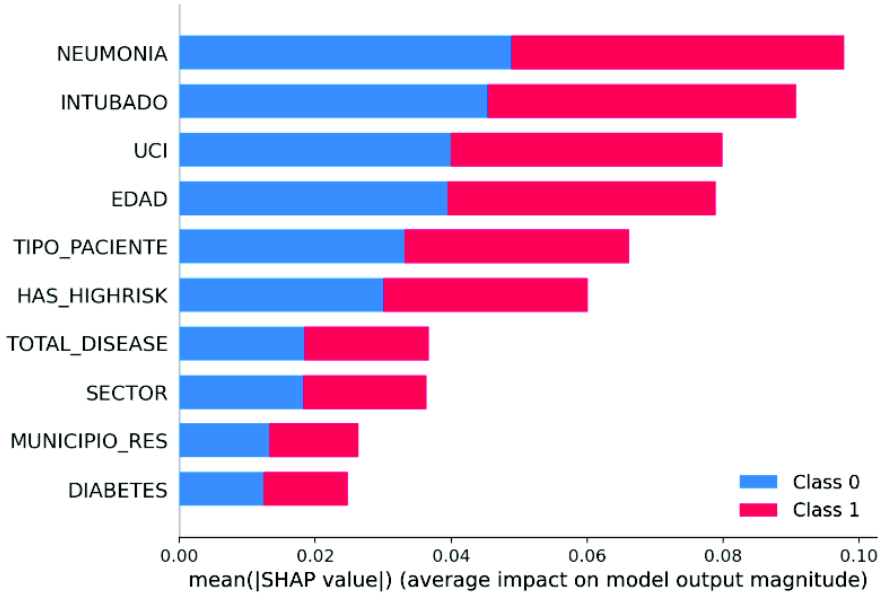

Figura 39 – Gráfico SHAP do experimento 2 - Covid19 - México

Tabela 11 – Questões de consultas de proveniência

Exemplos de consultas de Proveniência	
Q1	"Dado um modelo treinado, quais foram os atributos que derivaram o conjunto de treinamento de um experimento específico?"
Q2	"Dado um modelo treinado, quais atributos participantes do conjunto de treinamento foram construídos?"
Q2.a	"Dado um modelo treinado, qual foi a origem de construção dos atributos que foram construídos ?"
Q3	"Dado um modelo treinado, quais foram os 10 atributos que mais contribuíram para o resultado do modelo, por ordem de importância de contribuição? "
Q4	"Dado um modelo treinado, de um fluxo de trabalho específico, quais foram os valores para todos os parâmetros usados e os valores de medida de avaliação associados ao modelo?"
Q5	"Dado um modelo treinado qual o significado (descrição do atributo) de atributos específicos, que mais contribuíram com o resultado do modelo? "
Q6	"Dado um modelo treinado, quais foram as operações de pré-processamento realizadas nos sete primeiros atributos que mais contribuíram com o resultado desse modelo? "

Ao responder a Q1 (Figura 40), é possível identificar os atributos utilizados na construção do modelo treinado. Saber quais atributos são utilizados para construir um modelo é importante para entender quais características disponíveis no conjunto de dados foram consideradas para obter o resultado do modelo. Nesta consulta, além do *label* do atributo é exibida também a informação da origem do atributo, ou seja, se é um atributo

original do conjunto de dados ou um atributo construído. Um dos critérios de consulta é o número do workflow, que é identificado como *workflow.id*. A informação do id do workflow é exibida no início do fluxo de trabalho, de modo a facilitar as posteriores consultas. Note que, para os atributos construídos é exibida também a informação de construção do atributo.

```
SELECT experiment_attribute_label,
experiment_attribute_origin
FROM  experiment_attribute, experiment,
dataset, workflow
WHERE experiment_attribute.experiment.id=
experiment.id
AND experimentid.datasetid=dataset.id
AND dataset.workflowid=workflow.id
AND workflow.id = 146
```

	label	*origin*
0	ENTIDAD_UM	atributo original do dataset
1	ENTIDAD_RES	atributo original do dataset
2	ENTIDAD_REGISTRO	atributo original do dataset
	- - -	
25	MIGRANTE	atributo original do dataset
26	UCI	atributo original do dataset
27	DEAD	atributo original do dataset
28	TOTAL_DISEASE	df['TOTAL_DISEASE'] = df[df.columns[22:31]].sum(axis=1)
29	FAIXA_ETARIA	df['FAIXA_ETARIA'] = df.apply(lambda df: (0 if df['EDAD'] <= 39 else 1), axis=1)
30	HAS_HIGHRISK	df['HAS_HIGHRISK'] = df.apply(lambda df: (1 if (df['FAIXA_ETARIA'] == 1 and df['TIPO_PACIENTE']==2) else 0), axis=1)

Figura 40 – Consulta Q1 - Atributos que derivaram o conjunto de treinamento

```
SELECT "operatorsActivity".label as Atributo
FROM  "operatorsActivity", workflow
WHERE "operatorsActivity".workflowid=
workflow.id AND  workflow.id = 146
AND "operatorsActivity".function=
'Attribute Construction';
```

	Atributo
0	TOTAL_DISEASE
1	FAIXA_ETARIA
2	HAS_HIGHRISK

Figura 41 – Consulta Q2 - Atributos construídos

A resposta das Q2 (Figura 41) informa quais dos atributos que geraram o modelo não são originais do conjunto de dados, ou seja, foram derivados de outros atributos já existentes. Porém, é importante também conhecer a derivação desses atributos, principalmente se esses novos atributos tiveram uma boa contribuição no resultado do modelo, dessa forma, a consulta Q2a (Figura 42), complementa a Q2.

```
SELECT  experiment_attribute.label,
experiment_attribute.origin
FROM  experiment_attribute, experiment,
dataset,workflow
WHERE experiment_attribute.experimentid=
experiment.id AND experiment.datasetid=
dataset.id AND dataset.workflowid=
workflow.id AND workflow.id=146
and experiment_attribute.label IN
(SELECT "operatorsActivity".label_attribute
AS Atributo
FROM "operatorsActivity",workflow
WHERE  "operatorsActivity".workflowid=
workflow.id AND workflow.id=146
AND "operatorsActivity".function='Attribute␣Construction')
```

	label	*origin*
0	TOTAL_DISEASE	df['TOTAL_DISEASE'] = df[df.columns[22:31]].sum(axis=1)
1	FAIXA_ ETARIA	df['FAIXA_ETARIA'] = df.apply(lambda df: (0 if df['EDAD'] <= 39 else 1), axis=1)
2	HAS_ HIGHRISK	df['HAS_HIGHRISK'] = df.apply(lambda df: (1 if (df['FAIXA_ETARIA'] == 1 and df['TIPO_PACIENTE']==2) else 0), axis=1)

Figura 42 – Consulta Q2A - Derivação de atributos construídos

A Q3 (Figura 43) responde quais atributos mais contribuíram com o modelo treinado. Essa informação é possível uma vez que o modelo de dados (Figura 30) armazena os atributos que mais contribuíram com o resultado do modelo, por ordem de contribuição. Além disso, os valores de contribuição podem também ser consultados. Esta informação pode servir para transformar o resultado do gráfico XAI em uma informação textual, por exemplo.

A Q4 responde quais os parâmetros e os valores de medida de desempenho para um específico modelo treinado. Essas informações ajudam a analisar o desempenho e a respectiva configuração do modelo. Porém, para melhor visualização dos resultados, a consulta Q4 foi dividida em duas partes: Q4A, que corresponde a Figura 44 e Q4B, que corresponde a Figura 45 e respondem sobre as medidas de desempenho e os parâmetros utilizados num determinado modelo treinado, respectivamente. Esta informação ajuda a realizar análises sobre o desempenho e a respectiva configuração usada em cada modelo.

```
SELECT "xai_Results".label_feature_importance,
FROM  "xai_Results", xai, experiment,
dataset, workflow
WHERE "xai_Results"."int"=xai.id
AND xai.experimentid=experiment.id
AND experiment.datasetid=dataset.id
AND dataset.workdlowid=workflow.id
AND workflow.id = 146 limit 10
```

	label_feature_importance
0	NEUMONIA
1	INTUBADO
2	UCI
3	EDAD
4	TIPO_PACIENTE
5	HAS_HIGHRISK
6	TOTAL_DISEASE
7	SECTOR
8	MUNICIPIO_RES
9	DIABETES

Figura 43 – Consulta Q3 - Atributos com maior contribuição no modelo

```
SELECT  experiment.accuracy,experiment.recall, experiment.
   precision,experiment.f1score FROM  experiment,
dataset,workflow
WHERE experiment.datasetid=
dataset.id AND dataset.workflowid= workflow.id
AND workflow.id = 146
```

	acurácia	sensibilidade	Precisão	f1score
0	0.8452	0.8452	0.8381	0.8397

Figura 44 – Consulta Q4A - Medidas de avaliação do Modelo

A Q5 (Figura 46) ajuda a responder qual a descrição dos atributos. Neste trabalho foi considerado guardar a descrição dos atributos originais no repositório de dados. A informação da descrição dos atributos, ou seja, seu significado pode contribuir para um melhor entendimento do resultado de um gráfico XAI, especificamente em alguns conjuntos de dados, quando apenas o nome do atributo não for suficiente para o entendimento do seu significado. Dessa forma, por intermédio de consultas ao repositório de dados, pode-se resgatar essa descrição, complementando, dessa forma, esse entendimento.

