Ellen Leal dos Santos

Aprendizado de Máquina Automatizado na Preparação e Manipulação de Conjuntos de Treinamento

Relatório Final de Projeto de Seis Meses de Iniciação Científica Voluntária

São Paulo

2025

Ellen Leal dos Santos

Aprendizado de Máquina Automatizado na Preparação e Manipulação de Conjuntos de Treinamento

Relatório de trabalho de Iniciação Científica voltada à automatização do aprendizado de máquina para automatizar a criação e manipulação de conjuntos de treinamento. Este trabalho foi desenvolvido voluntariamente ao longo do segundo semestre de 2024 sob a orientação do Professor Doutor Luciano Antonio Digiampietri.

Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Bacharelado em Sistemas de Informação

São Paulo

2025

Resumo: Atualmente, muitas decisões que impactam a vida das pessoas são tomadas com o auxílio de algoritmos de inteligência artificial. Existem diversos tipos de algoritmos de aprendizado de máquina, cada um com capacidades diferentes para aprender padrões e com desempenhos variados dependendo do tipo e da quantidade de dados utilizados. Normalmente, esses algoritmos são usados para criar modelos que maximizam uma única métrica, como a acurácia. Preparar dados para treinamento, realizar a engenharia de características, escolher algoritmos e avaliar o desempenho são tarefas complexas e custosas. Este projeto de pesquisa explorou a Automatização do Aprendizado de Máquina (AutoML) para automatizar a criação e manipulação de conjuntos de treinamento, focando no balanceamento de dados (seleção de instâncias, criação de instâncias sintéticas, etc.) e manipulação de atributos sensíveis (exclusão, aleatorização, etc.). Para avaliar o desempenho dos modelos, o projeto avaliou medidas tradicionais, como acurácia e F1, e também de justiça algorítmica. A comparação de desempenho entre modelos considerados mais complexos e de difícil interpretação e aqueles considerados inerentemente explicáveis (ou interpretáveis) permite uma melhor avaliação entre de custo e benefício entre diferentes soluções. Assim, este projeto de AutoML visou a automatizar etapas da Mineração de Dados e contribuir para as áreas de Justiça Algorítmica e Inteligência Artificial Explicável.

Palavras-chave: Inteligência Artificial Explicável; Discriminação Algorítmica; Mineração de Dados

1. Introdução

Atualmente, diversas decisões que afetam o dia a dia das pessoas são tomadas parcialmente ou totalmente com auxílio de algoritmos de inteligência artificial. Existem diversos tipos de algoritmos de aprendizado de máquina que variam no tipo de padrões que são capazes de “aprender”, possuindo diferentes desempenhos de acordo com o tipo e a quantidade de dados que estão tratando. Adicionalmente, esses algoritmos tipicamente são utilizados para criar modelos que tentam maximizar uma única métrica, por exemplo, a acurácia. Neste contexto, a preparação dos conjuntos de treinamento de um algoritmo, o processo de extração, criação e seleção de características, a escolha do algoritmo, bem como a avaliação de seu desempenho (sob diferentes perspectivas) se tornam tarefas bastante complexas e custosas.

A Automatização das tarefas de Aprendizado de Máquina (AutoML) visa a automatizar algumas destas tarefas, possibilitando a escolha de boas combinações de estratégias de preparação dos dados, engenharia de características, seleção de algoritmos, escolha de hiperparâmetros, entre outras.

Este projeto explorou o AutoML em tarefas relacionadas a criação e manipulação do conjunto de treinamento visando a construção de modelos com melhor desempenho. No que diz respeito ao desempenho dos modelos treinados, eles foram avaliados não apenas com as medidas mais tradicionais aplicadas a modelos de classificação (acurácia e medida F1, por exemplo), mas também medidas relacionadas à justiça algorítmica (ASHOKAN e HAAS, 2021; BISWAS e RAJAN, 2020; LEE e FLORIDI, 2021; MANDHALA et al., 2022; SALAZAR et al., 2021).

