UMA ABORDAGEM NÃO SUPERVISIONADA PARA RECONSTRUÇÃO JUSTA DE DADOS

Felipe Bezerra de Melo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Setembro de 2023

UMA ABORDAGEM NÃO SUPERVISIONADA PARA RECONSTRUÇÃO JUSTA DE DADOS

Felipe Bezerra de Melo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Geraldo Zimbrão da Silva

Aprovada por: Prof. Geraldo Zimbrão da Silva
Prof. Geraldo Bonorino Xexéo
Prof. Filipe Braida do Carmo
Prof. Leandro Guimarães Marques Alvim

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2023

Bezerra de Melo, Felipe

Uma Abordagem não Supervisionada Para Reconstrução Justa de Dados/Felipe Bezerra de Melo. – Rio de Janeiro: UFRJ/COPPE, 2023.

XIII, 57 p.: il.; 29, 7cm.

Orientador: Geraldo Zimbrão da Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2023.

Referências Bibliográficas: p. 54 – 57.

1. Fairness. 2. Autoencoder. 3. Fair Representation. I. Zimbrão da Silva, Geraldo. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Esta dissertação conclui um ciclo muito importante na minha vida pessoal e acadêmica. Os desafios para sua conclusão foram muitos, porém tanto os conhecimentos adquiridos nesta jornada quanto as pessoas que estiveram ao meu lado contribuíram para meu crescimento.

Inúmeras pessoas estiveram presentes e foram fundamentais nesta jornada, cada uma à sua maneira. Portanto, sinto uma eterna gratidão àqueles que estiveram ao meu lado. Primeiramente, à minha mãe, que foi meu suporte ao longo da vida e cuja força e determinação sempre foram uma fonte de inspiração para mim.

À minha esposa, Bianca, que me apoiou em todos os momentos. Sua paciência, compreensão e amor foram essenciais durante este processo. Você não apenas acreditou em mim, mas também me lembrou do meu valor nos momentos em que duvidei de mim mesmo.

Ao professor Geraldo Zimbrão, pela orientação e apontamentos precisos no decorrer da pesquisa, sempre buscando extrair o melhor de nós.

Ao Leandro Alvim e Filipe Braida, pela orientação e ajuda durante o processo desta dissertação.

Um agradecimento especial aos meus amigos de graduação da UFRRJ: Pedro, Wilson, Lucas, Anderson e Marcus, para citar alguns. Os momentos de estudo e descontração foram essenciais.

Por último, agradeço a todos os meus professores. Cada um deles contribuiu de alguma forma para eu me tornar a pessoa que sou hoje. E aos professores da UFRRJ de Ciência da Computação, não havia lugar melhor para me formar como profissional. Lá, encontrei docentes dedicados e comprometidos com a formação de todos os alunos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA ABORDAGEM NÃO SUPERVISIONADA PARA RECONSTRUÇÃO JUSTA DE DADOS

Felipe Bezerra de Melo

Setembro/2023

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

A utilização de sistemas inteligentes de tomada de decisão, que se baseiam em conjuntos de dados de indivíduos, pode, se não tratada corretamente, perpetuar vieses preconceituosos presentes nos dados, oriundos de decisões historicamente enviesadas. Esse viés frequentemente está associado a atributos demográficos, como gênero, idade, raça, entre outros. Nesse contexto, pesquisas na área de *fairness* buscam propor métodos e métricas para identificar e mitigar o impacto negativo que algoritmos de aprendizado de máquina podem causar em grupos historicamente desfavorecidos, especialmente quando trabalham com bases de dados potencialmente enviesadas. Este trabalho apresenta um método de pré-processamento destinado a conjuntos de dados que possuam vieses relacionados à *fairness*. Por meio de um autoencoder com uma função de custo personalizada, o método é capaz de extrair variáveis latentes desses dados que apresentem um menor grau de injustiça em comparação aos dados originais. Para validar a hipótese do método proposto, foram realizados experimentos com dados reais, e as principais métricas da literatura foram empregadas para medir sua eficiência.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AN UNSUPERVISED APPROACH TO FAIR DATA RECONSTRUCTION

Felipe Bezerra de Melo

September/2023

Advisor: Geraldo Zimbrão da Silva

Department: Systems Engineering and Computer Science

The use of intelligent decision-making systems, which rely on individual data sets, can, if not properly addressed, perpetuate prejudiced biases present in the data, stemming from historically biased decisions. This bias is often associated with demographic attributes, such as gender, age, race, among others. In this context, research in the field of fairness seeks to propose methods and metrics to identify and mitigate the negative impact that machine learning algorithms can have on historically disadvantaged groups, especially when dealing with potentially biased data sets. This work introduces a pre-processing method designed for data sets that have biases related to fairness. Through an autoencoder with a customized cost function, the method is capable of extracting latent variables from these data that show a lower degree of injustice compared to the original data. To validate the hypothesis of the proposed method, experiments were conducted with real data, and the main metrics from the literature were used to measure its efficiency.

Sumário

Lista de Figuras **x**

Lista de Tabelas **xiii**

1 Introdução **1**

1.1 Motivação 1
1.2 Objetivos 2
1.3 Resumo dos Resultados 3
1.4 Contribuições 3
1.5 Organização do Trabalho 3

2 *Fairness* **4**

2.1 Contexto 4
2.2 Causas 6
2.3 Métricas 8
2.4 Abordagens e Soluções 9
2.5 Fair Representation 11

3 Aprendizado de Máquina **13**

3.1 Contexto 13
3.2 Aprendizado supervisionado 14
3.3 Aprendizado Não Supervisionado 15
3.4 Aprendizado por Reforço 16
3.5 Modelos de Aprendizado 17
3.5.1 Redes Neurais Artificiais 17
3.5.1.1 *Perceptron* 19
3.5.1.2 Arquiteturas de Redes Neurais 20
3.5.1.3 Autoencoders 21
3.5.2 *Support Vector Machine* (SVM) 24
3.6 Aprendizado com Viés 25

4 *Covariance Fair Autoencoder* 27
4.1 Definição do Problema 27
4.2 *Covariance Fair Autoencoder* 28
4.2.1 *Fair Autoencoder* 29
4.2.2 Variáveis Latentes 30
4.2.3 *Covariance Loss Function* 31
4.3 Trabalhos Relacionados 33

5 Resultados e Discussões 36
5.1 Base de Dados 36
5.2 Organização dos Experimentos 38
5.2.1 Configuração dos Hiper-parâmetros 39
5.3 Resultados 42

6 Conclusões 51
6.1 Objetivo do trabalho 51
6.2 Contribuições 52
6.3 Limitações e possíveis trabalhos futuros 52

Referências Bibliográficas 54

Lista de Figuras

3.1 Representação de como o aprendizado supervisionado funciona, onde o modelo tenta generalizar a partir dos dados de treino. 15
3.2 Representação de como o aprendizado não supervisionado funciona, onde o modelo tenta agrupar os dados de treino sem conhecer o rótulo. 16
3.3 Representação do neurônio biológico. Onde os dendritos funcionam como entrada de sinais de outros neurônios, o núcleo que processa e produz ou não uma saída de sinal pelo axônio. 18
3.4 Representação gráfica do perceptron, onde um neurônio artificial combina um conjunto de entradas x_1, x_2, x_3 com os parâmetros w_1, w_2, w_3 para produzir uma saída. 19
3.5 Representação de uma Rede Neural Artificial com 5 entradas, uma camada escondida e uma saída. 20
3.6 Representação de uma Rede Neural Recorrente onde a saída dos neurônios da camada interna também servem de entrada para a própria camada interna. 22
3.7 Representação de um *autoencoder*, com uma camada de *encoder*, uma camada oculta destacada como variáveis latentes, uma camada de *decoder* com objetivo de fazer a reconstrução dos dados original na camada de saída. 23
3.8 Representação gráfica de como SVM define uma margem para separar os dados em duas classes. 25

4.1 Processo de *fair representation* com um autoencoder capaz de gerar dados mais justos que pode ser utilizado por outros modelos de *machine learning*. 28
4.2 Relações lineares possíveis que duas variáveis podem ser determinadas pela função de covariância. (a) Relação direta (b) Relação inversa (c) Independente 32

5.1 Quantidade de indivíduos do grupo protegido, mulheres, e não protegido, homens, pertencentes a classe positiva e negativa da Base de dados *Adult Income*. 37
5.2 (a) Quantidade de indivíduos do grupo protegido, idade menor ou igual a 25, e não protegido, idade maior do que 25, pertencentes a classe positiva e negativa da base *German Credit* (b) Quantidade de indivíduos do grupo protegido, Mulheres, e não protegido, Homens, pertencentes a classe positiva e negativa da base *German Credit*. . . . 38
5.3 (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC, para diferentes valores de λ no conjunto de teste da base *Adult Income* após a classificação pelo SVM. 43
5.4 (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC para diferentes valores de λ no conjunto de teste da base *German Credit* considerando gênero como atributo sensível, após a classificação pelo SVM. 44
5.5 (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC para diferentes valores de λ para o conjunto de teste da base *German Credit* considerando idade como atributo sensível, após a classificação pelo SVM. 45
5.6 (a) t-SNE dos dados originais da base *Adult Income* considerando gênero como atributo sensível (b) t-SNE dos dados reconstruídos da base *Adult Income* considerando gênero como atributo sensível e $\lambda = 3 \times 10^2$. 46
5.7 (a) t-SNE dos dados originais da base *German Credit* considerando idade como atributo sensível (b) t-SNE dos dados reconstruídos da base *German Credit* considerando idade como atributo sensível e $\lambda = 50$. 47
5.8 (a) t-SNE dos dados originais da base *German Credit* considerando gênero como atributo sensível (b) t-SNE dos dados reconstruídos da base *German Credit* considerando gênero como atributo sensível e $\lambda = 50$. 48
5.9 (a) Comparativo de acurácia entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). (b) Comparativo de *statisitcal parity* entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (Learning Fair Representation). 49

5.10 (a) Comparativo de acurácia entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). (b) Comparativo de *statisitcal parity* entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). Considerando o gênero como atributo sensível. . . . 49

Lista de Tabelas

2.1 *Toy example* do funcionamento do método de *massaging*, onde os indivíduos com menor confiança na classificação podem ter a *label* trocada para gerar um conjunto mais equilibrado. 10

5.1 Configuração do *fair autoencoder* para cada base de dados, mostrando o intervalo de valores utilizado pelo *bayesian search* para se determinar a taxa de aprendizado (*learning rate*) que melhor reduzissem o erro entre a saída da rede e os dados originais. Bem como outros valores utilizados como tamanho dos *batches* e quantidade de neurônios em cada camada. 40

5.2 Configurações do SVM utilizado após o processo de extração de variáveis latentes pelo *fair autoencoder*, mostrando o *kernel* utilizado, e os intervalos para escolhidos para os hiper-parâmetros C e *gamma* que melhor otimizassem o MCC final. 41

5.3 Configuração dos intervalos de λ utilizados por cada base de dados e os respectivos atributos sensível considerados 42

5.4 *Statistical parity*, acurácia, F1 score e MCC para as base de dados: Adult Income e German Credit usando gênero; gênero e idade respectivamente, para diferentes abordagens 46

Capítulo 1

Introdução

1.1 Motivação

Nos últimos anos, temos testemunhado um aumento significativo na adoção de sistemas inteligentes, os quais empregam modelos preditivos que processam grandes volumes de dados de indivíduos. Esse cenário tem gerado preocupações sobre a forma como essas informações são manipuladas, bem como os potenciais impactos sociais e éticos resultantes de seu uso. Essas novas demandas estão principalmente relacionadas ao uso inadequado dessas informações, possíveis exposições indesejadas de dados sensíveis, a forma como esses dados são coletados e uma potencial amplificação de vieses indesejáveis.