Finalmente, a Q6 (Figura 47) responde ao objetivo principal da abordagem xMML-PPP. Após o modelo estar treinado, é possível saber quais as operações de pré-processamento efetuadas para cada atributo que participou no modelo, incluindo as que mais contribuíram para o desempenho do modelo.

```
SELECT  parameter.label, parameter.value
FROM   parameter,  experiment, dataset, workflow
WHERE  parameter.experimentid=experiment.id AND  experiment.
   datasetid=dataset.id AND dataset.workflowid = workflow.id AND
   workflow.id = 146
GROUP BY experiment.id, parameter.label, parameter.value
```

	label	value
0	max_depth	16
1	max_features	sqrt
2	min_samples_leaf	1
3	min_samples_split	2
4	n_estimators	200
5	random_state	42

Figura 45 – Consulta Q4B- Parâmetros utilizados no Modelo

```
SELECT dataset_attribute.label,
dataset_attribute.description
FROM  dataset_attribute,dataset,workflow
WHERE dataset_attribute.datasetid=
dataset.id
AND dataset.workflowid=workflow.id
AND workflow.id=146
AND dataset_attribute.label IN
('NEUMONIA', 'INTUBADO','UCI','SECTOR')
```

	label	*description*
0	SECTOR	Identify the type of institution of the National Health System that provided the care.
1	INTUBADO	Identifies if the patient required intubation.
2	NEUMONIA	Identifies if the patient was diagnosed with pneumonia.
3	UCI	Identifies if the patient was admitted to Intensive Care Unit.

Figura 46 – Consulta Q5 - Descrição de atributos específicos

Uma vez que estas operações influenciam o resultado do modelo, o conhecimento desta informação pode ajudar a complementar a compreensão do comportamento do modelo. Por exemplo, no gráfico representado na Figura 39, os atributos *NEUMONIA*, *INTUBADO* e *UCI* aparecem com a contribuição mais significativa para o modelo do experimento 2.

Aplicando Q6, podemos obter informações sobre quais operações de pré-processamento foram efetuadas sobre estes atributos, um exemplo de explicabilidade dos dados, fornecida pelo resultado da consulta Q6, apoiando a explicabilidade do modelo, fornecida pelo gráfico SHAP. Também é possível complementar a análise de resultados dos experimentos, unindo o resultado das buscas Q1, Q4 e Q6, por exemplo, para obter informações sobre a

```
SELECT "operatorsActivity".name,  "operatorsActivity".function,
   "operatorsActivity".label_attribute
FROM "operatorsActivity",workflow
WHERE "operatorsActivity".
workflowid=workflow.id
AND workflow.id=146
AND "operatorsActivity".label_attribute
IN (SELECT  "xai_Results".
label_feature_importance
FROM "xai_Results",xai,experiment,
dataset,workflow
WHERE"xai_Results"."int"=xai.id
AND xai.experimentid = experiment.id
AND experiment.datasetid=dataset.id
AND dataset.workflowid=workflow.id
AND workflow.id= 146 limit 7)
```

	name	*function*	label_attribute
0	IncludeColumn	AttributeConstruction	TOTAL_DISEASE
1	IncludeColumn	AttributeConstruction	HAS_HIGHRISK
2	TrainTestSplit	DataPartition	TIPO_PACIENTE
3	TrainTestSplit	DataPartition	INTUBADO
4	TrainTestSplit	DataPartition	NEUMONIA
5	TrainTestSplit	DataPartition	EDAD
6	TrainTestSplit	DataPartition	UCI
7	TrainTestSplit	DataPartition	TOTAL_DISEASE
8	TrainTestSplit	DataPartition	HAS_HIGHRISK

Figura 47 – Consulta Q6 - Operações de pré-processamento dos atributos de maior contribuição

configuração do modelo, o desempenho e as operações de pré-processamento aplicadas, além da importância da contribuição de cada atributo, de acordo com essas configurações. Assim sendo, as análises por meio da recuperação das informações existentes no repositório de dados podem ainda ser mais expandidas como, por exemplo:

- realizar análises por comparações dos resultados de diferentes modelos treinados, em termos de métricas de desempenho (e esta análise pode ser realizada por alguma métrica específica), sendo possível, dessa forma, relacionar os prováveis fatores que contribuíram para esse resultado, tais como a alteração dos parâmetros, a alteração dos valores desses parâmetros de treino do modelo, as operações de pré-processamento realizadas nos atributos; e
- realizar análises com dados de treino de vários modelos com operações de pré-processamento diferentes para um atributo específico e realizar a comparação do comportamento daquele atributo no modelo permitindo entender a importância

relativa de cada atributo, conforme o tratamento que ele recebeu, e como esse tratamento pode influenciar nos resultados finais, conforme os operadores utilizados para esse atributo que tenha sido alterado.

Em resumo, a realização dessas análises permite uma comparação referente às contribuições dos atributos entre as diferentes configurações de modelos treinados. Isso permitirá identificar se existem atributos consistentemente importantes ou se a relevância varia significativamente em diferentes configurações, e principalmente, se o comportamento altera conforme as operações de pré-processamento são alteradas.

Assim, mediante consultas no repositório de dados, como exemplificado em Q1 a Q6, cujas análises podem ser expandidas conforme exemplificado, obtém-se a explicabilidade dos dados, que contribui com informações adicionais com vistas a expandir a compreensão do resultado do modelo. A explicabilidade de dados, como visto, se refere às informações sobre os dados que derivaram um modelo de AM e, por meio da explicabilidade de dados, é possível inferir como o tratamento dos dados, realizado pelas operações de pré-processamento, contribui para um determinado resultado de um modelo, complementando, dessa forma, a explicabilidade do modelo, provida pelas técnicas XAI.

7 CONCLUSÃO

À medida que as soluções de IA se expandem, o seu impacto e a sua importância se tornam cada vez mais evidentes, especialmente em setores críticos nos quais a confiabilidade e a transparência são fundamentais. A XAI busca fornecer explicações e interpretações compreensíveis sobre o modo pelo qual os modelos de IA alcançam suas decisões, destacando-se sobretudo em casos nos quais os modelos de caixa-preta, caracterizados por baixa interpretabilidade, estão presentes.

No entanto, os dados utilizados para derivar um modelo também são frequentemente tratados antes do treino e contribuem com o processo de decisão. Nesse sentido, considerar a proveniência dos dados pode contribuir significativamente para tornar o processo de tomada de decisão dos modelos de IA mais compreensível e explicável.

Neste trabalho, propomos a xMML-PPP, uma abordagem que utiliza informações sobre a proveniência dos dados para melhorar a compreensão da explicação dos modelos de AM. No entanto, nossa proposta não se trata apenas de mais uma abordagem para rastrear a proveniência no contexto de AM. Nossa proposta visa, de maneira mais significativa, aprimorar a compreensão da explicabilidade dos modelos, a qual é provida utilizando-se técnicas de explicabilidade (XAI). Isso é alcançado ao incorporar a proveniência dos dados, com especial atenção à fase de pré-processamento, que chamamos de explicabilidade de dados. Assim, unimos a proveniência à utilização do XAI com vistas a agregar explicabilidade aos sistemas de classificação em AM.