2. Contextualização

Métricas de justiça no aprendizado de máquina são essenciais para avaliar e mitigar possíveis vieses nos modelos de inteligência artificial, garantindo que as decisões automatizadas sejam justas e imparciais. À medida que o aprendizado de máquina se torna cada vez mais integrado em decisões críticas, como admissões em universidades, concessão de empréstimos e processos de recrutamento, a justiça se torna uma preocupação central.

No presente trabalho foram calculadas as seguintes 19 métricas para medir a justiça algorítmica no treinamento dos modelos (ASHOKAN e HAAS, 2021; BISWAS e RAJAN, 2020; LEE e FLORIDI, 2021; MANDHALA et al., 2022; SALAZAR et al., 2021). As funções para o cálculo destas métricas foram implementadas em Python.

- **Equal Opportunity:** Grupos privilegiados e não privilegiados devem possuir as mesmas taxas de verdadeiros e falsos positivos.
- **Demographic Parity:** A probabilidade de obter resultados positivos deve ser igual entre grupos privilegiados e não privilegiados.
- **Disparate Impact:** Calculado pela razão entre a probabilidade de um grupo não privilegiado receber uma predição favorável e a do grupo privilegiado.
- **Equalized Odds:** Exige que as previsões de um classificador sejam independentes do atributo sensível, garantindo que ambos os grupos tenham taxas equilibradas de verdadeiros e falsos positivos.
- **Theil Index:** Mede a iniquidade entre todos os indivíduos, avaliando a distribuição de recursos ou predições.
- **Predictive Equality:** Busca o balanceamento da taxa de falsos positivos entre os dois grupos.
- **Average Odds Difference:** Reflete a média das diferenças entre as taxas de falsos positivos e verdadeiros positivos entre os grupos.
- **Disparate Mistreatment:** Analisa a discrepância na taxa de erro de classificação de grupos durante o processo de treinamento.
- **Non-Parity:** Representa a diferença absoluta na média das taxas preditas entre os grupos, sendo uma adaptação do conceito de *"statistical parity"*.
- **Positive Predictive Parity:** Dentro de um grupo favorável, as predições devem resultar em igualdade nos resultados, independentemente do atributo sensível.
- **Positive Class Balance:** Dentro de um grupo favorecido, deve haver igualdade entre indivíduos com diferentes atributos sensíveis.

- **Negative Class Balance:** Dentro de um grupo desfavorecido, também deve haver igualdade entre indivíduos com diferentes atributos sensíveis.
- **Individual Fairness:** Indivíduos que diferem apenas em seus atributos sensíveis devem receber a mesma previsão.
- **Counterfactual Fairness:** A decisão deve ser a mesma tanto no mundo real quanto em um cenário "paralelo", onde o indivíduo tem um atributo sensível favorável.
- **Equal Opportunity Difference:** Mede a diferença na taxa de verdadeiros positivos entre grupos privilegiados e não privilegiados.
- **Error Rate Difference:** Soma as taxas de falsos positivos e falsos negativos para medir discrepâncias entre os grupos.
- **Recall Difference:** Representa a diferença nas métricas de *recall* entre os grupos privilegiados e não privilegiados.
- **Difference in Positive Proportion:** É a diferença entre as proporções de predições positivas entre os grupos.
- **Difference in Rejection Rates:** Calcula a diferença entre as razões de verdadeiros positivos e negativos previstos entre grupos privilegiados e não privilegiados.

3. Objetivo

O objetivo principal desse projeto é explorar a Automatização do Aprendizado de Máquina (AutoML) para automatizar a criação e manipulação de conjuntos de treinamento, visando a construção de modelos com melhor desempenho. Isso inclui o balanceamento das instâncias do conjunto de treinamento (incluindo a seleção automatizada de subconjuntos de instâncias, criação de instâncias sintéticas) e manipulação do atributo considerado sensível. Além de maximizar métricas tradicionais como acurácia e medida F1, o projeto avaliou as medidas de justiça algorítmica apresentadas na seção 2. Ao automatizar etapas complexas e custosas da Mineração de Dados, e avaliar os resultados com diferentes métricas e pensando em diferentes contextos, o projeto busca contribuir para a Justiça Algorítmica e a Inteligência Artificial Explicável, promovendo a criação de

modelos que sejam não apenas eficazes, mas também transparentes e mais justos.