A utilização desses modelos de inteligência artificial possibilita a prestação de serviços como recomendações, detecção de anomalias, predição de comportamentos, entre outros. Frequentemente, esses conjuntos de dados são formados a partir de decisões tomadas por indivíduos no passado, como um histórico de concessão de crédito em bancos ou histórico de salários de uma empresa. Com isso em mente, ao aplicar modelos de tomada de decisões que utilizam dados provenientes de decisões individuais, é importante considerar que esses modelos podem incorporar possíveis preconceitos e vieses inerentes a essas decisões tomadas no passado, a menos que sejam devidamente tratados.

Portanto, utilizar modelos preditivos de aprendizado supervisionado com dados potencialmente comprometidos em termos de preconceito, pode reforçar esse viés, resultando potencialmente em uma tomada de decisão que prejudique um determinado grupo de indivíduos. Para evitar essa situação, é necessário o uso de técnicas de aprendizado de máquina que considerem essa questão ao fazer uso de conjuntos de dados que contenham informações de pessoas reais.

Dentro do campo de *machine learning*, *fairness* é um problema relacionado a um tipo de viés associado a um grupo de pessoas que são desfavorecidas em um

conjunto de dados. Esse desfavorecimento está diretamente ligado a um atributo sensível, como gênero, raça, idade ou orientação sexual, refletindo os preconceitos presentes na sociedade. Esses atributos, muitas vezes, têm uma relação direta com a classificação do indivíduo. (FOULDS *et al.*, 2020)

Nesse contexto, existem diversas questões a serem abordadas na área, como a utilização adequada de dados para evitar que modelos de aprendizado reproduzam vieses indesejados e, ao mesmo tempo, extrair conhecimento deles. Além disso, é necessário medir a injustiça desses algoritmos de forma apropriada, considerando os *trade-offs* envolvidos e as questões éticas relacionadas ao uso de dados sensíveis de pessoas reais. Também é importante explorar possíveis relações entre a interpretabilidade dos dados, a compreensão das decisões dos algoritmos e a relevância dos atributos envolvidos.

Como os modelos de aprendizado frequentemente buscam identificar e reproduzir padrões presentes nos dados, se esses dados indicarem, por meio de informações sensíveis, um padrão que prejudique determinado grupo — mesmo que a pertença a esse grupo não deveria implicar diretamente nesse prejuízo —, enfrentamos um problema de *fairness*. Neste trabalho, abordamos essa questão e propomos uma técnica de aprendizado de máquina para atenuar esse tipo de viés, visando resultados mais equitativos e justos.

1.2 Objetivos

Este trabalho tem como objetivo propor um método genérico a ser aplicado a qualquer base de dados com características de injustiça, visando promover um uso mais justo dessas bases por outros modelos de aprendizado de máquina. O método proposto utiliza aprendizado de máquina de forma não supervisionada para mitigar o impacto do atributo sensível, que normalmente está associado ao viés de *fairness*. Para isso método proposto utiliza exclusivamente nos dados originais, exceto pelo rótulo de classificação.

Com o objetivo de avaliar a capacidade da proposta em remover parte da informação sensível de conjuntos de dados injustos, este trabalho foi dividido da seguinte forma:

(i) Propor um método de pré-processamento para bases de dados com viés de injustiça. O método é capaz de extrair variáveis latentes o conjunto de dados original mantendo suas características, removendo parte da informação sensível associada à injustiça.

(ii) Propor um *autoencoder*, com uma função de custo personalizada, capaz de extrair variáveis latentes desassociadas dos atributos sensíveis, gerando uma

representação mais justa dos dados.

(iii) Realizar experimentos utilizando conjuntos de dados reais com viés de injustiça, com objetivo de avaliar a capacidade do método proposto em remover parte do da informação sensível desses conjuntos.

1.3 Resumo dos Resultados

Foram selecionadas duas bases de dados amplamente utilizadas pela literatura de *fairness*, com diferentes características para validar a proposta. Os resultados indicam um sucesso em reduzir a injustiça, relacionado ao atributo sensível em questão, em detrimento de uma perda nas métricas de desempenho como Acurácia, F1 *score* e *Matthews Correlation Coefficient*. Também foram realizados experimentos comparativos para determinar a capacidade do método proposto em remover parte informação sensível.

1.4 Contribuições

Esta dissertação foi desenvolvida no contexto de *fairness*, mais especificamente na área de *fair representation*, com as seguintes contribuições: (i) Um método de pré–processamento para dados injustos; (ii) Um *autoencoder* que faz uso de uma função de custo com uma penalidade para reconstruir os dados removendo parte da informação sensível; (iii) Avaliação da capacidade da proposta por meio de experimentos com bases de dados reais; (iv) Avaliar a capacidade da proposta em remover parte da informação sensível.

1.5 Organização do Trabalho

Esta dissertação está organizada da seguinte forma: no Capítulo 2, apresentamos definições, causas e algumas abordagens de soluções sobre *fairness*. No Capítulo 3, é apresentado a fundamentação teórica de aprendizado de máquina para um melhor entendimento do presente trabalho. No Capítulo 4, é detalhado o método *Covariance Fair Autoencoder*, que busca aprender uma representação mais justa dos dados. Os experimentos realizados, bem como as discussões sobre os resultados, foram detalhados no Capítulo 5. Por fim, no Capítulo 6, a conclusão do trabalho foi desenvolvida, juntamente com considerações sobre possíveis trabalhos futuros.

Capítulo 2

Fairness

Este capítulo tem como objetivo apresentar alguns conceitos de *Fairness*, bem como as suas possíveis causas e métricas. Por último, abordaremos as principais abordagens e soluções utilizadas na literatura.

2.1 Contexto

Com o avanço significativo da tecnologia e a proliferação de dispositivos digitais, resultando em uma produção cada vez maior de dados pessoais, as pessoas estão constantemente compartilhando informações em redes sociais, realizando compras online, usando aplicativos de transporte, entre outras atividades, o que gera enormes quantidades de dados. Esses dados são valiosos para empresas e governos, permitindo a criação de produtos personalizados, recomendações precisas e aprimoramento de serviços.

Pela capacidade de análise e extração de conhecimento de algoritmos de aprendizado de máquina, esses dados vêm sendo utilizados para alimentar sistemas de tomada de decisão que fazem uso de *machine learning*. Essas aplicações envolvem empréstimos bancários, campanhas de *marketing*, detecção de características para sistema de reconhecimento de identidade e muitos outros. No entanto, essa crescente coleta e uso de dados pessoais também levantam questões éticas e sociais. Um dos problemas mais preocupantes é a questão de *fairness* ou equidade. (LE QUY *et al.*, 2022)

O problema emerge quando algoritmos e sistemas automatizados de tomada de decisão usam dados pessoais para aprendizado e, inadvertidamente, perpetuam ou amplificam preconceitos e discriminações existentes na sociedade. Uma vez que os dados coletados podem conter vieses implícitos ou refletir desigualdades históricas, a utilização direta ou inadequada dessa informação em algoritmos pode levar a decisões discriminatórias. Por exemplo, se um banco se baseia no histórico de decisões anteriores de concessão de crédito para tomar decisões atuais e esse histórico inclui

informações demográficas dos clientes, como gênero, raça, idade e orientação sexual, e se no passado o banco tinha uma tendência a negar crédito a pessoas de um determinado gênero, decisões baseadas nesses dados podem classificar indivíduos desse grupo como mais arriscados com base apenas no gênero (LE QUY *et al.*, 2022).

Além disso, certos grupos minorizados e vulneráveis podem ser sub-representados nos dados de treinamento, o que pode levar a modelos que não são adequadamente ajustados para suas características e necessidades. Esta sub-representação pode resultar em algoritmos que são menos precisos e injustos para essas populações, perpetuando assim as desigualdades existentes. Por exemplo, sistemas de reconhecimento facial, que são cada vez mais utilizados em diversas aplicações, podem ter um desempenho inferior em relação a pessoas de pele mais escura ou mulheres. Se esses sistemas forem predominantemente treinados com um conjunto de dados que tem uma maioria de homens brancos, eles podem não ser capazes de identificar corretamente indivíduos de outros grupos demográficos. Isso não apenas compromete a eficácia do sistema, mas também pode levar a consequências discriminatórias e injustas para aqueles que já são frequentemente marginalizados. (YUCER *et al.*, 2020)

O viés relacionado ao *fairness* tem características próprias. Enquanto os outros tipos de vieses podem estar relacionados à imprecisão ou distorção dos dados em geral, o viés de *fairness* concentra-se especificamente em como os atributos demográficos, como raça, gênero, idade, entre outros, são tratados na tomada de decisões. Por exemplo, um viés de amostragem pode ocorrer quando os dados de treinamento não representam adequadamente a diversidade da população, mas isso pode não afetar diretamente as decisões relacionadas a atributos sensíveis. Já o viés de *fairness* se refere à possibilidade de que os algoritmos, ao considerarem os atributos sensíveis, perpetuem discriminações e desigualdades injustas. (MEHRABI *et al.*, 2021)

Neste contexto, *fairness* tem ganhado considerável destaque nos trabalhos da literatura. A preocupação com a equidade tem impulsionado pesquisas e discussões sobre como criar modelos mais inclusivos, que considerem as diversas realidades e perspectivas dos indivíduos. A busca por soluções que promovam a equidade, tem se tornado uma prioridade na área de ciência de dados e inteligência artificial. O objetivo é construir um futuro em que a tecnologia seja utilizada para impulsionar a igualdade e não perpetuar ou amplificar desigualdades existentes na sociedade. (ZAFAR *et al.*, 2015)

Ao longo desses estudos, surgiram diferentes definições relevantes para esse tema. E embora não seja possível ter uma definição única de *fairness*, por conta dos diferentes cenários sociais e das possíveis aplicações, é importante destacar algumas definições relevantes (PESSACH e SHMUELI, 2020; VERMA e RUBIN, 2018). Como os conceitos de *disparate treatment* e *disparate impact*, onde o primeiro refere-se à

prática de utilizar diretamente informações sensíveis em sistemas de tomada de decisão. Já o segundo ocorre quando o atributo sensível afeta a decisão final de um sistema de tomada de decisão, mesmo que o conjunto de dados utilizado no treinamento não tenha a informação sensível diretamente, mas de alguma forma tenha acesso a essa informação de forma indireta (ZAFAR *et al.*, 2015).