A abordagem proposta introduz um modelo de dados que possibilita a conexão entre informações relacionadas ao pré-processamento de dados e dados de explicabilidade do modelo, juntamente com a inclusão de dados sobre a configuração e o desempenho do modelo. Além disso, uma arquitetura foi desenvolvida para viabilizar a captura e recuperação dos dados pela ferramenta construída, “xMML-PPP — Tool”, conforme a estrutura estabelecida. A “xMML-PPP — Tool” permite a realização dos passos de preparação dos dados para treino a partir da escolha de operadores de pré-processamento definidos, além do treino do modelo e da explicação por ferramenta XAI. Os dados capturados pela “xMML-PPP — Tool” são enviados e armazenados no “xMML-PPP — Prov”, o repositório de proveniência da ferramenta, conforme detalhado no Capítulo 4.

A avaliação da abordagem foi realizada por meio da condução de quatro experimentos utilizando dois conjuntos de dados, o conjunto de dados do Titanic e um conjunto de dados epidemiológicos para casos de COVID-19 no México. Verificou-se que o conhecimento acerca das operações de pré-processamento, com foco especial na engenharia de atributos, desempenha um papel crucial na compreensão da formação dos atributos que constituem

um modelo. Esse entendimento é especialmente importante quando tais atributos exercem uma influência substancial no desempenho do modelo. Segundo a análise realizada no Apêndice C, foi evidenciado que ao introduzir etapas progressivas de pré-processamento, especialmente focadas na criação de atributos, e ao construir um modelo para cada atributo adicionado, houve modificações no desempenho do modelo, assim como impactos na interpretação desse modelo. Tornou-se evidente que a proveniência da formação dos atributos desempenhou um papel significativo na compreensão de como esses atributos se desenvolvem, contribuindo para a clareza dos resultados do modelo.

Além disso, foi observado que é viável conduzir análises de dados por meio de consultas que estabelecem relações entre as etapas de pré-processamento, o desempenho e os resultados. Isso amplia a compreensão da explicabilidade do modelo através da análise dos dados. Por meio de questionamentos específicos de origem, apresentados no Capítulo 6, foi possível, através da execução de consultas ao repositório de proveniência, obter informações como: quais atributos deram origem ao conjunto de treinamento de um experimento particular; quais desses atributos eram resultantes de derivação; qual foi a origem de construção desses atributos; quais parâmetros e seus valores foram utilizados no treinamento do modelo; quais métricas de desempenho foram associadas a um determinado experimento; quais operações de pré-processamento foram aplicadas a cada atributo que teve maior impacto nos resultados do modelo. É importante ressaltar que essas consultas ilustradas no capítulo podem ser ampliadas conforme a demanda por conhecimento.

Dentre as contribuições da presente dissertação, foram observadas:

- Especificação da abordagem xMML-PPP visando contribuir com a explicabilidade em AM;
- Implementação da arquitetura e o desenvolvimento dos elementos arquiteturais definidos: a ferramenta xMML-PPP Tool e do repositório de proveniência xMML-PPP Prov; e
- Aplicação da xMML-PPP por meio de estudos de caso para validação da abordagem. Mediante a condução dos estudos de caso foi possível:
  - a validação da abordagem por meio da utilização da ferramenta xMML-PPP Tool em todas as etapas do ciclo de vida do AM. Nesse processo, são capturadas as informações de proveniência conforme o escopo definido, visando alcançar a explicabilidade de dados;
  - a execução de consultas de proveniência à camada xMML-PPP Prov e recuperação dos dados relativos à instanciação de cada fase do fluxo de trabalho, inclusive das operações de pré-processamento e do resultado da ferramenta XAI aplicada ao modelo treinado; e

– apresentação de um protótipo de interface gráfica para visualização interativa da proveniência através do gráfico SHAP.

Certamente, o trabalho realizado possui algumas limitações e aspectos que podem ser explorados em futuras pesquisas com o intuito de aprimorar a avaliação de maneira mais abrangente. Uma das limitações que podemos citar é em relação à aplicação dos experimentos, uma vez que todos foram aplicados utilizando um único tipo de algoritmo caixa-preta, o *Random Forest*. Dessa forma, não foi possível avaliar se ocorreria alguma variação no comportamento dos modelos em relação à aplicação das operações de pré-processamento e seus respectivos resultados, quando diferentes algoritmos fossem empregados.

No que diz respeito aos trabalhos futuros, uma oportunidade que se destaca é a exploração da viabilidade de aplicar a abordagem proposta a uma variedade de tipos de dados, tais como imagens e vídeos, ampliando seu alcance além dos dados tabulares. Além desse aspecto, é relevante considerar a avaliação da abordagem em conjunção com a utilização de outros algoritmos, como as redes neurais. Isso provavelmente demandará ajustes para a adaptação ao modelo, composto por camadas interconectadas de nós.

Ademais, é de interesse desenvolver uma versão interativa do gráfico SHAP, com o propósito de integrá-la à xMML-PPP Tool. Essa versão interativa possibilitaria a representação visual das informações de proveniência capturadas diretamente do gráfico de XAI. Dessa forma, seria possível aos usuários selecionar os atributos de interesse, conforme ilustrado nos protótipos de interfaces apresentados subseção 6.3.1, e obter a linhagem da proveniência relativa ao atributo desejado.

Em relação ao armazenamento de dados, outra possibilidade a ser considerada é a substituição da base de dados relacional atual por uma base de dados em grafo. Essa mudança traria maior flexibilidade, especialmente para a inclusão de novas entidades, além de permitir o armazenamento de informações de treinamento em tempo de execução de uma rede neural sem a necessidade de preocupações com alterações no esquema da base de dados.

Outra perspectiva para trabalhos futuros reside na exploração do emprego de ontologias. Tal abordagem visaria ampliar a expressividade semântica e facilitaria a interoperabilidade dos dados, por intermédio de ontologias já existentes e expandindo-as para atender aos objetivos da explicabilidade dos dados e do modelo propostos na xMML-PPP. Essa iniciativa promoveria uma compreensão mais profunda e um maior alcance na interpretação dos resultados obtidos.

Finalmente, é importante ressaltar que a adoção de uma base de dados em grafo e a incorporação de ontologias estão alinhadas com a adesão aos princípios FAIR (Findable, Accessible, Interoperable, Reusable), que têm o objetivo de tornar os dados mais acessíveis, interoperáveis e reutilizáveis. Assim, a adoção dessas tecnologias é um desejável trabalho

futuro.

A integração de ontologias, por exemplo, aprimora a compreensão dos dados por diferentes sistemas, promovendo a interoperabilidade entre eles. Ao expandir as ontologias existentes para atender aos objetivos de explicabilidade dos dados e do modelo propostos na xMML-PPP, ampliamos o escopo da interpretação dos resultados, o que, por sua vez, aumenta a transparência e a confiabilidade da abordagem. Adicionalmente, as ontologias facilitam a reutilização, fornecendo uma base de entendimento semântico comum, tornando os dados mais valiosos em diversos contextos.

Da mesma forma, a adoção de bancos de dados em grafo também promove a reutilização dos dados, graças à sua estrutura altamente flexível e adaptável, além de possibilitar o armazenamento de dados no formato RDF (Resource Description Framework), o que amplia a acessibilidade e a compatibilidade com padrões semânticos.

REFERÊNCIAS

1 SHAH, N.; ENGINEER, S.; BHAGAT, N. et al. Research trends on the usage of machine learning and artificial intelligence in advertising. *Augmented Human Research*, v. 5, n. 1, p. 19, 2020. Disponível em: <https://doi.org/10.1007/s41133-020-00038-8>.

2 FEDORKO, R.; KRAL, S.; BACIK, R. Artificial intelligence in e-commerce: A literature review. In: _____. [S.l.: s.n.], 2022. p. 677–689. ISBN 978-981-16-9112-6.

3 BERRADA, I. R.; BARRAMOU, F. Z.; ALAMI, O. B. A review of artificial intelligence approach for credit risk assessment. In: *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. [S.l.: s.n.], 2022. p. 1–5.

4 TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, Nature Publishing Group, v. 25, n. 1, p. 44–56, 2019.

5 KNAPIC, S.; MALHI, A.; SALUJA, R.; FRäMLING, K. *Explainable Artificial Intelligence for Human Decision-Support System in Medical Domain*. 2021.

6 KALE, A.; NGUYEN, T.; HARRIS FREDERICK C., J.; LI, C.; ZHANG, J.; MA, X. Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence*, v. 5, n. 1, p. 139–162, 03 2023. ISSN 2641-435X. Disponível em: <https://doi.org/10.1162/dint_a_00119>.