4. Metodologia

As primeiras semanas de desenvolvimento do projeto foram dedicadas a revisão bibliográfica, estudando artigos científicos focados na construção de conjuntos de treinamento e engenharia de características, de forma a entender o escopo, os objetivos específicos, as técnicas utilizadas, os métodos de validação e medição de desempenho, os resultados obtidos e as limitações, além de artigos voltados para métricas relacionadas à justiça algorítmica.

4.1 Conjuntos Dados

Após a revisão da literatura, foram selecionados os quatro conjuntos de dados públicos mais frequentemente utilizados nos trabalhos correlatos.

4.1.1 Adult Income

O conjunto traz informações de renda anual de uma pessoa, seu nível de educação, idade, gênero, ocupação, entre outros, somando 14 atributos ao total, sendo a renda anual a variável alvo do conjunto.

Foi considerado o atributo “Raça” como o atributo o sensível, sendo “Branca” ou *True*, a classe privilegiada.

4.1.2 Breast Cancer

O conjunto possui informações de pacientes com câncer de mama. Com 9 atributos, por exemplo: tamanho do tumor, idade e lado do tumor. O atributo alvo é a re-ocorrência do câncer.

O atributo sensível desse conjunto foi o “irradiat_yes” , sendo 1 ou *True*, a classe privilegiada.

4.1.3 German Credit

Com 9 atributos, este conjunto de dados inclui informações de crédito alemão, como: idade, sexo, ocupação, risco de crédito, etc. Sendo o atributo alvo o risco da concessão de crédito.

O atributo “Sex_male” foi usado como atributo sensível e o grupo privilegiado composto por indivíduos com o valor desse atributo como 1 ou *True*.

4.1.4 Ricci

Os atributos desse conjunto de dados são: raça, cargo, no exame oral, nota no exame escrito e nota final (combinação das notas dos dois exames de promoção de bombeiros como parte do processo judicial Ricci. O atributo alvo indica se a pessoa foi ou não promovida a capitão.

O atributo sensível dele é a raça, sendo 1 ou *True* a classe favorecida.

4.2 Balanceadores

Foram pesquisados e selecionados nove algoritmos para o balanceamento do conjunto de treinamento. A implementação utilizada foi a disponível na biblioteca *imblearn* do Python.

4.2.1 Oversampling

Realiza cópias (reamostragem) de amostras da classe minoritária para equilibrar o conjunto de dados.

4.2.2 Undersampling

Reduz o número de amostras da classe majoritária para equilibrar o conjunto de dados, balanceando o conjunto de dados (com o mesmo número de instâncias para cada classe).

4.2.3 SMOTE (Synthetic Minority Over-sampling Technique)

Cria novas amostras sintéticas da classe minoritária interpolando entre amostras existentes.

4.2.4 ADASYN (Adaptive Synthetic Sampling)

Gera amostras sintéticas da classe minoritária, mas dá mais peso a amostras difíceis de classificar, adaptando-se às distribuições locais.

4.2.5 Tomek Links

Identifica pares de amostras de classes opostas próximas entre si e remove as amostras da classe majoritária para reduzir o *overlap*.

4.2.6 SMOTETomek

Combina SMOTE para sobreamostrar a classe minoritária com Tomek Links, que remove amostras ambíguas da classe majoritária.

4.2.7 SMOTEENN

Combina SMOTE para sobreamostrar com Edited Nearest Neighbors (ENN), que remove amostras mal classificadas da classe majoritária.

4.2.8 BorderlineSMOTE

Gera amostras sintéticas apenas nas regiões de fronteira entre as classes, onde os dados são mais difíceis de classificar.

4.2.9 KMeansSMOTE

Aplica K-Means para dividir a classe minoritária em clusters e, em seguida, aplica SMOTE dentro de cada cluster para gerar novas amostras.