Outro conceito fundamental no contexto de *fairness* é o de *equal opportunity* (Oportunidade Igual). Esse conceito sugere que, em um sistema de tomada de decisão, a probabilidade de indivíduos, que verdadeiramente pertençam a classe positive, de diferentes grupos demográficos serem classificados nas categorias ou *targets* positivos deve ser a mesma. Em outras palavras, não deve existir discriminação baseada em atributos sensíveis, como gênero, raça, idade ou origem étnica. Já o princípio de *Statistical Parity* remete à ideia de que a classificação final de indivíduos deve ser independente do grupo ao qual uma pessoa pertence. (MEHRABI *et al.*, 2021)

Além disso, outro conceito relevante é o de *fairness through awareness*, que envolve a ideia de que indivíduos com características similares devem ser classificados de forma similar. Esse conceito busca garantir que pessoas que compartilham algumas características tenham tratamento similares, independentemente de quaisquer outros aspectos não relevantes para a decisão em questão. Por exemplo, se dois indivíduos têm perfis semelhantes em relação a variáveis significativas para a decisão, como histórico de crédito ou qualificações profissionais, eles devem receber tratamento similar no processo de aprovação de crédito ou contratação de emprego. No entanto, é importante destacar que essa igualdade de tratamento deve ocorrer dentro dos limites legais e éticos, de forma a não violar direitos individuais ou perpetuar estereótipos injustos. (MEHRABI *et al.*, 2021)

2.2 Causas

Um sistema de aprendizado de máquina pode enfrentar desafios de *fairness* decorrentes de várias fontes. MEHRABI *et al.* (2021) apresentam algumas causas para esse tipo de viés. Uma das principais causas surgir devido a um viés histórico presente nos dados de treinamento. Isso se manifesta quando os dados utilizados para treinar o modelo refletem desigualdades e preconceitos presentes na sociedade. Se esses dados apresentarem viés em relação a atributos sensíveis, o algoritmo de aprendizado de máquina pode assimilar e reproduzir esses padrões discriminatórios nos resultados.

Outra causa significativa para a ocorrência de viés de dados pode estar relacionada a um viés de população, que emerge quando as características dos indivíduos pertencentes à população-alvo diferem das características presentes no conjunto de dados coletado ou utilizado para análise. Esse tipo de viés pode surgir de várias

fontes, como diferenças socioeconômicas, localização geográfica, preferências culturais ou comportamentais, e até mesmo pela presença de grupos sub-representados ou pouco representados na amostra de dados. Consequentemente, essa falta de correspondência entre as características da população-alvo e as do conjunto de dados pode gerar resultados imprecisos, dificultando a generalização das conclusões. (MEHRABI *et al.*, 2021)

O viés de conteúdo é uma questão importante que surge quando há discrepâncias na produção ou coleta de dados, resultando em diferentes grupos produzindo conjuntos de dados distintos, muitas vezes influenciados por suas classes sociais, culturas ou hábitos específicos. Esse tipo de viés pode ocorrer em várias situações, como em plataformas de mídia social, onde determinados grupos podem ter maior visibilidade ou participação em relação a outros, levando a uma representação desigual dos diferentes segmentos da população. Além disso, o viés de conteúdo pode se manifestar em pesquisas, estudos de mercado ou até mesmo na coleta de dados governamentais, onde as metodologias empregadas podem favorecer certos grupos ou aspectos, resultando em uma falta de representatividade completa da diversidade da sociedade. (MEHRABI *et al.*, 2021)

Essa disparidade nos dados pode levar a interpretações enviesadas e conclusões imprecisas, afetando negativamente tomadas de decisões que fazem uso desses dados. Para abordar o viés de conteúdo, é crucial adotar estratégias de coleta de dados mais inclusivas e representativas, bem como aplicar métodos de correção ou ajuste para garantir que os resultados sejam mais abrangentes e fiéis à realidade. A consciência e o combate ao viés de conteúdo são fundamentais para assegurar que as informações e *insights* derivados dos dados reflitam uma visão mais completa e equitativa da sociedade, impulsionando a construção de soluções mais justas e igualitárias para os desafios que enfrentamos. (MEHRABI *et al.*, 2021)

É de extrema importância que essas causas sejam cuidadosamente abordadas em sistemas de aprendizado de máquina, a fim de evitar problemas relacionados *fairness* e assegurar resultados mais equitativos. Isso requer uma seleção criteriosa dos dados de treinamento, com atenção especial para a representatividade da amostra em relação à população-alvo, evitando assim viés de população e de conteúdo. Além disso, é imprescindível o uso de algoritmos e métricas apropriadas que levem em conta questões de justiça, como a equidade em termos de grupos protegidos. Ao adotar tais práticas, podemos aumentar a confiança nas decisões automatizadas e garantir que elas sejam guiadas por princípios éticos e sociais, em benefício de toda a sociedade.

2.3 Métricas

Ter uma forma confiável e precisa de medir ou avaliar o nível de *fairness* de uma solução ou abordagem é de extrema importância para a área de inteligência artificial. Isso permite que os pesquisadores, desenvolvedores e tomadores de decisão compreendam melhor como as estratégias estão funcionando em termos de justiça e equidade. Ao ter métricas e métodos de avaliação bem definidos, torna-se possível comparar diferentes abordagens, identificar possíveis vieses ou desigualdades e, mais importante, avaliar o sucesso ou fracasso em lidar com o problema.

A existência de métricas de *fairness* também é crucial para orientar o desenvolvimento de sistemas mais éticos e responsáveis, pois possibilita a identificação de áreas de melhoria e aperfeiçoamento das soluções existentes. Além disso, as métricas de *fairness* fornecem uma base sólida para a tomada de decisões informadas sobre o uso de um modelo ou algoritmo em contextos reais, como em áreas sensíveis já citadas, como saúde, justiça criminal ou empréstimos financeiros.

Ao longo dos anos, diversos pesquisadores propuseram uma variedade de métricas para avaliar e garantir a justiça nos modelos de aprendizado de máquina. Cada métrica tem sua própria abordagem e foco, refletindo as nuances e especificidades dos problemas que buscam abordar. E, embora não exista uma única métrica que consiga contemplar a complexidade dos diferentes problemas relacionados à *fairness*, algumas delas se destacam por sua relevância e frequência de uso na literatura. Essas métricas tornaram-se referências no campo, servindo como padrão para muitos estudos e aplicações práticas. Além disso, a escolha da métrica adequada é crucial, pois pode influenciar significativamente os resultados e a interpretação dos modelos, bem como as decisões tomadas com base neles.

Statistical parity, também conhecida como *demographic parity*, refere-se a uma medida que avalia se os grupos de indivíduos, protegidos e não protegidos, têm chances iguais de pertencer à classe positiva. Por exemplo, em um conjunto de dados onde o atributo sensível é o gênero, homens e mulheres devem ter chances iguais de serem classificados positivamente. A Equação 2.1 mostra como calcular o *statistical parity* por meio da diferença entre duas probabilidades: a primeira é a probabilidade de um indivíduo pertencer à classe positiva, $\hat{y} = 1$, dado que ele pertence ao grupo não protegido, $s \in S^+$; a segunda é a probabilidade de um indivíduo pertencer à classe positiva, $\hat{y} = 1$, dado que ele está no grupo protegido, $s \in S^-$. A ideia é que, quanto mais próximas estiverem essas duas probabilidades, mais independente do atributo sensível será a probabilidade de um indivíduo pertencer à classificação positiva. (VERMA e RUBIN, 2018)

$$\text{statistical parity} = P(\hat{y} = 1 | s \in S^+) - P(\hat{y} = 1 | s \in S^-) \tag{2.1}$$

O *Equalized odds* é uma métrica que estipula que a diferença entre a probabilidade de um indivíduo, que verdadeiramente pertença classe positiva, ser corretamente classificado e a probabilidade de um indivíduo, que verdadeiramente pertença à classe negativa, ser erroneamente classificado deve ser o mais próxima possível de zero. Isso visa garantir resultados mais justos entre os diferentes grupos. A Equação 2.2 exemplifica essa definição. Nela, y representa a classe verdadeira, $\hat{y}$ a classificação predita, S^+ o grupo privilegiado e S^- o grupo protegido. (VERMA e RUBIN, 2018)

$$\text{equalized odds} = P(\hat{y} = 1|y = i, s \in S^+) - P(\hat{y} = 1|y = i, s \in S^-), i \in 0, 1 \quad (2.2)$$

2.4 Abordagens e Soluções

Com o aumento do interesse no tema de *fairness*, naturalmente surgem abordagens e propostas na literatura com o objetivo de resolver ou mitigar o problema. Na seção 4.3, alguns desses trabalhos foram detalhados. Porém, nesta seção, será discutido de forma mais abrangente as diferentes categorias das soluções de forma mais geral.

Existem três tipos diferentes de abordagens para lidar com o problema de *fairness*: pré-processamento, em processamento e pós-processamento. No pré-processamento, busca-se transformar os dados que contêm o viés de *fairness* em um novo conjunto de dados, de forma que esse novo conjunto esteja livre do viés presente nos dados originais. Esse processo envolve técnicas de modificação e tratamento dos dados para reduzir ou eliminar o impacto de atributos sensíveis na tomada de decisão. (MEHRABI *et al.*, 2021)

Já no Em Processamento, procura-se incluir no próprio processo de aprendizado dos modelos regras ou restrições que garantam que a decisão final seja mais equitativa. Isso pode ser feito através da modificação de funções de custo ou por inclusão de restrições no aprendizado que penalizem a perpetuação preconceito, e dessa forma consigam generalizar e aprender parâmetros que não reflitam esse problema. (MEHRABI *et al.*, 2021)

Por último, técnicas de pós-processamento ocorre após o treinamento do modelo. Nessa etapa, são aplicadas regras que modifiquem as classificações do conjunto de teste que tenham um grau de confiança menor feito pelo modelo de treino, garantindo a equidade nas decisões finais. Ao identificar as previsões do modelo de treino que possuem menor confiança, é possível realizar intervenções específicas para tentar remover ou amenizar possíveis tendências discriminatórias. (MEHRABI *et al.*, 2021)

Uma solução que surge naturalmente é a simples remoção do atributo sensí-

vel para evitar que este seja utilizado por modelos de aprendizagem. No entanto, diversos estudos, como os apresentados por KAMISHIMA *et al.* (2012, 2011), demonstraram que o viés inerente à informação sensível pode permanece nos atributos restantes. Isso ocorre porque os atributos muitas vezes não são independentes entre si, exigindo mais do que a simples exclusão dessa informação do conjunto de dados. Esse fenômeno é conhecido como *redlining effect*, como a proposta deste trabalho busca remover tal característica da base de dados com um todo alguns experimentos e comparativos foram realizados nesse sentido, mais detalhes são apresentados no Capítulo 5.