7 JENTZSCH, S. F.; HOCHGESCHWENDER, N. Don't forget your roots! using provenance data for transparent and explainable development of machine learning models. In: *2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW)*. [S.l.: s.n.], 2019. p. 37–40.

8 CHAPMAN, A.; MISSIER, P.; SIMONELLI, G.; TORLONE, R. Capturing and querying fine-grained provenance of preprocessing pipelines in data science. *Proc. VLDB Endow.*, VLDB Endowment, v. 14, n. 4, p. 507–520, dec 2020. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/3436905.3436911>.

9 SCHERZINGER, S.; SEIFERT, C.; WIESE, L. The best of both worlds: Challenges in linking provenance and explainability in distributed machine learning. In: *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. [S.l.: s.n.], 2019. p. 1620–1629.

10 JAIGIRDAR, F. T.; RUDOLPH, C.; OLIVER, G.; WATTS, D.; BAIN, C. What information is required for explainable ai? : A provenance-based research agenda and future challenges. In: *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*. [S.l.: s.n.], 2020. p. 177–183.

11 TSAKALAKIS, N.; STALLA-BOURDILLON, S.; CARMICHAEL, L.; HUYNH, T. D.; MOREAU, L.; HELAL, A. The dual function of explanations: Why it is useful to compute explanations. *Computer Law & Security Review*, v. 41, p. 105527, 2021. ISSN 0267-3649. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0267364920301321>.

12 CHAPMAN, A.; LAURO, L.; MISSIER, P.; TORLONE, R. Dpds: Assisting data science with data provenance. *Proc. VLDB Endow.*, VLDB Endowment, v. 15, n. 12, p. 3614–3617, sep 2022. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/3554821.3554857>.

13 NAMAKI, M. H.; FLORATOU, A.; PSALLIDAS, F.; KRISHNAN, S.; AGRAWAL, A.; WU, Y.; ZHU, Y.; WEIMER, M. Vamsa: Automated provenance tracking in data science scripts. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, Jul 2020. Disponível em: <http://dx.doi.org/10.1145/3394486.3403205>.

14 MOURA, L. d. A. L.; SILVA, M. A. A. da; CORDEIRO, K. d. F.; CAVALCANTI, M. C. R. A well-founded ontology to support the preparation of training and test datasets. In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 2*. SCITEPRESS, 2021. p. 99–110. Disponível em: <https://doi.org/10.5220/0010460000990110>.

15 HARTLEY, M.; OLSSON, T. S. dtoolai: Reproducibility for deep learning. *Patterns*, v. 1, n. 5, p. 100073, 2020. ISSN 2666-3899. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666389920300933>.

16 SOUZA, R.; AZEVEDO, L.; LOURENÇO, V.; SOARES, E.; THIAGO, R.; BRANDÃO, R.; CIVITARESE, D.; BRAZIL, E.; MORENO, M.; VALDURIEZ, P. et al. Provenance data in the machine learning lifecycle in computational science and engineering. In: IEEE. *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*. [S.l.], 2019. p. 1–10.

17 W3C. *The PROV Data Model*. 2013. <https://www.w3.org/TR/PROV-DM>. (Acessado em 01/11/2021).

18 HERSCHEL, M.; DIESTELKäMPER, R.; LAHMAR, H. B. A survey on provenance: What for? what form? what from? *The VLDB Journal*, Springer-Verlag, Berlin, Heidelberg, v. 26, n. 6, p. 881–906, dec 2017. ISSN 1066-8888. Disponível em: <https://doi.org/10.1007/s00778-017-0486-1>.

19 FREIRE, J.; KOOP, D.; SANTOS, E.; SILVA, C. T. Provenance for computational tasks: A survey. *Computing in Science Engineering*, v. 10, n. 3, p. 11–21, 2008.

20 DAVIDSON, S.; FREIRE, J. Provenance and scientific workflows: Challenges and opportunities. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. [S.l.: s.n.], 2008. p. 1345–1350.

21 PéREZ, B.; RUBIO, J.; SáENZ-ADáN, C. A systematic review of provenance systems. *Knowl. Inf. Syst.*, Springer-Verlag, Berlin, Heidelberg, v. 57, n. 3, p. 495–543, dec 2018. ISSN 0219-1377. Disponível em: <https://doi.org/10.1007/s10115-018-1164-3>.

22 KHAN, F. Z.; SOILAND-REYES, S.; SINNOTT, R. O.; LONIE, A.; GOBLE, C.; CRUSOE, M. R. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. *GigaScience*, v. 8, n. 11, p. giz095, 11 2019. ISSN 2047-217X. Disponível em: <https://doi.org/10.1093/gigascience/giz095>.

23 GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining, Conceitos, Técnicas, algoritmos, orientações e aplicações*. 2. ed. Brasil: Elsevier, 2015. 296 p.

24 MOURA, L. d. A. L. *Preparados: Uma abordagem baseada em ontologia de fundamentação para apoiar a preparação de conjuntos de dados de treinamento e de teste*. 0–143 p. Tese (Doutorado) — IME, 2021.

25 GAMA, J.; FACELI, K.; LORENA, A. C.; CARVALHO, A. C. P. L. F. de. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2. ed. Brasil: Grupo Gen - LTC, 2021. 400 p.

26 GARCíA, S.; LUENGO, J.; HERRERA, F. *Data preprocessing in data mining*. 2. ed. [S.l.]: Springer, 2015.

27 HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. 2. ed. [S.l.]: Elsevier, 2011.

28 RUSSELL, P. N. S. *Inteligência artificial*. 3. ed. Brasil: Grupo GEN, 2013. 400 p.

29 MITCHEL, T. M. *Machine Learning*. [S.l.]: Mcgraw-hill, 1997.

30 BREIMAN, L. Machine learning, volume 45, number 1 - springerlink. *Machine Learning*, v. 45, p. 5–32, 10 2001.

31 LIAW, A.; WIENER, M. Classification and regression by randomforest. *Forest*, v. 23, 11 2001.

32 VAPNIK, V. N. *The Nature of Statistical Learning Theory*. [S.l.]: Springer, 1995.

33 ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018.

34 LENT, M. van; FISHER, W.; MANCUSO, M. An explainable artificial intelligence system for small-unit tactical behavior. In: *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*. [S.l.]: AAAI Press, 2004. (IAAI'04), p. 900–907. ISBN 0262511835.

35 TUREK, D. M. *XAI*. 2018. <https://www.darpa.mil/program/explainable-artificial-intelligence>. (Acessado em 20/10/2021).

36 ALICIOGLU, G.; SUN, B. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 2021. ISSN 0097-8493. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0097849321001886>.

37 Barredo Arrieta, A.; DíAZ-RODRíGUEZ, N.; Del Ser, J.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCIA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, v. 58, p. 82–115, 2020. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.

38 DOSHI-VELEZ, F.; KIM, B. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017.

39 GILPIN, L. H.; BAU, D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2019.

40 SRIHARI, S. Explainable artificial intelligence: An overview. *J. Wash. Acad. Sci*, 2020.

41 ISLAM, S. R.; EBERLE, W.; GHAFOOR, S. K.; AHMED, M. *Explainable Artificial Intelligence Approaches: A Survey*. 2021.

42 ARRIETA, A. B.; DíAZ-RODRíGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCíA, S.; GIL-LóPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*. 2019.

43 RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939778>.

44 LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. Disponível em: <http://arxiv.org/abs/1705.07874>.

45 LUNDBERG, S. *Welcome to the SHAP documentation*. 2018. <https://shap.readthedocs.io/en/latest/index.html>. (Acessado em 02/12/2021).

46 FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001. ISSN 00905364. Disponível em: <http://www.jstor.org/stable/2699986>.

47 DEMŠAR, J.; CURK, T.; ERJAVEC, A.; GORUP, Č.; HOČEVAR, T.; MILUTINOVIČ, M.; MOŽINA, M.; POLAJNAR, M.; TOPLAK, M.; STARIČ, A.; ŠTAJDOHAR, M.; UMEK, L.; ŽAGAR, L.; ŽBONTAR, J.; ŽITNIK, M.; ZUPAN, B. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, v. 14, n. 71, p. 2349–2353, 2013. Disponível em: <http://jmlr.org/papers/v14/demsar13a.html>.

48 GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. 2014.

49 WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018.

50 KIM, B.; RUDIN, C.; SHAH, J. *The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification*. 2015.

51 PINA, D.; NEVES, L.; PAES, A.; OLIVEIRA, D. de; MATTOSO, M. Análise de hiperparâmetros em aplicações de aprendizado profundo por meio de dados de proveniência. In: *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*. Porto Alegre, RS, Brasil: SBC, 2019. p. 223–228. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/8827>.