4.3 Classificadores

Foram utilizados nove algoritmos de classificação disponíveis na biblioteca sklearn do Python, utilizando os valores padrão para os hiperparâmetros, exceto aqueles explicitados a seguir: Dummy, DecisionTree, RandomForest, SVC Linear e polinomial, LogisticRegression com número máximo de iterações igual a 5.000, MLP com iteração máxima igual a 5.000, GaussianNB, KNeighbors com número de vizinhos igual a 5.

4.4 Métodos

Foi desenvolvido um conjunto de ferramentas que, dado cada conjunto de dados e o conjunto de classificadores, o conjunto de treinamento é construído e manipulado, treinando os modelos e

avaliando de forma a se investigar os efeitos de diferentes abordagens de manipulação do conjunto de treinamento no desempenho desses modelos.

Os quatro conjuntos de dados selecionados (apresentados na seção 4.1) foram divididos de forma estratificada aleatória em 70% das instâncias para treinamento e 30% para testes. Realizamos diferentes manipulações no conjunto de treinamento, conforme apresentado nas próximas subseções.

4.4.1 Balanceamento e treinamento

O primeiro conjunto de ferramentas implementado renomeia as colunas de atributo sensível e de variável alvo, seleciona as variáveis numéricas, cria conjuntos de treinamento para cada um dos balanceadores apresentados na seção 4.2, em seguida treina esses conjuntos com os classificadores e por fim, monta a matriz de confusão, as taxas de verdadeiros positivos e falsos positivos e calcula as métricas apresentadas na seção 2.

4.4.2 Balanceamento e treinamento sem o atributo sensível

Nessa abordagem o objetivo foi entender se remover o atributo sensível seria uma boa estratégia para melhorar as métricas de justiça. Ela consistiu em excluir a coluna do atributo sensível do conjunto de treinamento na fase de seleção de características, fazendo com que todos os processos a seguir, que são exatamente iguais aos da subseção 4.4.1, sejam feitos sem o atributo sensível. Ao fim, no conjunto de teste, o atributo sensível é mantido, pois é necessário para o cálculo de algumas métricas.

4.4.3 Balanceamento e treinamento com atributo sensível desorganizado

Na última abordagem, foi calculada a proporção de instância com cada valor do atributo sensível do conjunto de treinamento e com o método *choice* da função *random* da biblioteca NumPy, foi feita a desorganização do atributo sensível de cada um dos conjuntos de treinamento. As outras etapas foram as mesmas explicitadas na seção 4.4.2.

5. Resultados

A presente seção apresenta e discute os resultados. As tabelas 1, 2, 3 e 4 contêm os resultados para cada um dos quatro conjuntos de dados utilizados. Estas tabelas contêm, das colunas 2 a 7, a variação das medidas Acurácia e F1 macro, em relação aos valores das respectivas métricas considerando o conjunto de treinamento original, considerando tanto a estratégia de balanceamento quanto mudanças no atributo sensível (exclusão ou aleatorização no conjunto de treinamento). Já as três últimas colunas de cada tabela contém a quantidade de métricas de justiça que obtiveram melhorias em seus valores em relação aos valores obtidos pelo modelo utilizando o conjunto de treinamento original (sem balanceamento e sem nenhuma manipulação no atributo sensível). As cores de cada célula da tabela indicam os maiores valores da respectiva coluna em tons de verde mais intensos e os menores em tons de vermelho mais intenso.

Os resultados de cada abordagem serão discutidos individualmente nas subseções a seguir.

As tabelas completas com todos os resultados podem ser encontradas nos seguintes links: Adult Income, Breast Cancer, German Credit e Ricci.