Um dos primeiros trabalhos a propor uma solução baseada na abordagem de pré-processamento foi apresentado por PEDRESHI *et al.* (2008). A abordagem proposta ficou conhecida como *massaging*. Esse tipo de abordagem procura utilizar um modelo probabilístico de classificação e, através desse primeiro resultado, faz uso dessa probabilidade para trocar as *labels* de classificação dos dados originais, especificamente de indivíduos com menor grau de confiança. Essa troca é realizada até que se atinja um equilíbrio de classificação positiva entre grupos demográficos, normalmente utilizando *equalized odds* ou *statistical parity* para isso. O resultado produzir um novo conjunto de dados mais justo.

Na Tabela 2.1, é apresentado um exemplo de como as técnicas de *massaging* funcionam, onde o atributo sensível é o gênero, com o grupo protegido sendo as mulheres. Após a aplicação de um modelo de aprendizado, os indivíduos com menor confiança na classificação terão suas *labels* de classificação trocadas. Nesse caso, as entradas da segunda e da quarta linha, porque embora no primeiro caso a classificação positiva tenha ficado abaixo dos 50%, esse foi o indivíduo do grupo protegido com menos confiança na classificação. Já no segundo caso, a classificação foi positiva, mas com menor confiança dentro dessa amostra. A troca dessas classificações proporciona um equilíbrio entre ambos os grupos, mulheres e homens, em relação à pertencerem à classe positiva.

Tabela 2.1: *Toy example* do funcionamento do método de *massaging*, onde os indivíduos com menor confiança na classificação podem ter a *label* trocada para gerar um conjunto mais equilibrado.

Gênero	**Profissão**	**Renda**	**Escolaridade**	**Classe Possitiva (%)**
Homem	Designer	\$4.000	Superior	85
Mulher	Professora	\$3.800	Superior	**45**
Mulher	Vendedora	\$3.000	Médio	30
Homem	Vendedor	\$2.500	Médio	**55**
Homem	Veterinário	\$5.000	Superior	9
Mulher	Programadora	\$7.000	Superior	86

2.5 Fair Representation

Representation learning é uma subárea emergente do aprendizado de máquina que se dedica a identificar e extrair representações latentes dos dados, transformando-os em formatos mais compactos e informativos. Em vez de depender de características pré-definidas, o *representation learning* busca identificar automaticamente padrões e características intrínsecas que são cruciais para tarefas específicas, como classificação, regressão ou agrupamento. (LOUIZOS *et al.*, 2016)

Por exemplo, em imagens, uma representação aprendida pode identificar bordas, texturas e padrões que são cruciais para distinguir entre diferentes objetos. Em processamento de linguagem natural, pode identificar temas ou tópicos subjacentes em documentos.

Uma das principais vantagens do *representation learning* é a capacidade de aprender representações compactas e informativas dos dados. Estas representações podem proporcionar uma melhor compreensão dos dados, facilitando tarefas subsequentes, como classificação, agrupamento ou até mesmo geração de novos dados (KENFACK *et al.*, 2021). Em um conjunto de dados médicos por exemplo, uma representação aprendida pode identificar padrões subjacentes relacionados a doenças específicas, tornando mais fácil para os médicos diagnosticar e tratar pacientes no futuro.

Além disso, uma vez que estas representações são aprendidas, elas podem ser transferidas para diferentes tarefas ou conjuntos de dados, resultando em benefícios de generalização e eficiência em problemas relacionados. Isso é particularmente útil em cenários onde os dados são escassos.

Em certos cenários, o objetivo do *representation learning* pode envolver a remoção de características indesejáveis presentes nos dados originais, que são comumente tratadas como ruído a ser eliminado. Por outro lado, as características úteis que se pretende preservar devem permanecer intactas, como destacado por LOUIZOS *et al.* (2016).

Fair representation, por sua vez, é um tipo de *representation learning*. O seu objetivo é encontrar uma representação dos dados que não apenas capture as características essenciais, mas que também esteja livre de informações sensíveis (KENFACK *et al.*, 2021). Em um mundo onde a discriminação e o viés são preocupações crescentes, a busca por representações justas é crucial. Por exemplo, em um sistema de recomendação de empregos, uma representação justa garantiria que candidatos sejam avaliados com base em suas habilidades e experiências, e não em características sensíveis como gênero ou etnia.

Como o uso de informações sensíveis pode prejudicar pessoas de grupo protegidos, levando a decisões discriminatórias ou injustas. Por esse motivo, uma técnica de *fair representation* deve ser capaz de preservar informações essenciais de forma

eficiente dos dados, e ao mesmo tempo procurar remover ou dificultar a indicação que possa determinar o pertencimento dos indivíduos a um grupo protegido.

Contudo, existem diversos desafios para encontrar uma representação mais justa dos dados. Uma vez que os dados originais contêm um viés de injustiça, uma representação latente dessa informação naturalmente herdará características similares (KENFACK *et al.*, 2021). Portanto, um dos maiores desafios é justamente encontrar uma forma de extrair conhecimento útil desses dados, tentando remover a informação sensível.

Outro desafio gira em torno da interpretabilidade dos dados. Uma representação dos dados originais, embora carregue informações relevantes, pode perder a estrutura e o formato original, tornando mais difícil a interpretação e análise dos dados em cenários onde isso é necessário.

Apesar dos desafios apresentados, o *fair representation* emerge como uma abordagem promissora e válida para lidar com questões de *fairness*. O *fair representation* não apenas busca neutralizar os vieses presentes nos dados, mas também se esforça para garantir que as representações derivadas sejam justas e não discriminatórias. Isso é particularmente crucial em aplicações de aprendizado de máquina que têm impactos diretos na vida das pessoas.

Capítulo 3

Aprendizado de Máquina

Este capítulo tem como objetivo apresentar conceitos teóricos fundamentais para o entendimento da proposta desta dissertação. As seções subsequentes estão divididas da seguinte forma: Aprendizado de Máquina, onde apresentamos diferentes abordagens e tipos de modelos, com foco especial em redes neurais artificiais, detalhando seu funcionamento e tipos de arquitetura. Em seguida, apresentamos o funcionamento básico do modelo SVM, que foi o modelo escolhido para validar a proposta. Por último, exploramos alguns aspectos do aprendizado com viés.

3.1 Contexto

Aprendizado de Máquina (*Machine Learning* - ML) tem emergido como uma disciplina essencial na era da informação, promovendo avanços significativos em diversas áreas, como medicina, finanças, automação industrial, entre outras. A capacidade de extrair informações úteis e padrões complexos de dados, combinada com a flexibilidade de adaptação a cenários diversos, torna o ML uma ferramenta poderosa para a tomada de decisões e solução de problemas complexos.

Em busca de solucionar os desafios e questões do mundo real por meio de modelos de inteligência artificial, duas tarefas fundamentais têm se mostrado de extrema importância e amplamente exploradas pelos algoritmos de aprendizagem de máquina: a classificação e a regressão. A classificação visa atribuir uma classe ou categoria específica a uma instância de dados com base em suas características distintivas. Essa tarefa encontra aplicação em uma variedade de cenários, como detecção de fraudes em transações financeiras, diagnóstico médico de doenças a partir de exames ou classificação de imagens em tarefas de visão computacional.

Por outro lado, a regressão se destina a prever um valor contínuo ou quantitativo com base nas características conhecidas de uma instância. Essa tarefa é útil em diversas áreas, desde previsão de preços de ativos financeiros, como ações e commodities, até estimativa de demanda em vendas, auxiliando em tomadas de decisões

estratégicas.

Ambas as tarefas são cruciais para o desenvolvimento de sistemas de inteligência artificial que possam auxiliar em tomadas de decisão ou fornecer *inputs* valiosos em inúmeros domínios, seja na área médica, financeira, de marketing ou em tarefas complexas de reconhecimento de padrões e análise de dados. A classificação e a regressão desempenham um papel central na evolução e na aplicação bem-sucedida do aprendizado de máquina em problemas reais. No entanto, para garantir a eficácia e a confiabilidade desses modelos, é essencial enfrentar desafios como a seleção apropriada de algoritmos, a qualidade e adequação dos dados de treinamento, e a interpretabilidade dos resultados gerados, visando sempre aprimorar a tomada de decisão e proporcionar soluções cada vez mais precisas e responsáveis.

3.2 Aprendizado supervisionado

O aprendizado supervisionado é um dos principais paradigmas de aprendizado de máquina. Nesse tipo de abordagem, um algoritmo é treinado usando um conjunto de dados rotulados, ou seja, exemplos em que as respostas corretas ou rótulos são fornecidos junto com as características. O objetivo é extrair padrões e relações entre as características e os rótulos conhecidos, para que o modelo possa fazer previsões precisas para novos exemplos não rotulados.

Durante o treinamento, os algoritmos ajustam seus parâmetros para minimizar o erro entre as previsões feitas pelo modelo e os rótulos verdadeiros do conjunto de treinamento. Uma vez que o modelo é treinado, ele pode ser usado para fazer previsões em dados desconhecidos, onde os rótulos são desconhecidos. Espera-se que, com o treinamento, os parâmetros aprendidos sejam capazes de generalizar a relação entre as características e o *target* com uma confiança razoável. (CUNNINGHAM *et al.*, 2008)

A principal vantagem do aprendizado supervisionado é sua capacidade de generalizar para novos dados, permitindo que o modelo faça previsões precisas mesmo em exemplos não vistos durante o treinamento. No entanto, a qualidade das previsões depende em grande parte da qualidade e representatividade do conjunto de treinamento. Além disso, o aprendizado supervisionado requer um esforço considerável na rotulagem dos dados, o que pode ser caro e demorado em algumas aplicações, na Figura 3.1 é apresentado uma representação desse tipo de aprendizado, onde o conjunto de dados representados por X_1 e X_2 são conhecido e a seta em azul representa o modelo aprendendo a relação dos dados. (GHAHRAMANI, 2003)

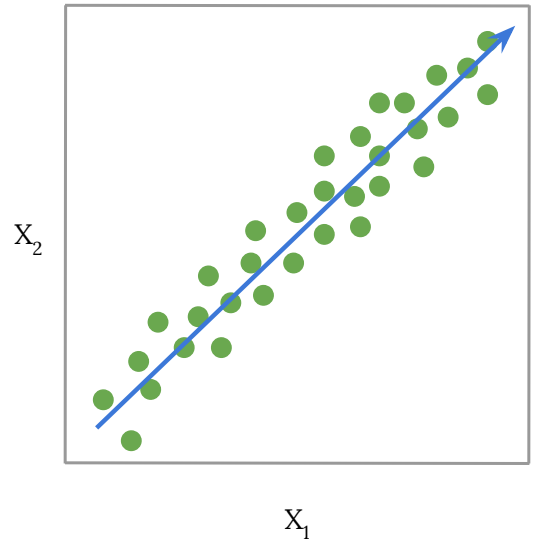


Figura 3.1: Representação de como o aprendizado supervisionado funciona, onde o modelo tenta generalizar a partir dos dados de treino.