52 PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, n. Oct, p. 2825–2830, 2011.

53 RUPPRECHT, L.; DAVIS, J. C.; ARNOLD, C.; GUR, Y.; BHAGWAT, D. Improving reproducibility of data science pipelines through transparent provenance capture. *Proc. VLDB Endow.*, VLDB Endowment, v. 13, n. 12, p. 3354–3368, aug 2020. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/3415478.3415556>.

54 SPINNER, T.; SCHLEGEL, U.; SCHÄFER, H.; EL-ASSADY, M. explainer: A visual analytics framework for interactive and explainable machine learning. *CoRR*, abs/1908.00087, 2019. Disponível em: <http://arxiv.org/abs/1908.00087>.

55 PUBLIO, G.; ESTEVES, D.; ŁAWRYNOWICZ, A.; PANOV, P.; SOLDATOVA, L.; SORU, T.; VANSCHOREN, J.; ZAFAR, H. Ml-schema: Exposing the semantics of machine learning with schemas and ontologies. 07 2018.

56 SOUZA, R.; AZEVEDO, L.; THIAGO, R.; SOARES, E.; NERY, M.; NETTO, M. A. S.; VITAL, E.; CERQUEIRA, R.; VALDURIEZ, P.; MATTOSO, M. Efficient runtime capture of multiworkflow data using provenance. In: *2019 15th International Conference on eScience (eScience)*. [S.l.: s.n.], 2019. p. 359–368.

57 HAN, R.; BYNA, S.; TANG, H.; DONG, B.; ZHENG, M. Prov-io: An i/o-centric provenance framework for scientific data on hpc systems. In: *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*. New York, NY, USA: Association for Computing Machinery, 2022. (HPDC '22), p. 213–226. ISBN 9781450391993. Disponível em: <https://doi.org/10.1145/3502181.3531477>.

58 SUPREM, A.; VAIDYA, S.; VENUGOPAL, A.; FERREIRA, J. E.; PU, C. *EdnaML: A Declarative API and Framework for Reproducible Deep Learning*. 2022.

59 KENNEDY, O.; GLAVIC, B.; FREIRE, J.; BRACHMANN, M. The right tool for the job: Data-centric workflows in vizier. *Bulletin of the Technical Committee on Data Engineering*, v. 45, n. 3, 2022.

60 NAKAGAWA, P. I.; PIRES, L. F.; MOREIRA, J. L. R.; SANTOS, L. O. Bonino da S.; BUKHSH, F. Semantic description of explainable machine learning workflows for improving trust. *Applied Sciences*, v. 11, n. 22, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/22/10804>.

61 GUIZZARDI, G.; WAGNER, G. A unified foundational ontology and some applications of it in business modeling. In: *CAiSE Workshops (3)*. [S.l.: s.n.], 2004. p. 129–143.

62 ALI, S.; ABUHMED, T.; EL-SAPPAGH, S.; MUHAMMAD, K.; ALONSO-MORAL, J. M.; CONFALONIERI, R.; GUIDOTTI, R.; Del Ser, J.; DíAZ-RODRíGUEZ, N.; HERRERA, F. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, v. 99, p. 101805, 2023. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253523001148>.

63 DWIVEDI, R.; DAVE, D.; NAIK, H.; SINGHAL, S.; OMER, R.; PATEL, P.; QIAN, B.; WEN, Z.; SHAH, T.; MORGAN, G.; RANJAN, R. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 9, jan 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3561048>.

64 ROSSUM, G. van. *Python reference manual*. 1995.

65 STONEBRAKER, M.; ROWE, L.; HIROHAMA, M. The implementation of postgres. *IEEE Transactions on Knowledge and Data Engineering*, v. 2, n. 1, p. 125–142, 1990.

66 Kaggle. *Titanic - Machine Learning from Disaster*. 2022. Disponível em: <https://www.kaggle.com/c/titanic/>. Acesso em: 26 de fevereiro 2022.

67 WOLLENSTEIN-BETECH, S.; CASSANDRAS, C. G.; PASCHALIDIS, I. C. Personalized predictive models for symptomatic covid-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an icu or ventilator. *International Journal of Medical Informatics*, v. 142, p. 104258, 2020. ISSN 1386-5056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S138650562030616X>.

68 SALUD, S. de. *Datos Abiertos Dirección General de Epidemiología*. 2020. <https://www.gob.mx/salud/documentos/datos-abiertos-152127>. (Acessado em 04/08/2021).

69 MUHAMMAD, L.; ALGEHYNE, E. A.; USMAN, S. S.; AHMAD, A.; CHAKRABORTY, C.; MOHAMMED, I. A. Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN computer science*, Springer, v. 2, n. 1, p. 1–13, 2021.

70 FRANKLIN, M. R. *"Kaggle: Mexico COVID-19 clinical data"*. 2020. Https://www.kaggle.com/marianarfranklin/mexico-COVID19-clinical-data/metadata. (Acessado em 02/10/2021).

71 SAúDE, M. da. *Entenda as diferenças entre RT-PCR, antígeno e autoteste*. 2022. <https://www.gov.br/saude/pt-br/assuntos/noticias/2022/fevereiro/entenda-as-diferencas-entre-rt-pcr-antigeno-e-autoteste>. (Acessado em 22/04/2023).

72 YADAV, A. Predicting covid-19 using random forest machine learning algorithm. In: *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. [S.l.: s.n.], 2021. p. 1–6.

73 CORNELIUS, E.; AKMAN, O.; HROZENCIK, D. Covid-19 mortality prediction using machine learning-integrated random forest algorithm under varying patient frailty. *Mathematics*, v. 9, n. 17, 2021. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/9/17/2043>.

74 LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, v. 23, n. 1, 2021. ISSN 1099-4300. Disponível em: <https://www.mdpi.com/1099-4300/23/1/18>.

APÊNDICE A – DESCRIÇÃO DAS TABELAS

Tabela Workflow: refere-se às informações do workflow.

Atributo	Descrição
id	Identificador da tabela Workflow
label	rótulo do workflow
timestamp	marca de tempo de início do workflow

Tabela Dataset: refere-se às informações do conjunto de dados.

Atributo	Descrição
id	Identificador da tabela Dataset
local	Local de armazenamento do dataset
Size	Tamanho do dataset
N_line	Número de linhas do dataset
N_column	Número de colunas do dataset

Tabela Operator: refere-se aos operadores de pré-processamento utilizados durante o workflow.

Atributo	Descrição
id	Identificador da tabela Operator
id_workflow	Identificador da tabela Workflow
name_operator	Nome do operador de pré-processamento
function_operator	Função do operador de pré-processamento

Tabela OperatorsActivity: refere-se às operações de pré-processamento realizadas nos atributos do dataset.

Atributo	Descrição
id	Identificador da tabela OperatorActivity
id_Operator	Identificador da tabela Operator
id_Dataset_Attribute	Id da tabela Dataset_Attribute

Tabela Dataset_Attribute: Refere-se aos atributos originais do dataset.

Atributo	Descrição
id	Identificador da tabela Dataset_attribute
id_dataset	Identificador da tabela Dataset
label	Rótulo do atributo
type	Tipo do atributo (int, object, float, etc.)
description	Descrição referente ao atributo do dataset

Tabela Experiment: Refere-se às informações de medidas de desempenho do

experimento.

Atributo	Descrição
Id	Identificador da tabela Experiment
timestamp	marca de tempo da execução do experimento
method	Algoritmo usado no experimento
accuracy	Medida da Acurácia do experimento
recall	Medida da abrangência do experimento
precision	Medida da precisão do experimento
F1-Score	Medida f1 do experimento

Tabela Experiment_Attribute: Refere-se aos atributos utilizados no treino do experimento.

Atributo	Descrição
Id	Identificador da tabela Experiment_Attribute
Experiment_id	Identificador da tabela Experiment
label	Rótulo do atributo
Type	Tipo do atributo (int, object, float, etc.)
origin	Indica se o atributo é original do dataset.

Tabela Parameter: Refere-se aos parâmetros usados no treinamento do experimento.

Atributo	Descrição
Id	Identificador da tabela Parameter
Experiment_id	Identificador da tabela Experiment
label	Rótulo do parâmetro
value	Valor do parâmetro

Tabela XAI_Method: Refere-se a Configuração do gráfico XAI.