Tabela 1 - Resultados das abordagens aplicadas ao conjunto de dados Adult Income

	Acurácia			F1 macro			Melhoria em Métricas de Justiça		
Estratégia de Balanceamento	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível
Oversampling	-0,025	-0,031	-0,024	-0,003	0,001	-0,004	1	5	3
Undersampling	-0,041	-0,035	-0,033	-0,014	-0,002	-0,010	2	6	3
SMOTE	-0,026	-0,023	-0,030	-0,005	0,005	-0,009	2	6	4
ADASYN	-0,042	-0,050	-0,054	-0,016	-0,014	-0,025	7	9	9
SMOTETomek	-0,026	-0,031	-0,025	-0,010	0,001	-0,005	1	6	6
SMOTEENN	-0,021	-0,008	-0,001	-0,001	0,013	0,006	10	6	4
Tomek Links	-0,004	0,002	0,001	-0,014	0,009	-0,005	16	7	15
BorderlineSMOTE	-0,050	-0,047	-0,046	-0,022	-0,012	-0,020	7	11	11
KMeansSMOTE	-0,004	-0,001	0,003	-0,021	-0,007	0,000	14	7	7

Tabela 2 - Resultados das abordagens aplicadas ao conjunto de dados Breast Cancer

	Acurácia			F1 macro			Melhoria em Métricas de Justiça		
Estratégia de Balanceamento	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível
Oversampling	0,003	0,008	0,001	0,007	0,013	0,006	1	9	5
Undersampling	0,009	0,006	0,003	0,015	0,012	0,008	1	10	2
SMOTE	0,008	0,010	0,010	0,014	0,016	0,016	1	9	2
ADASYN	0,012	0,014	0,009	0,018	0,021	0,015	7	9	8
SMOTETomek	0,008	0,013	0,008	0,014	0,019	0,014	4	10	1
SMOTEENN	0,001	0,001	-0,004	0,008	0,007	0,003	7	15	7
Tomek Links	0,000	0,001	-0,004	0,001	0,001	-0,007	3	17	3
BorderlineSMOTE	0,006	0,013	0,010	0,012	0,020	0,017	5	10	5
KMeansSMOTE	0,003	0,004	-0,001	0,007	0,009	0,000	2	9	3

Tabela 3 - Resultados das abordagens aplicadas ao conjunto de dados German Credit

	Acurácia			F1 macro			Melhoria em Métricas de Justiça		
Estratégia de Balanceamento	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível
Oversampling	-0,029	-0,014	-0,016	0,028	0,011	0,010	5	6	7
Undersampling	-0,049	-0,026	-0,009	0,007	0,034	0,034	7	5	5
SMOTE	-0,014	-0,014	-0,019	0,036	0,040	0,036	5	5	10
ADASYN	-0,027	-0,038	-0,040	0,033	0,023	0,020	4	5	6
SMOTETomek	-0,019	-0,017	-0,058	0,034	0,008	-0,005	5	6	6
SMOTEENN	-0,110	-0,114	-0,122	-0,055	-0,030	-0,036	12	7	7
Tomek Links	-0,036	-0,010	-0,044	0,004	0,026	-0,009	5	5	4
BorderlineSMOTE	-0,042	-0,011	-0,007	0,016	0,042	0,047	4	6	5
KMeansSMOTE	-0,004	-0,014	-0,024	0,003	0,018	0,011	11	7	7

Tabela 4 - Resultados das abordagens aplicadas ao conjunto de dados Ricci

Estratégia de Balanceamento	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível	Apenas Balanceamento	Remoção do Atributo Sensível	Aleatorização do Atributo Sensível
Oversampling	0,028	0,009	0,028	0,113	0,068	0,113	0	2	0
Undersampling	0,040	0,000	0,040	0,126	0,061	0,126	0	4	0
SMOTE	0,040	-0,046	0,040	0,127	0,013	0,127	2	2	2
ADASYN	0,046	0,003	0,046	0,133	0,064	0,133	2	0	2
SMOTETomek	0,046	0,000	0,046	0,132	0,060	0,132	2	2	2
SMOTEENN	-0,052	-0,090	-0,052	0,007	-0,054	0,007	0	2	0
Tomek Links	0,043	0,000	0,043	0,098	0,029	0,098	0	2	0
BorderlineSMOTE	0,052	0,012	0,052	0,139	0,074	0,139	0	1	0
KMeansSMOTE	0,040	0,003	0,040	0,120	0,055	0,120	1	0	1

5.1 Balanceamento e treinamento

Foi observado que a combinação de classificador e estratégia de balanceamento teve efeitos variados de acordo com o conjunto de

dados, não sendo possível encontrar uma combinação que seja a mais adequada para todos os conjuntos de dados.