3.3 Aprendizado Não Supervisionado

O Aprendizado não supervisionado também é uma das principais técnicas de ML. Nessa abordagem, ao contrário do aprendizado supervisionado, os modelos não têm acesso às classificações ou rótulos das instâncias no conjunto de treinamento. Em vez disso, o algoritmo é desafiado a encontrar padrões e estruturas ocultas diretamente nas características dos atributos, sem a orientação de rótulos previamente conhecidos.

A principal tarefa do aprendizado não supervisionado é a clusterização, onde o algoritmo agrupa instâncias similares em *clusters*, com base na proximidade entre suas características. O objetivo é descobrir agrupamentos naturais ou estruturas intrínsecas nos dados, que podem ser úteis para entender as relações entre as instâncias ou identificar grupos distintos. (GHAHRAMANI, 2003)

Uma das principais vantagens do aprendizado não supervisionado é sua capacidade de explorar e encontrar padrões em grandes volumes de dados não rotulados, o que pode levar a *insights* valiosos e descobertas inesperadas. Isso é especialmente útil em casos em que é difícil ou impraticável obter rótulos para as instâncias ou quando se deseja explorar a estrutura dos dados antes de realizar outras tarefas, como a classificação em aprendizado supervisionado, na Figura 3.2 é apresentado uma representação desse tipo de aprendizado, onde os modelos procura agrupar os dados em grupo distintos. (GHAHRAMANI, 2003)

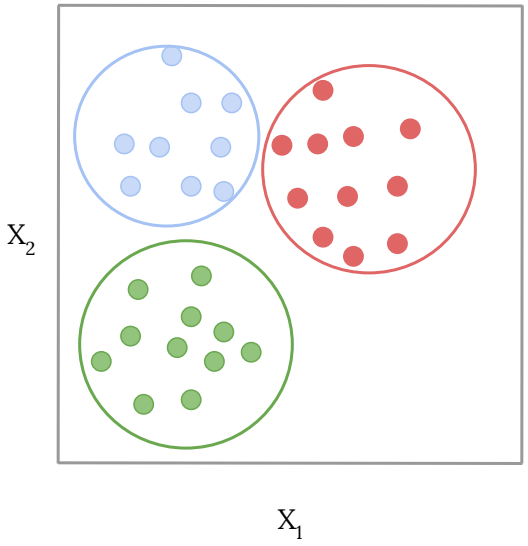


Figura 3.2: Representação de como o aprendizado não supervisionado funciona, onde o modelo tenta agrupar os dados de treino sem conhecer o rótulo.

3.4 Aprendizado por Reforço

O aprendizado por reforço é um paradigma de aprendizado de máquina que se baseia em um modelo de interação agente-ambiente. Nesse contexto, um agente é responsável por tomar decisões em um ambiente para alcançar um objetivo específico. O ambiente é um sistema que o agente interage e no qual ele pode executar ações. A cada ação realizada pelo agente, o ambiente responde com um *feedback* em forma de recompensa ou penalidade, indicando o quão boa foi a ação tomada pelo agente para alcançar seu objetivo. (LI, 2017)

Diferentemente do aprendizado supervisionado, onde os dados de treinamento contêm exemplos rotulados com as respostas corretas, o aprendizado por reforço requer uma exploração ativa do agente no ambiente para aprender a melhor estratégia. Essa exploração pode levar o agente a realizar ações que resultem em recompensas negativas temporárias, mas que sejam necessárias para descobrir uma política mais eficaz a longo prazo. (GHAHRAMANI, 2003)

Essa abordagem de aprendizado é amplamente utilizada em aplicações onde o ambiente é dinâmico e incerto, como em jogos, robótica, sistemas de controle e até mesmo na otimização de recursos. Além disso, o aprendizado por reforço tem sido aplicado com sucesso em cenários complexos, como treinamento de agentes virtuais para jogar xadrez ou jogos de estratégia, mostrando seu potencial para resolver problemas desafiadores de tomada de decisão. (LI, 2017)

Apesar de suas vantagens, o aprendizado por reforço também apresenta desafios,

como a necessidade de lidar com o problema de exploração versus explotação, ou seja, a decisão de se deve realizar ações conhecidas por gerar recompensas ou explorar novas ações para descobrir estratégias melhores. Além disso, o aprendizado por reforço requer um grande número de interações com o ambiente, o que pode tornar o processo de treinamento mais demorado e computacionalmente custoso. (GHAHRAMANI, 2003)

3.5 Modelos de Aprendizado

Alguns exemplos de modelos de aprendizado de máquina incluem árvores de decisão, aprendizado *bayesiano*, redes neurais, entre outros. As árvores de decisão são modelos que buscam estabelecer limites nos domínios dos atributos utilizados no treinamento para determinar os caminhos que melhor representam as classificações do conjunto de treino. Esses modelos empregam técnicas de particionamento recursivo para dividir o espaço de atributos em regiões distintas, de forma que cada região corresponda a uma decisão ou classe específica. Por meio de processos de divisão baseados em regras lógicas, as árvores de decisão conseguem fazer predições precisas em problemas de classificação.

Outro modelo bastante explorado é o aprendizado *bayesiano*, que se baseia em uma abordagem probabilística para estimar as probabilidades de ocorrência de diferentes classes com base nas informações disponíveis. Esse método utiliza o teorema de Bayes para atualizar as probabilidades à medida que novas evidências são apresentadas, permitindo que o modelo aprenda a partir dos dados e ajuste suas estimativas de maneira iterativa. Com isso, o aprendizado *bayesiano* é especialmente útil em problemas com incertezas e quando se possui informações a priori sobre as classes.

Na Equação 3.1, é apresentada uma forma de se determinar a probabilidade de se pertencer a C_k dado um conjunto de características x, aplicando o teorema de Bayes. Essa probabilidade é determinada pela probabilidade de se encontrar classe C_k no conjunto de treino ($p(C_k)$), prior, combinado com a probabilidade de se observar as características x dado a classe C_k ($p(x|C_k)$), sobre a probabilidade de se observar o conjunto x independente da classe ($p(x)$).

$$p(C_k|x) = \frac{p(C_k)p(x|C_k}{p(x)} \tag{3.1}$$

3.5.1 Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs) são modelos de aprendizado de máquina compostos por unidades simples de processamento interconectadas, chamadas neurônios. Cada conexão possui um peso que representa a força daquela conexão, permitindo que a

rede processe uma entrada pelos neurônios e produza uma saída. Esses neurônios utilizam os pesos das conexões combinados com os dados em uma função matemática, geralmente não linear, para produzir uma saída adequada.

Esse modelo foi inspirados no funcionamento do cérebro biológico, que é uma rede interconectada de neurônios que se comunicam por meio de sinapses. Um determinado padrão de ativação de sinapses é capaz de reconhecer um padrão e gerar uma saída. A partir disso, as RNAs são redes de neurônios artificiais conectados com o objetivo de aprender padrões a partir de exemplos, e assim, são capazes de generalizar para novas entradas.

A principal característica que motivou o estudo das RNAs foi a possibilidade de replicar a capacidade de aprendizado e generalização do cérebro biológico. Esse aprendizado está intimamente relacionado às conexões entre os neurônios, pois, uma vez que um padrão é bem aprendido, uma rede neural é capaz de reconhecer esse padrão em novos exemplos, com um certo grau de confiança dependendo da quantidade e qualidade dos exemplos, tornando as RNAs um método muito robusto. (KRIESEL, 2013)

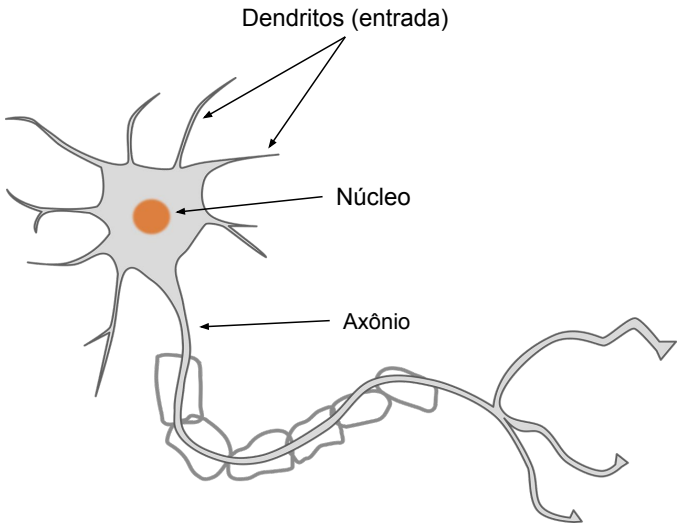


Figura 3.3: Representação do neurônio biológico. Onde os dendritos funcionam como entrada de sinais de outros neurônios, o núcleo que processa e produz uma saída pelo axônio. [2]

O neurônio biológico opera transmitindo sinais provenientes dos dendritos, que recebem estímulos de neurônios vizinhos. Esses sinais são processados no núcleo do neurônio e, por meio do axônio, um sinal de saída é gerado, possibilitando a conectividade com outros neurônios, como ilustrado na Figura 3.3. Quando um neurônio estimula outro, ocorre a formação de uma sinapse, ativando o neurônio

[2]Figura feita a partir da original acessada em https://pixabay.com/vectors/axon-brain-cell-dendrites-nerve-1294021/ no dia 04/06/2023, distribuída sob a licença Content License

seguinte e dando continuidade ao processo. O cérebro humano possui aproximadamente 10^{11} neurônios, o que possibilita inúmeras sinapses e conexões capazes de processar diferentes tipos de estímulos. (KRIESEL, 2013)

3.5.1.1 *Perceptron*

O perceptron é a representação mais simples de um neurônio artificial. Assim como seu análogo biológico, o perceptron é composto por um conjunto de n entradas, onde cada entrada é ponderada por um peso correspondente w, e possui um terminal de saída y. O núcleo do perceptron executa o cálculo da soma das entradas x_1, x_2, ..., x_n, ponderadas pelos respectivos pesos w_1, w_2, ..., w_n. O resultado dessa operação é a saída do perceptron. Ao longo do tempo, surgiram variações do perceptron, como o *sigmoid perceptron*, *tanh perceptron* e outros, os quais incorporam funções não lineares ao processamento. A representação gráfica de um perceptron pode ser visualizada na Figura 3.4. (NIELSEN, 2015)

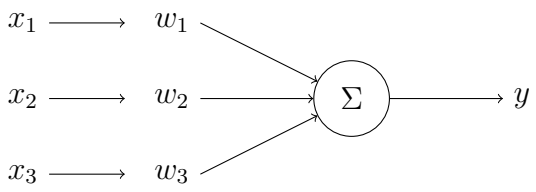


Figura 3.4: Representação gráfica do perceptron, onde um neurônio artificial combina um conjunto de entradas x_1, x_2, x_3 com os parâmetros w_1, w_2, w_3 para produzir uma saída.