Atributo	Descrição
Id	Identificador da tabela XAI
Method	Nome do método utilizado
Set_input	Dataframe utilizado para construção do gráfico
Idx_instance	Índice da instância a ser avaliada no dataframe
Max_features	Máximo de atributos a serem considerados

Tabela XAI_Method_Results: Refere-se aos resultados da XAI por *feature importance* e valor.

Atributo	**Descrição**
id	Identificador da tabela Results_XAI
id_XAI	Identificador da tabela XAI
label_Feature_importance	Nome do atributo usado no gráfico XAI
feature_importance_val	Valor de contribuição atribuído ao atributo no modelo

APÊNDICE B – DICIONÁRIO DE DADOS

BASE DE DADOS - TITANIC

Tabela 12 – Titanic Dataset

Atributo	Descrição	Categorias
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings/spouses	
parch	# of parents/children	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

BASE DE DADOS - COVID-19 - MÉXICO

Tabela 13 – Descrição dos atributos COVID-19 - México

Atributo	Descrição
id	Unique id to keep track of data while preprocessing
FECHA_ARCHIVO	Date where record was last updated
ID_REGISTRO	Unique id used by the original source to identify cases
ENTIDAD_UM	Region where hospital performed the admission
ENTIDAD_RES	Region of residence of the patient
RESULTADO	Result of the lab test for RT-PCR of COVID-19
DELAY	Lag (in days) between a reported case and the lab result confirming either positive or negative COVID-19 result
ENTIDAD_REGISTRO	The actual region where the case was officially assigned
ENTIDAD	Name of the region (state)
ABR_ENT	Simplified name of the region (state)
FECHA_ACTUALIZACION	This variable allows to identify the date of the last update (YYYY-MM-DD)

ORIGEN	Sentinel surveillance is carried out through the respiratory disease monitoring health unit system (USMER). The USMER includes medical units of the first, second or third level of care, and third level units also participate as USMERs
SECTOR	Identify the type of institution of the National Health System that provided the care
SEXO	Gender of the patient
ENTIDAD_NAC	The patient's birth region (state)
MUNICIPIO_RES	The patient's birth city
TIPO_PACIENTE	Type of care the patient received in the unit. It is called an outpatient if you returned home or it is called an inpatient if you were admitted to the hospital
FECHA_INGRESO	Date of admission of the patient to the care unit (YYYY-MM-DD)
FECHA_SINTOMAS	Date on which the patient's symptoms began (YYYY-MM-DD)
FECHA_DEF	Date the patient died
INTUBADO	Identifies if the patient required intubation
NEUMONIA	Identifies if the patient was diagnosed with pneumonia
EDAD	Age of the patient
NACIONALIDAD	Identifies if the patient is Mexican or foreign
EMBARAZO	Identifies if the patient is pregnant
HABLA_LENGUA_INDIG	Identifies if the patient speaks an indigenous language
DIABETES	Identifies if the patient is diabetic
EPOC	Identifies if the patient presents EPOC
ASMA	Identifies if the patient has asthma
INMUSUPR	Identifies if the patient is immunosuppressed
HIPERTENSION	Identifies if the patient has hypertension
OTRA_COM	Identifies if the patient presents another disease
CARDIOVASCULAR	Identifies if the patient has cardiovascular disease
OBESIDAD	Identifies if the patient has obesity
RENAL_CRONICA	Identifies if the patient presents chronic renal insufficiency
TABAQUISMO	Identifies if the patient has a tobacco addiction
OTRO_CASO	Identifies if the patient had contact with any other case diagnosed with SARS CoV-2
MIGRANTE	Identifies if the patient is immigrant

PAIS_NACIONALIDAD	Country of nationality
PAIS_ORIGEN	Country of origin
UCI	Identifies if the patient was admitted to Intensive Care Unit
DEAD	Identifies if the patient has died

APÊNDICE C – A INFLUÊNCIA DOS ATRIBUTOS NA EXPLICABILIDADE E NO RESULTADO DO MODELO

Neste apêndice é apresentado o comportamento do modelo e da sua explicabilidade após a criação de cada atributo construído. Para isso, a base de dados Titanic será utilizada. O processo de análise desse comportamento será conduzido conforme o seguinte procedimento. Inicialmente, o primeiro experimento é executado, focando exclusivamente nas operações de pré-processamento, tais como a remoção de atributos e a aplicação de técnicas para o preenchimento de valores ausentes. Nessa etapa, não houve a necessidade de criação de novos atributos nem de utilização de técnicas de codificação adicionais.

A partir do segundo experimento, operações adicionais foram gradualmente incorporadas ao processo, de forma incremental. Essas etapas incluem a implementação de outras técnicas de pré-processamento que permitem a construção progressiva de cada atributo, com o propósito de ressaltar o impacto que tais modificações causam no desempenho do modelo e, consequentemente, na capacidade de explicação do mesmo. Esse processo evolutivo culminou no resultado obtido no relatório de classificação do experimento 2 observável na Tabela 7, da seção 6.1.

Para o primeiro experimento, o objetivo consiste em simplificar o processo de treinamento do modelo, focando na redução significativa da etapa de pré-processamento. Assim, opta-se por evitar a aplicação de técnicas avançadas de codificação de atributos, direcionando a abordagem para a seleção de atributos numéricos.

Dentro dessa abordagem, procede-se à exclusão dos atributos *Name, Sex, Ticket, Cabin e Embarked* do conjunto de dados. Adicionalmente, o atributo *Age*, que inicialmente possuía valores ausentes, foi tratado por meio do preenchimento dos valores faltantes com a média dos valores disponíveis. O desempenho deste experimento pode ser observado no relatório de classificação da Tabela 14.

Tabela 14 – Relatório de Classificação - Base do Titanic - Primeiro experimento

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0 (Negativa)	71,00	94,00	81,00	157
1 (Positiva)	84,00	47,00	60,00	111
Acurácia			74,00	268
Média Macro	78,00	70,00	71,00	268
Média Ponderada	77,00	74,00	72,00	268

Ao comparar este relatório de classificação com o relatório de classificação (Tabela 6) referente ao primeiro experimento realizado com os dados do Titanic (seção 6.1) fica

evidente que há uma significativa redução no desempenho neste experimento. Isso pode ser observado pela grande queda da acurácia e pela redução dos valores das métricas Precisão, Revocação e Medida-F1 para a classe 1, e da precisão e Medida-F1 para a classe 0. O principal motivo para essa queda está relacionado à exclusão do atributo *Sex*. Esse atributo é de fundamental importância, uma vez que o atributo *Sex_male*, que se trata de um atributo *dummy* criado a partir da codificação do atributo *Sex*, desempenha um papel crucial na compreensão daquele experimento, como pode ser claramente observado no gráfico SHAP da Figura 31, onde este atributo aparece como o de maior contribuição para o modelo.

A Figura 48 exibe o gráfico SHAP para esse experimento, onde pode-se perceber que os atributos *Pclass* e *Fare* foram os que que mais contribuíram com o resultado do modelo.

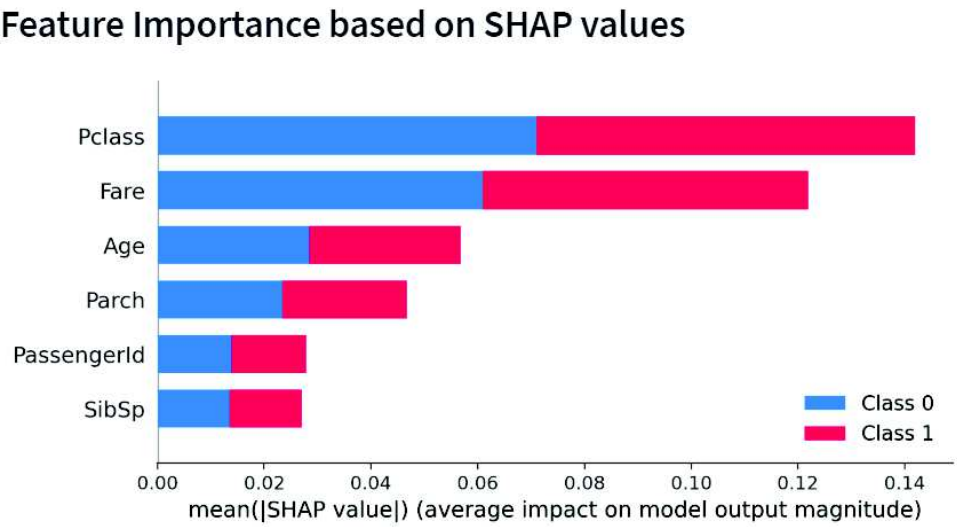

Figura 48 – Gráfico SHAP - Primeiro experimento

Para o segundo experimento, os atributos *Name, Ticket e Cabin* também foram removidos, porém os *Embarked e Sex* foram mantidos. Como estes atributos são categóricos, foi realizada a sua codificação. Esse experimento corresponde ao experimento 1 do Titanic, cujos resultados podem ser verificados pelo relatório de classificação da Tabela 6 (seção 6.1) e cuja explicação do modelo encontra-se no gráfico SHAP, representado pela Figura 31 (subseção 6.3.1).