Considerando o conjunto Adult Income, utilizado para prever a renda anual de uma pessoa com base em algumas características, observa-se que a não utilização de balanceamento produziu os melhores resultados de acurácia para três dos oito modelos testados. O uso do algoritmo de balanceamento KMeansSMOTE obteve as melhores acurácias para três modelos e do Tomek Links para dois. Em termos de medida F1 macro, o treinamento sem balanceamento atingiu o melhor valor para apenas um modelo. Destaca-se o balanceamento usando SMOTEENN que atingiu os melhores resultados em três. Em termos de métricas de justiça, o balanceamento usando Tomek Links fez com que 16 das 19 métricas ligadas à justiça tivessem seus valores melhorados em relação ao treinamento sem balanceamento. Destacam-se, em seguida, os algoritmos KMeansSMOTE (melhoria em 14 métricas) e SMOTEENN (melhoria em 10 métricas).

Os resultados para o conjunto Breast Cancer foram bastante diferentes. Trata-se de um conjunto simplificado com informações para tentar prever a ocorrência de câncer de mama. A estratégia de balanceamento ADASYN atingiu a melhor acurácia para todos os modelos, empatando com outras estratégias para alguns classificadores. Em termos de medida F1 macro, essa estratégia obteve o melhor desempenho para sete dos oito modelos. Considerando as medidas de justiça, duas abordagens de balanceamento (ADASYN e SMOTEENN) obtiveram resultados semelhantes, melhorando o valor de sete métricas em relação ao treinamento sem balanceamento dos dados.

O conjunto Ricci é relacionado à promoção de bombeiros em sua carreira. Para este modelo, no geral, as melhores métricas de justiça foram atingidas sem o balanceamento do conjunto de treinamento. Porém, o balanceamento usando BorderlineSMOTE permitiu um aumento de acurácia, na média, superior a 5% e de medida F1 macro superior a 13,9%.

No conjunto German Credit os melhores resultados de acurácia consideraram o conjunto sem balanceamento. Já a medida F1 macro teve leve melhora (entre 3% e 4% utilizando os algoritmos SMOTE,

ADASYN e SMOTETomek). Em termos de justiça, destacam-se as estratégias SMOTEENN (melhora em 12 métricas em relação ao treinamento original) e KMeansSMOTE (melhora em 11 métricas).

5.2 Balanceamento e treinamento sem o atributo sensível

Nos resultados para o conjunto German Credit, os melhores valores de acurácia foram obtidos com o conjunto sem balanceamento. Na métrica F1 macro, o balanceador TomekLinks proporcionou os melhores resultados. Em termos de métricas de justiça, os balanceadores SMOTEENN e KMeansSMOTE se destacaram, melhorando 7 das 19 métricas avaliadas.

Para o conjunto Ricci, a estratégia BorderlineSMOTE mostrou-se superior tanto em acurácia quanto em F1 macro, alcançando os melhores resultados em 7 dos 8 modelos. Já nas métricas de justiça, o método Undersampling apresentou melhorias em 4 delas.

No conjunto Adult Income, o balanceador TomekLinks garantiu os melhores resultados de acurácia em 6 dos 8 modelos testados, enquanto o SMOTEENN obteve os maiores valores para a medida F1 macro em 4 modelos. Em relação às métricas de justiça, o BorderlineSMOTE se destacou, melhorando 11 delas em comparação com o conjunto sem balanceamento.

Para o conjunto Breast Cancer, os balanceadores ADASYN, SMOTEENN e BorderlineSMOTE atingiram os melhores valores de acurácia e F1 macro em 6 dos 8 modelos analisados. Quanto às métricas de justiça, o TomekLinks liderou com melhorias em 17 métricas, seguido pelo SMOTEENN, que melhorou 15.