Um único neurônio é capaz de executar uma computação muito simples. Portanto, a ideia por trás das redes neurais é combinar vários neurônios em uma rede interconectada, em que cada camada aplica transformações com base na camada anterior. A Figura 3.5 apresenta o exemplo mais simples de rede neural com uma camada oculta, com um conjunto de entrada, uma camada oculta e uma saída. Porém redes neurais mais complexas podem ser configuradas, com várias camadas ocultas e diversas saídas. (NIELSEN, 2015)

A quantidade de camadas internas e de neurônios em cada camada alinhado com o problema em questão é crucial para que a rede neural seja capaz de aprender padrões. Não há uma regra fixa a ser seguida; normalmente, são realizados testes experimentais com diferentes arquiteturas até que o resultado desejado seja alcançado.

Antes que uma rede neural consiga generalizar seu aprendizado para novos exemplos ela precisa passar por um treinamento, se esse treinamento for feito de forma supervisionada serão fornecidos diversos exemplos de entradas para a rede junto com a saída esperada. A diferença entre a saída esperada e a saída da rede é utilizada

para ajustar os parâmetros w_n de cada conexão e esse ciclo se repete até que se atinja um resultado satisfatório. Esse treinamento é conhecido como *back-propagation*, é o mais conhecido e utilizado na literatura de RNAs. (GÜNTHER e FRITSCH, 2010)

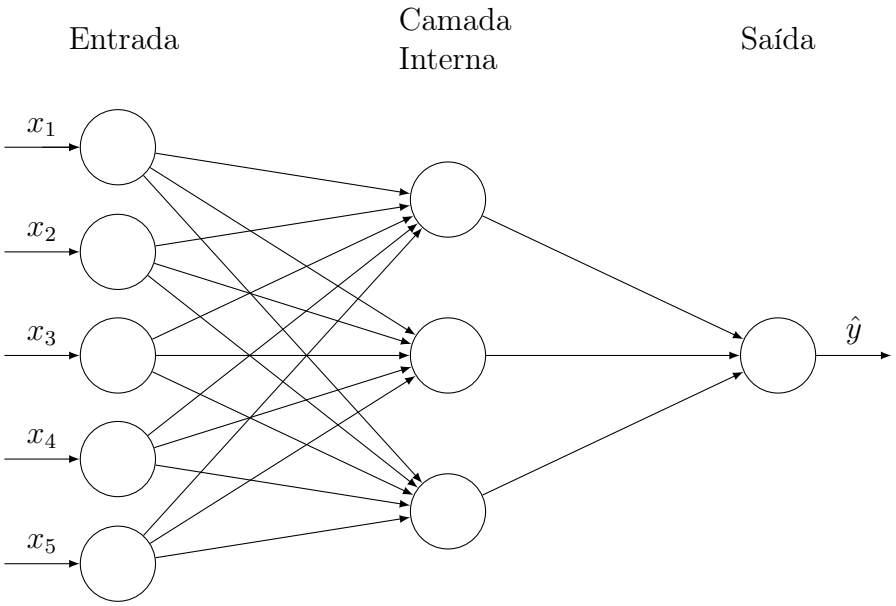


Figura 3.5: Representação de uma Rede Neural Artificial com 5 entradas, uma camada escondida e uma saída.

A Equação 3.2 representa a forma mais comum de se calcular o erro entre a saída da rede, o erro médio quadrático (MSE), representada por $\hat{y}$, e o rotulo conhecido, representado por y. Como dito anteriormente, esse erro é propagado por todas as conexões da rede de forma a corrigir os parâmetros das conexões e frequentemente um outro parâmetro η, chamado de *learning rate*, pode ser empregado para controlar a correção nos parâmetros. (NIELSEN, 2015)

$$error = \frac{1}{2n} \sum_{i=1}^{n} ||y - \hat{y}||^2 \tag{3.2}$$

3.5.1.2 Arquiteturas de Redes Neurais

A escolha da arquitetura da rede neural é muito particular do problema em questão, do tipo de dados, do tamanho do conjunto de dados disponível e dos recursos computacionais disponíveis. Cada arquitetura tem suas próprias características que as tornam adequadas para diferentes tarefas e cenários. Além disso, é possível ainda combinar diferentes arquiteturas ou criar arquiteturas personalizadas para atender a requisitos específicos de um problema particular. À medida que a pesquisa em aprendizado de máquina avança, novas arquiteturas são desenvolvidas para lidar com desafios mais complexos e melhorar o desempenho em várias aplicações.

O *perceptron* apresentado na Seção 3.5.1.1, embora tenha sido o precursor do neurônio artificial, em redes de perceptrons de múltiplas camadas (*Multilayer Perceptrons - MLP*), não é mais tão utilizado por não conseguir generalizar bem para problemas não lineares, e por consequência mais complexos. Por esse motivo, neurônios com funções não lineares se tornaram mais populares pela capacidade de aprendizagem de padrão de maneira mais eficiente para problema mais complicados. (NIELSEN, 2015)

A arquitetura mais utilizada de redes neurais é a *Feedforward*, onde o fluxo das conexões vai sempre em uma direção, da camada de entrada em direção à camada de saída, sem nenhum *loop* entre as conexões. Essa característica torna a *feedforward* simples de ser implementada e eficiente para tarefas de classificação e regressão. Em uma rede *feedforward* típica, cada neurônio em uma camada está conectado a todos os neurônios na camada seguinte, formando uma rede densamente conectada. As camadas intermediárias entre a entrada e a saída são chamadas de camadas ocultas e são responsáveis por aprender representações hierárquicas dos dados. A camada de saída da rede normalmente contém os neurônios responsáveis por produzir as saídas desejadas, que podem ser probabilidades de classes em problemas de classificação ou valores numéricos em problemas de regressão. (NIELSEN, 2015)

Por outro lado, as Redes Neurais Recorrentes (RNNs) permitem *loops* em suas conexões, o que é uma vantagem quando se trabalha com dados sequenciais, como texto, áudio, vídeos ou séries temporais. Essa capacidade de conexões retroativas permite que a saída de um neurônio em uma determinada etapa de tempo sirva como entrada para um neurônio na mesma camada ou em camadas anteriores em um passo subsequente. Esse mecanismo de realimentação cria uma espécie de "memória"na rede, permitindo que informações importantes de etapas anteriores influenciem as etapas posteriores do processamento. Na Figura 3.6 é apresentada uma representação de uma rede neural recorrente. (NIELSEN, 2015)

3.5.1.3 Autoencoders

Os *autoencoders* são modelos de redes neurais artificiais que empregam um mecanismo de treinamento não supervisionado, em que a saída desejada é igual à entrada fornecida. Essencialmente, um *autoencoder* é projetado para aprender uma representação latente dos dados originais. Embora isso possa parecer contra intuitivo à primeira vista, revela-se uma abordagem eficaz para a redução de dimensionalidade e outras aplicações de extração de informação. (BALDI, 2012)

O *autoencoder* consiste em duas partes principais: o codificador (*encoder*) e o decodificador (*decoder*). O codificador recebe os dados de entrada e os mapeia para um espaço latente de dimensionalidade reduzida, por meio de uma série de camadas ocultas. Essas camadas, comumente chamadas de camadas de redução de

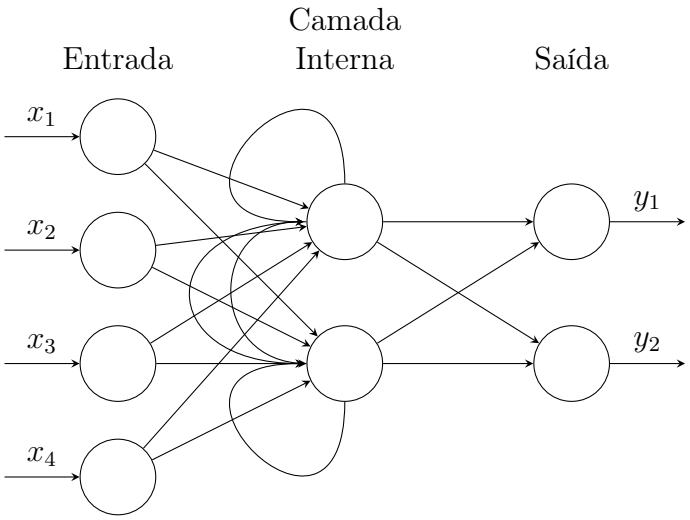


Figura 3.6: Representação de uma Rede Neural Recorrente onde a saída dos neurônios da camada interna também servem de entrada para a própria camada interna.

dimensionalidade, têm a tarefa de extrair características significativas dos dados, capturando padrões e estruturas importantes. A função de ativação utilizada nas camadas ocultas pode variar, desde funções não lineares, como a função de ativação ReLU (*Rectified Linear Unit*), até funções de ativação mais suaves, como a função sigmoidal. A Figura 3.7 apresenta uma representação de um *autoencoder* com uma camada oculta. (ZHANG *et al.*, 2020)

Após a etapa de codificação, o decodificador entra em ação. Ele recebe a representação latente dos dados gerada pelo codificador e tenta reconstruir a entrada. Essa etapa envolve uma série de camadas de decodificação, que gradualmente aumentam a dimensionalidade da representação latente até que a saída seja reconstruída. É importante mencionar que a arquitetura do autoencoder é projetada de forma a forçar o decodificador a aprender uma representação fiel da entrada original, minimizando o erro de reconstrução entre a saída reconstruída e a entrada real. Para isso, é comumente utilizada uma função de perda, como a função de erro médio quadrático (MSE), visto na Equação 3.2. (ZHANG *et al.*, 2020)

O processo de treinamento de um *autoencoder* envolve a otimização iterativa dos pesos das camadas ocultas, buscando minimizar uma função de perda específica. Para isso, são aplicados algoritmos de otimização, como o Gradiente Descendente, que ajustam gradualmente os pesos da rede com base no gradiente da função de perda. Esses algoritmos permitem que o *autoencoder* encontre os pesos ideais que melhor representam os dados de entrada.