Para o terceiro experimento, além das operações realizadas no segundo experimento, foi realizada a criação do atributo *Title* que é extraído do nome do passageiro e a sua codificação. Para este atributo, conforme realizado no experimento 2 da seção 6.1, também foram mantidos apenas os valores Master, Miss, Mr., Mrs., que apresentavam uma maior frequência. As instâncias, cujos valores diferiam destes, tiveram o valor deste atributo substituído por “Others” e, após a sua criação, o atributo foi codificado através do operador

OneHotEncoder. A Tabela 15 exibe o resultado do relatório de classificação para este experimento e a Figura 49 exibe o gráfico SHAP.

Tabela 15 – Relatório de Classificação - Base do Titanic - Terceiro experimento

Classe	Precisão	Revocação	Medida-F1	Suporte
0 (Negativa)	81,00	89,00	85,00	157
1 (Positiva)	82,00	71,00	76,00	111
Acurácia			82,00	268
Média Macro	82,00	80,00	81,00	268
Média Ponderada	82,00	82,00	81,00	268

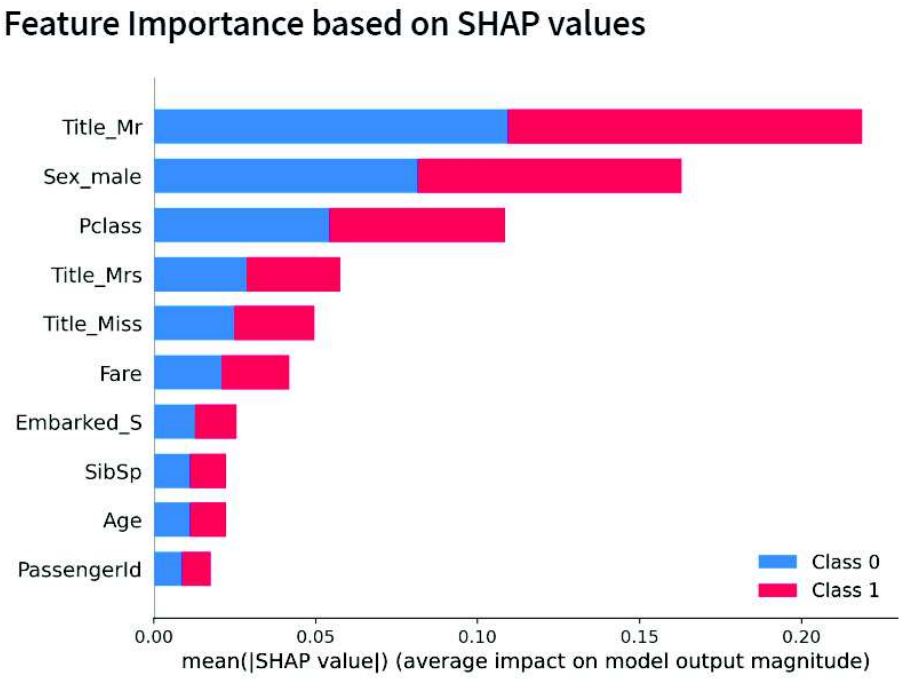

Figura 49 – Gráfico SHAP - Terceiro experimento

O resultado do desempenho apresentado nesta tabela ao ser comparado com o apresentado na Tabela 6 (seção 6.1) evidencia que a acurácia permaneceu inalterada. Entretanto, é possível identificar pequenas alterações nos valores de outras métricas. Em relação à classe 0, observou-se um aumento na precisão, mas houve uma diminuição nos valores de revocação e medida F1. Para a classe 1, houve um aumento na métrica de revocação, enquanto as demais métricas apresentaram redução.

Ao compararmos o gráfico da Figura 49 com o gráfico SHAP do experimento mostrado na Figura 31 (subseção 6.3.1), é possível perceber que os atributos de maior importância acarretam em resultados diferentes devido à introdução do atributo *Title* e suas categorias derivadas (*Title_Mr, Title_Mr, Title_Miss*).

No gráfico da Figura 31, os cinco atributos mais relevantes são *Sex_male*, *Pclass*, *Fare*, *Age* e *Embarked_s*. Já no gráfico da Figura 49, os atributos de maior importância

passam a ser *Title_Mr*, *Sex_male*, *Pclass*, *Title_Mrs* e *Title_Miss*. É importante observar que, embora a criação do atributo *Title* e suas categorias derivadas tenham influenciado a interpretação do modelo, especialmente o *Title_Mr* neste experimento, essa criação não resultou em um aumento do desempenho do modelo.

Para o quarto experimento, além das operações realizadas no terceiro experimento, foi realizada a criação do atributo *Groupsize*. O atributo *Groupsize*, como visto na seção 6.1, se refere ao número de mulheres ou crianças com o mesmo sobrenome e, provavelmente, da mesma família. A Tabela 16 apresenta o relatório de classificação para este modelo.

Tabela 16 – Relatório de Classificação - Base do Titanic - Quarto experimento

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0 (Negativa)	84,00	89,00	86,00	157
1 (Positiva)	83,00	76,00	79,00	111
Acurácia			84,00	268
Média Macro	84,00	82,00	83,00	268
Média Ponderada	84,00	84,00	83,00	268

Como pode ser observado no gráfico SHAP apresentado na Figura 50, o atributo *Groupsize* apresenta uma significativa contribuição para a classificação do modelo, sendo o de maior influência em sua predição. Ao analisarmos o relatório de classificação representado na Tabela 16, nota-se um aumento nos valores das métricas de precisão, revocação e medida-F1 para ambas as classes, além do incremento na acurácia do modelo.

Adicionalmente, é válido observar que, no contexto deste modelo, o atributo *Fare* exibe maior importância em relação aos atributos *Title_Mrs* e *Title_Miss*. Essa diferença é notável ao compararmos com o gráfico SHAP do experimento anterior, conforme ilustrado na Figura 49.

Para o quinto experimento (que corresponde ao segundo experimento da seção 6.1), além das operações realizadas no quarto experimento, foi efetuada a criação do atributo *Family_size*. Como discutido na subseção 6.3.1, este atributo representa a soma do número total de membros da família a bordo do Titanic e deriva dos atributos *Parch* e *SibSp* em cada instância. O relatório de classificação referente a esse experimento é apresentado na Tabela 7 (seção 6.1).

A Figura 32 (subseção 6.3.1) exibe o gráfico SHAP para o modelo deste experimento. Como é possível observar, o atributo construído *Family_Size* apresenta uma contribuição reduzida, ocupando a sétima posição em importância. No que diz respeito à explicação do modelo, ao compará-la com o gráfico SHAP do experimento anterior (conforme ilustrado na Figura 50), é perceptível que a introdução do atributo *Family_size* tem um impacto sutil na interpretação do modelo. Esse acréscimo provoca uma alteração na importância