5.3 Balanceamento e treinamento com atributo sensível desorganizado

Nessa abordagem, o conjunto Adult Income obteve a melhor acurácia em 3 dos 8 modelos tanto na abordagem sem balanceamento quanto com o balanceador KMeansSMOTE. Em relação à medida F1 macro, o balanceador SMOTEENN apresentou os melhores resultados, liderando em 4 modelos. Nas métricas de justiça, o TomekLinks e o BorderlineSMOTE se destacaram, com melhorias em 15 e 11 métricas, respectivamente.

Para o conjunto Breast Cancer, os balanceadores SMOTE, ADASYN e SMOTEENN alcançaram os melhores valores de acurácia e F1 macro em 7 dos 8 modelos. No que diz respeito às métricas de justiça, o ADASYN obteve melhoria em 8 métricas, seguido pelo SMOTEENN, que aprimorou 7 delas.

No caso do conjunto German Credit, o conjunto sem balanceamento resultou na melhor acurácia em 6 dos 8 modelos. A medida F1 macro teve seus melhores valores também em 6 modelos, mas com a abordagem Tomek Links. Nas métricas de justiça, a técnica SMOTE melhorou 10 das 19 métricas avaliadas, se destacando entre os métodos.

No conjunto Ricci, tanto a acurácia quanto a F1 macro atingiram seus melhores valores com o balanceador BorderlineSMOTE. Nas métricas de justiça, as técnicas SMOTE, ADASYN e SMOTETomek apresentaram melhorias modestas, cada uma aprimorando apenas 2 métricas.

6. Conclusões

Algoritmos de aprendizado de máquina são cada vez mais utilizados em diversas aplicações, porém, a variedade de métodos, opções de parametrização, técnicas de balanceamento e seleção de atributos tornam o treinamento adequado desses modelos uma tarefa complexa. Neste projeto, desenvolvemos ferramentas que combinam diferentes abordagens de treinamento para identificar a configuração mais adequada para cada conjunto de dados, considerando múltiplas métricas de desempenho e justiça. Os resultados obtidos demonstram que técnicas específicas podem otimizar tanto a acurácia quanto a justiça dos modelos, dependendo das características dos dados. Assim, o trabalho contribui para um processo de treinamento mais direcionado e justo, facilitando a escolha de estratégias eficazes para diferentes cenários.

7. Bibliografia e Referências Bibliográficas

AGARWAL, A.; BELGRAVE, D.; CHO, K.; OH, A. (Ed.). Advances in Neural Information Processing Systems. Curran Associates, Inc., 2022. v. 35, p. 14170–14182.

ASHOKAN, A.; HAAS, C. Fairness metrics and bias mitigation strategies for rating predictions. Information Processing and Management, v. 58, n. 5, 2021.

BISWAS, S.; RAJAN, H. Do the machine learning models on a crowdsourced platform exhibit bias? an empirical study on model fairness. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. New York, NY, USA: Association for Computing Machinery, p. 642–653, 2020.

LEE, M. S. A.; FLORIDI, L. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds and Machines, v. 31, n. 1, p. 165–191, 2021.

MANDHALA, V. N.; BHATTACHARYYA, D.; MIDHUNCHAKKARAVARTHY, D.; KIM, H.-J. Mitigating bias by optimizing the variance between privileged and deprived data using post processing method. Rev. D Intell. Artif., International Information and Engineering Technology Association, v. 36, n. 1, p. 87–91, 2022.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, v. 1, n. 5, p. 206–215, May 2019. ISSN 2522-5839.

SALAZAR, T.; SANTOS, M. S.; ARAUJO, H.; ABREU, P. H. FAWOS: Fairness-aware oversampling algorithm based on distributions of sensitive attributes. IEEE Access, Institute of Electrical and Electronics Engineers (IEEE), v. 9, p. 81370–81379, 2021.

SHAMSABADI, A. S.; YAGHINI, M.; DULLERUD, N.; WYLLIE, S.; AÏVODJI, U.; ALAAGIB, A.; GAMBS, S.; PAPERNOT, N. Washing the unwashable : On the (im)possibility of fairwashing detection. In: KOYEJO, S.; MOHAMED, S.;