Além disso, é comum a utilização de técnicas de regularização durante o treinamento do *autoencoder*. Essas técnicas visam evitar o *overfitting*, um problema em

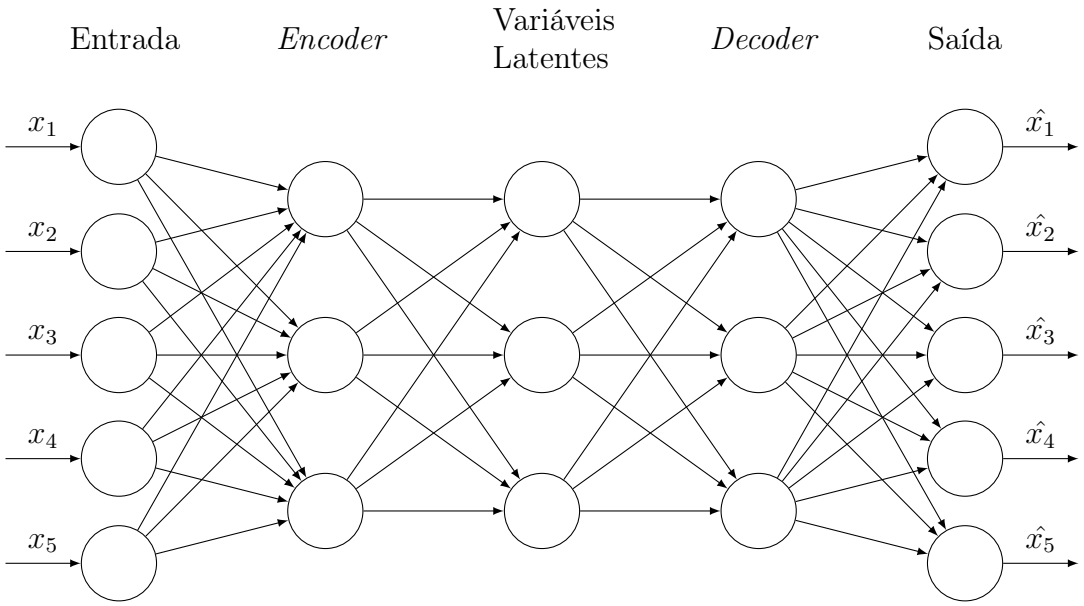


Figura 3.7: Representação de um *autoencoder*, com uma camada de *encoder*, uma camada oculta destacada como variáveis latentes, uma camada de *decoder* com objetivo de fazer a reconstrução dos dados original na camada de saída.

que o modelo se ajusta demasiadamente aos dados de treinamento, prejudicando sua capacidade de generalização para novos dados. Uma abordagem comum é a adição de termos de penalização L1 ou L2 à função de perda, que controlam a complexidade do modelo ao restringir os valores dos pesos (JING *et al.*, 2020; YU *et al.*, 2013; ZHAO *et al.*, 2018). Isso ajuda a evitar pesos excessivamente grandes e a promover a aprendizagem de representações mais robustas.

Uma das principais aplicações do *autoencoder* é a reconstrução e compressão de dados. Ao aprender uma representação compacta e informativa dos dados de entrada, o *autoencoder* pode ser utilizado para reconstruir os dados originais, bem como gerar novas amostras semelhantes aos dados de treinamento. Além disso, o *autoencoder* também é amplamente utilizado como um método de pré-treinamento em tarefas de aprendizado supervisionado, onde a representação latente aprendida pode ser utilizada como entrada para outras redes neurais, melhorando o desempenho geral do modelo.

Em resumo, o *autoencoder* é uma poderosa ferramenta no campo do aprendizado de máquina não supervisionado, capaz de aprender representações latentes significativas dos dados de entrada. Sua estrutura composta por um codificador e um decodificador permite a extração de características importantes e a reconstrução precisa dos dados originais, suas principais aplicações são redução de dimensionalidade e extração de características de dados.

3.5.2 *Support Vector Machine* (SVM)

Support Vector Machine é um método de aprendizado de máquina amplamente utilizado em problemas de reconhecimento de padrões, bem como em problemas de regressão. O SVM é reconhecido por sua eficácia e robustez no tratamento de problemas complexos. Ele se destaca ao lidar com problemas complexos devido à sua habilidade em encontrar hiperplanos que separam claramente as diferentes classes no espaço de características.

Uma das principais características do SVM é sua capacidade de trabalhar tanto em cenários de classificação linear como não linear. Em problemas linearmente separáveis, o modelo procura encontrar o hiperplano que maximiza a margem entre as classes, ou seja, a maior distância entre os exemplos mais próximos de cada classe. Essa margem ampla contribui para uma maior capacidade de generalização do modelo, reduzindo o risco de *overfitting*, especialmente em conjuntos de dados de treinamento pequenos. (BURGES, 1998)

Quando os dados não são linearmente separáveis, o SVM mapeia os exemplos de treinamento utilizando funções *kernel*, dessa forma, os pontos são projetados para um espaço de dimensão superior a fim de maximizar a largura da lacuna entre as duas categorias. Novos exemplos são então mapeados para esse mesmo espaço e previstos como pertencentes a uma categoria com base em qual lado da lacuna eles se encontram. Isso permite ao SVM lidar com problemas mais complexos, onde as fronteiras de decisão são não lineares. Diferentes funções de *kernel*, como o *kernel* polinomial, gaussiano (RBF) e sigmoide, podem ser usadas para a transformação dos dados, possibilitando uma adaptação mais flexível aos padrões presentes nos dados. (BURGES, 1998)

Outra característica importante do SVM é a sua capacidade de controlar o compromisso entre a margem e o erro de classificação através do parâmetro de regularização C. Valores maiores de C resultam em uma margem menor, mas com menos erros de classificação nos dados de treinamento. Por outro lado, valores menores de C buscam uma margem maior, mesmo que isso signifique permitir alguns erros de classificação. A escolha adequada desse parâmetro é fundamental para o desempenho do modelo.

O SVM também pode ser estendido para problemas de classificação multiclasse, utilizando estratégias como "um contra um"ou "um contra todos". Além disso, ele pode ser adaptado para tarefas de regressão, conhecido como *Support Vector Regression* (SVR), onde o objetivo é encontrar uma função que se ajuste aos dados com um erro máximo especificado.

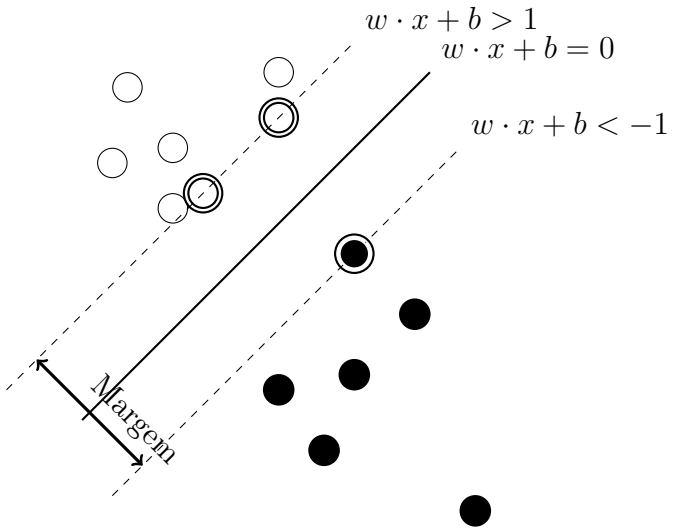


Figura 3.8: Representação gráfica de como SVM define uma margem para separar os dados em duas classes.

Além disso, o SVM é conhecido por sua capacidade de generalização em cenários com poucas amostras, tornando-o uma escolha confiável para tarefas que envolvem conjuntos de dados limitados. Ele também é eficiente em lidar com problemas de alta dimensão, como classificação de texto e análise de imagens, onde o número de características pode ser muito grande. A Figura 3.8 ilustra como o SVM é capaz de separar os dados.

3.6 Aprendizado com Viés

Dentro da área de aprendizado de máquina, o termo viés tem um significado específico e crucial. Ele representa qualquer desvio de generalização que não desconsidere a totalidade dos dados observados (DIETTERICH e KONG, 1995). Em outras palavras, o viés refere-se à tendência de um modelo em fazer suposições consistentes e sistemáticas sobre os dados, independentemente de serem corretas ou não.

O viés pode ser tanto benéfico quanto prejudicial. Em alguns casos, um certo grau de viés é necessário para que o modelo possa fazer previsões úteis quando enfrenta dados nunca vistos antes. No entanto, um viés indesejado ou excessivo pode levar a previsões incorretas ou injustas, especialmente quando os dados de treinamento não são representativos da realidade ou contêm preconceitos históricos. (HELLSTRÖM *et al.*, 2020)

Além disso, é essencial diferenciar entre viés e variância em aprendizado de máquina. Enquanto o viés se refere à precisão das previsões médias do modelo em relação aos valores reais, a variância se refere à consistência das previsões do modelo em diferentes conjuntos de dados. Um equilíbrio adequado entre viés e variância

é fundamental para criar modelos que sejam tanto precisos quanto generalizáveis. (DIETTERICH e KONG, 1995)

Diversas abordagens têm sido propostas na literatura para lidar com o viés. Por exemplo, técnicas de reamostragem, como o *oversampling* e o *undersampling*, são usadas para equilibrar conjuntos de dados desequilibrados, reduzindo assim o viés em relação à classe minoritária. Métodos de aprendizado justo, como o *adversarial training*, buscam minimizar o viés ao introduzir adversários que penalizam decisões injustas. Além disso, técnicas de regularização, como a regularização L1 e L2, podem ser aplicadas para reduzir o viés ao penalizar modelos complexos que se ajustam demais aos dados de treinamento.

Outra abordagem envolve a interpretabilidade do modelo. Ferramentas como LIME e SHAP têm sido usadas para entender como os modelos tomam decisões e identificar possíveis fontes de viés. Ao compreender as contribuições de cada característica para a decisão final, é possível identificar e corrigir preconceitos indesejados.

O viés de *fairness* é particularmente preocupante em aplicações de aprendizado de máquina que têm implicações diretas na vida das pessoas, como sistemas de recomendação de empregos, decisões de crédito ou diagnósticos médicos. Por exemplo, se um modelo de aprendizado de máquina treinado em dados históricos de contratação favorece candidatos de um gênero específico porque esse gênero foi historicamente mais contratado, isso é um exemplo de viés de *fairness*.

Em resumo, o viés de *fairness* em aprendizado de máquina é uma extensão do conceito geral de viés, focando especificamente em preconceitos e discriminações que resultam em tratamento injusto. Abordar esse viés requer uma combinação de técnicas de pré-processamento, modelagem e pós-processamento, todas voltadas para garantir que os modelos sejam justos e não perpetuem discriminações históricas ou sociais.

Antes de tudo, é necessário compreender profundamente as causas e origens do viés para lidar adequadamente com ele em modelos de aprendizado. Isso envolve identificar os atributos associados a essa tendência e sua possível relação com outros atributos. Além disso, é importante buscar métricas adequadas para medir o viés e, por fim, ajustar o algoritmo para minimizar o impacto indesejado nos resultados.

Capítulo 4

Covariance Fair Autoencoder

Neste capítulo, será apresentada a proposta desta dissertação, que consiste em um método de *fair representation* que faz uso de um *autoencoder* com uma função de custo com uma penalidade. O objetivo do método é remover parte da informação relacionadas ao atributo sensível, e com isso, extrair uma representação mais justa dos dados. Nas próximas seções, detalharemos como esse método funciona, conceitos teóricos relevantes e alguns trabalhos relacionados com a proposta deste trabalho.

4.1 Definição do Problema

O problema de *fairness* refere-se à preocupação com a equidade e imparcialidade na aplicação de algoritmos e sistemas computacionais (ZAFAR *et al.*, 2015). Essa questão pode impactar negativamente indivíduos pertencentes a grupos minorizadas da sociedade. Especificamente, busca-se evitar a perpetuação da discriminação com base em características sensíveis, como gênero, raça, idade ou orientação sexual. O objetivo é garantir que esses algoritmos não reproduzam vieses e preconceitos presentes nos dados, e que não resultem em tratamentos desiguais ou injustos para diferentes grupos.