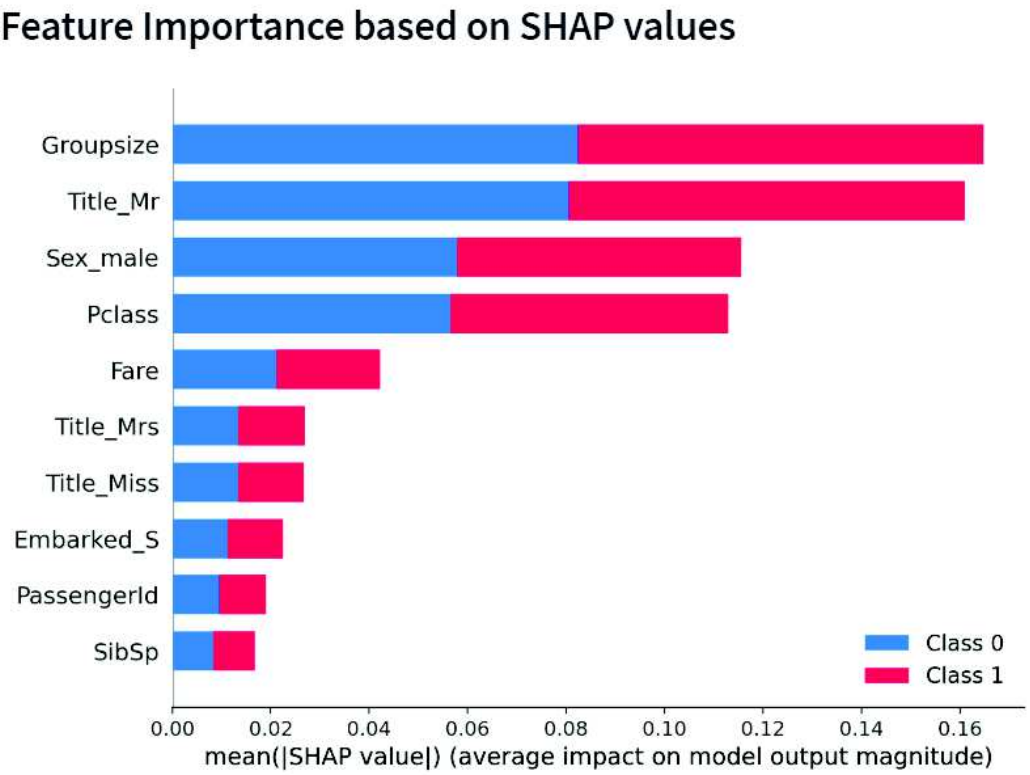

Figura 50 – Gráfico SHAP - Quarto Experimento

dos atributos, posicionando o atributo *Sex_male* na segunda posição e o novo atributo criado, *Family_size*, na sétima posição.

Além disso, este atributo não apresenta uma influência substancial no modelo, mantendo sua acurácia inalterada. Entretanto, para a classe 0, as métricas de revocação e medida-F1 apresentam uma melhora. Já para a classe 1, além dessas métricas, a precisão também registra um pequeno aumento em comparação com o experimento anterior (Tabela 16).

O sexto experimento foi conduzido com o intuito de exemplificar mais um benefício decorrente do entendimento das etapas iniciais de pré-processamento, especialmente no contexto da elaboração de atributos. É relevante salientar neste ponto que este atributo foi adicionado exclusivamente com o propósito de demonstrar uma instância da imperativa necessidade de compreensão acerca da construção dos dados. Isto se deve ao fato de que este atributo em particular não ter sido concretizado de maneira adequada, uma vez que ele emprega, em sua formulação, o atributo alvo denominado *Survived*.

Nesse contexto, introduz-se deliberadamente a criação de um novo atributo. Assim, para este experimento, além das operações realizadas no quinto experimento, foi realizada a criação do atributo *GroupSurvived*. Esse atributo representa a média de mulheres ou crianças sobreviventes que compartilham o mesmo sobrenome, sugerindo que possam pertencer à mesma família.

O Tabela 17 oferece o relatório de classificação para este experimento. Observa-se que o atributo em questão desempenha um papel de grande relevância no modelo, como indicado pelo impacto positivo nas métricas de ambas as classes.

No entanto, é fundamental compreender a origem da concepção desse novo atributo. Para tal, realizou-se uma consulta ao repositório de proveniência, visando rastrear a criação e a motivação por trás desse atributo recém-introduzido.

Tabela 17 – Relatório de Classificação - Base do Titanic - Sexto experimento

Classe	**Precisão**	**Revocação**	**Medida-F1**	**Suporte**
0 (Negativa)	86,00	99,00	92,00	157
1 (Positiva)	99,00	77,00	86,00	111
Acurácia			90,00	268
Média Macro	92,00	88,00	89,00	268
Média Ponderada	91,00	90,00	90,00	268

Pelo gráfico SHAP do experimento (Figura 51), podemos observar que o novo atributo, *GroupSurvived*, seguido do *Groupsize*, foram os atributos que mais contribuíram com o resultado do modelo. Na Figura 52, é apresentado o comando SQL com o resultado da consulta que revela a origem da criação do atributo *GroupSurvived*.

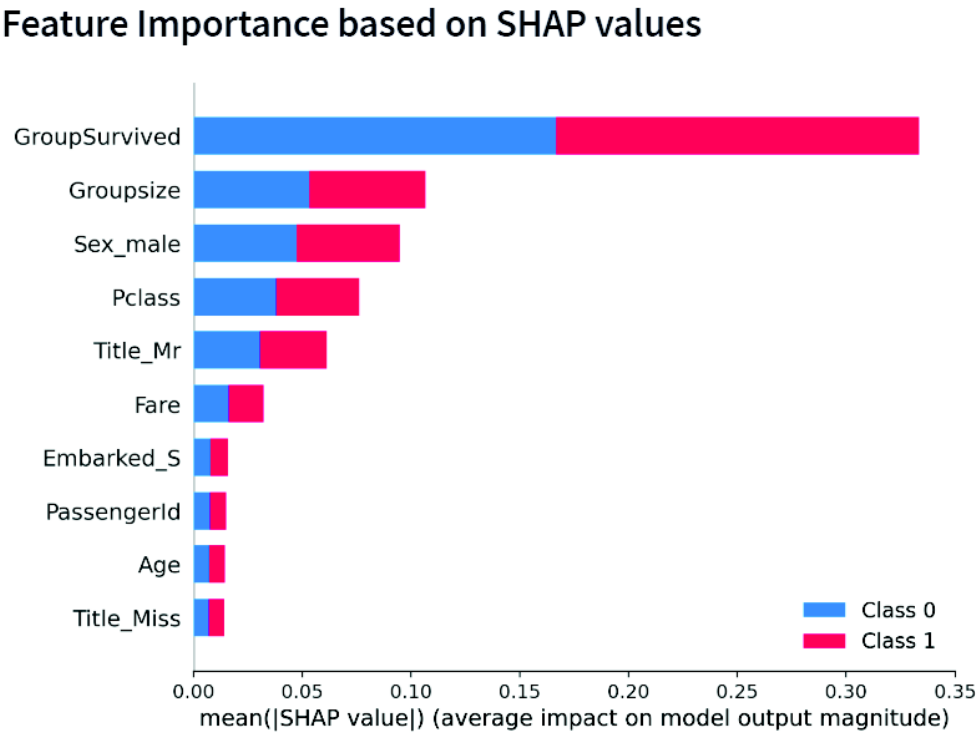

Figura 51 – Gráfico SHAP - Sexto experimento

Com base na análise da origem da construção do atributo, realizada ao consultar os dados de proveniência, é evidente que houve um erro na forma como o atributo foi

```
SELECT Experiment_attribute.label,
experiment_attribute.origin
FROM  experiment_attribute,
experiment, dataset, workflow
WHERE experiment_attribute.experimentid=
experiment.id
AND  experiment.datasetid=dataset.id
AND dataset.workflowid = workflow.id
AND  workflow.id = 163
AND experiment_attribute.label
LIKE 'GroupSurvived';
```

	attribute	source
0	GroupSurvived	df['GroupSurvived'] = df['Lastname'].apply(lambda x: df[['Title', 'Survived']].loc[df['Lastname']==x].loc[(df['Sex']=='female') | (df['Title']=='Master')].mean()['Survived'])

Figura 52 – Informações de origem do atributo GroupSurvived - o atributo alvo (Survived) é utilizado na sua construção

criado, uma vez que ele incorporou o próprio atributo alvo *Survived* em seu processo de construção. Isso, entretanto, pode conduzir a uma interpretação artificialmente positiva do desempenho do modelo. Essa situação ocorre devido à possibilidade de o modelo estar aprendendo com informações que não estariam disponíveis na prática. Além desse ponto, é importante ressaltar que a construção inadequada do referido atributo compromete sua utilidade no modelo. Portanto, torna-se inviável a sua utilização, visto que foi concebido de maneira imprópria.

Nesse contexto, evidencia-se a relevância do uso dos dados de proveniência não apenas para a condução de uma avaliação precisa e verdadeira dos dados empregados na construção do modelo, mas também para aumentar a confiança e a transparência em todo o processo. A compreensão da origem dos dados que compõem o modelo esclarece as decisões tomadas durante sua construção, o que aprimora a transparência e a validade das operações realizadas durante esse processo.