Os trabalhos nesta área geralmente têm como objetivo o desenvolvimento de técnicas e métricas para medir, mitigar e monitorar possíveis vieses e discriminações em algoritmos e sistemas de tomada de decisão. Isso envolve o estudo de algoritmos de aprendizado de máquina, *representation learning*, análise de dados e outros campos relacionados, a fim de promover a equidade e a justiça social em aplicações práticas (ZEMEL *et al.*, 2013).

Conforme discutido no Capítulo 2, existem três abordagens principais para lidar com a questão da *fairness*: pré-processamento, *in-processing* e pós-processamento. Neste trabalho, focaremos na abordagem de pré-processamento, na qual nosso objetivo é aplicar métodos de transformação aos dados injustos, tornando-os mais equitativos em relação aos atributos sensíveis.

A utilização de métodos de *fairness* na etapa de pré-processamento apresenta diversos desafios, incluindo o *trade-off* entre a equidade dos dados e o desempenho. Uma vez que a aplicação de técnicas de pré-processamento para garantir a equidade pode resultar em uma possível redução da precisão ou de outros aspectos do desempenho do modelo, uma vez que essas técnicas acabam alterando os dados originais de alguma forma.

Além disso, a utilização de métodos de pré-processamento em cenários de *fairness* também podem enfrenta questões da representatividade inadequada de indivíduos pertencentes a grupos protegidos. Grupos diferentes podem apresentar distribuições e características distintas. Desenvolver métodos de pré-processamento que levem em consideração essa diversidade e que sejam aplicáveis a uma variedade de tipos de dados e grupos não é uma tarefa simples.

A preocupação com a generalização também emerge como uma complexidade adicional, pois garantir a eficácia e equidade dos métodos de pré-processamento em diferentes conjuntos de dados e cenários representa um desafio notável. O que funciona bem em um contexto pode não ser prontamente generalizável para outros.

4.2 *Covariance Fair Autoencoder*

Este trabalho tem como objetivo apresentar um método de pré-processamento para bases de dados com viés de preconceito em relação a um grupo de indivíduos historicamente desfavorecidos. Esse pré-processamento é realizado por meio de um autoencoder com uma função de custo que inclui uma penalidade, visando remover a informação sensível e extrair as variáveis latentes da base de dados, gerando assim uma representação mais justa dos dados. O fluxograma da proposta denominada *Covariance Fair Autoencoder* (CFA) é ilustrado na Figura 4.1.

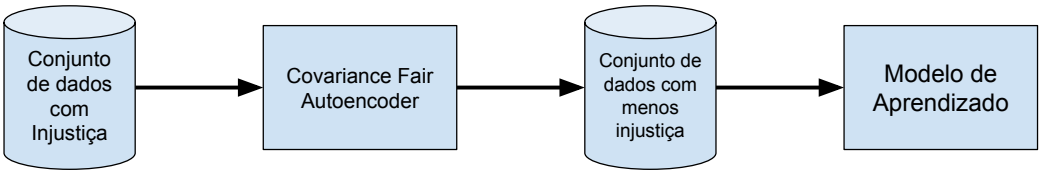


Figura 4.1: Processo de *fair representation* com um autoencoder capaz de gerar dados mais justos que pode ser utilizado por outros modelos de *machine learning*.

Com uma representação mais justa dos dados, torna-se possível extrair padrões e analises desses conjuntos sem impactar negativamente indivíduos de grupos protegidos, promovendo assim a equidade e a justiça em tarefas como classificação, regressão, agrupamento e geração de dados. Além disso, essa representação pode

ser facilmente incorporada em outros algoritmos de aprendizado de máquina existentes, proporcionando uma base sólida para a tomada de decisões automatizadas mais éticas e imparciais em uma variedade de cenários aplicados.

4.2.1 *Fair Autoencoder*

Autoencoder são modelos de aprendizado de máquina que têm aplicações em uma variedade de problemas. Alguns dos principais cenários em que os *autoencoders* são utilizados incluem: Redução de dimensionalidade, onde no processo de reconstrução as variáveis latentes de menor dimensionalidade conseguem capturar as principais características dos dados originais. Extração de características para se aprender as informação mais relevantes em um conjunto bruto de dados. Detecção de anomalias, pela capacidade de aprender os padrões dos dados de entrado os *autoencoders* podem detectar elementos considerados anomalias e *outliers* de um conjunto de dados.

O *autoencoder* é composto por duas partes principais: o *encoder* e o *decoder*. Onde o primeiro é responsável por mapear os dados de entrada em uma representação latente de dimensionalidade reduzida, utilizando técnicas de compressão e extração de características relevantes. Essa etapa é realizada por meio de camadas ocultas, como camadas convolucionais, recorrentes ou densas, que aplicam transformações não lineares nos dados para capturar relações complexas e reduzir sua dimensionalidade.

Pela capacidade dos *autoencoders* em extrair variáveis latentes de dados, e pela maneira como os parâmetros desse método são aprendidos, com a correção dos pesos a partir do gradiente do erro da função de custo, escolhemos o *autoencoder* como base para o método proposto. Porém, para corrigir os pesos de maneira a desassociar a informação sensível buscamos uma função de similaridade a ser incluída na função de custo de forma a forçar a modelo a aprender uma representação com menos influência do atributo sensível.

Devido à capacidade dos *autoencoders* de extrair variáveis latentes dos dados, e considerando a forma como os parâmetros desse método são aprendidos (com a correção dos pesos a partir do gradiente do erro da função de custo) visto no Capítulo 3, escolhemos o *autoencoder* como base para o método proposto. No entanto, para ajustar os pesos de modo a desassociar a informação sensível, buscamos uma função de similaridade a ser incluída na função de custo. Isso visa forçar o modelo a aprender uma representação com menor influência do atributo sensível, detalharemos mais a função de similaridade na Seção 4.2.3.

Outra característica do método proposto é o uso exclusivo dos dados de entrada de uma base de dados, sem a utilização dos rótulos. Com isso, adotamos uma abordagem não supervisionada, empregando um *autoencoder* para aprender

uma representação mais imparcial da base. Dessa forma, a reconstrução dos dados não é afetada por qualquer viés presente no *target* original, tornando a solução completamente desvinculada de um processo supervisionado.

O novo conjunto de dados resultante pode ser utilizado como entrada para outros modelos de aprendizado de máquina. Uma vez que parte dos dados sensíveis foi removida desse novo conjunto, o método proposto consegue mitigar o *disparate treatment*. Simultaneamente, lida com o *disparate impact*, uma vez que o método visa eliminar as características sensíveis da representação justa como um todo, evitando assim o uso indireto da informação sensível.

Portanto, o *fair autoencoder* proposto é capaz de extrair uma representação mais equitativa dos dados. Consequentemente, a solução proposta consegue remover eficazmente parte das informações sensíveis dos dados, sem perder informações cruciais para a tarefa em questão. Isso torna o método promissor para a construção de sistemas mais éticos e confiáveis, especialmente em contextos sensíveis onde a equidade é um objetivo primordial.

4.2.2 Variáveis Latentes

Variáveis latentes refere-se a um conceito fundamental da estatística e de outras áreas relacionadas, referindo-se a um conjunto de características que não podem ser diretamente observadas ou medidas. Em contraste, as variáveis observáveis são aquelas que podem ser observadas ou medidas diretamente e são geralmente utilizadas para coletar dados e informações. Embora as variáveis latentes não sejam diretamente observáveis, elas influenciam os dados observados de maneira indireta, moldando os padrões e relacionamentos entre as variáveis observáveis (BISHOP, 1998).

As variáveis latentes são amplamente utilizadas em várias disciplinas e contextos, incluindo estatística, ciência de dados, psicologia, economia, engenharia e aprendizado de máquina, entre outras. Elas são especialmente úteis quando lidamos com fenômenos complexos e multifacetados, nos quais não é possível medir todos os aspectos relevantes diretamente.

Para extrair e entender as variáveis latentes nos dados, são utilizadas diversas técnicas estatísticas e de aprendizado de máquina. Algumas dessas técnicas incluem a Análise de Componentes Principais (PCA), Decomposição de Valores Singulares (SVD), *autoencoders* e a Modelagem de Equações Estruturais (SEM). Cada técnica possui suas vantagens e é aplicada de acordo com o contexto e os objetivos específicos da análise (BISHOP, 1998).

No contexto de *fairness* e *fair representation*, a extração de variáveis latentes com menor discriminação em relação aos atributos sensíveis, é uma abordagem

amplamente adotada e de grande relevância, como nos trabalhos apresentados por LOUIZOS *et al.* (2016); SATTIGERI *et al.* (2019); MADRAS *et al.* (2018). O objetivo é encontrar uma representação latente dos dados que seja mais equitativa e imparcial em relação a grupos protegidos ou atributos sensíveis, ao mesmo tempo em que outras características úteis dos dados originais consigam ser extraídos.

Como as variáveis latentes estão relacionadas diretamente com as variáveis observáveis, elas compartilham as mesmas informações úteis. Com isso, este trabalho busca utilizar uma técnica de extração de variáveis latentes por intermédio de um *autoencoder*, ao mesmo tempo em que tentamos remover a informação sensível dessa representação.

Este trabalho propõe o uso de um *autoencoder* para extrair variáveis latentes de dados com viés de *fairness*. O objetivo é remover parte da informação sensível no processo, garantindo que as variáveis latentes apresentem menos viés que o conjunto original.

4.2.3 *Covariance Loss Function*

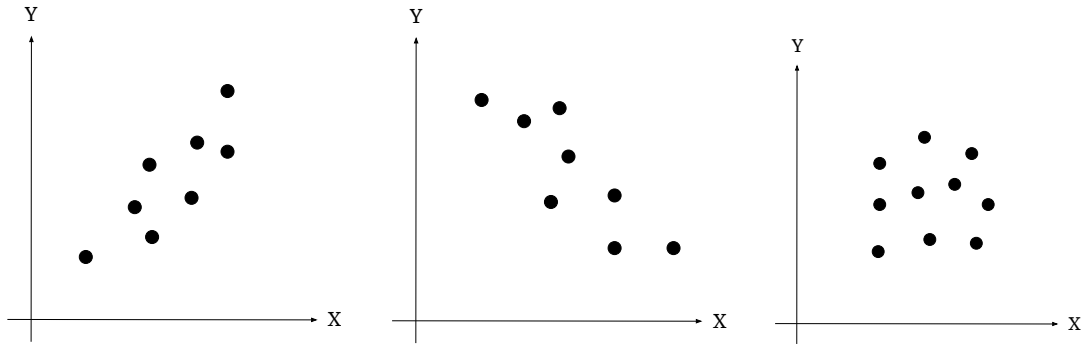
A covariância é uma medida estatística que permite avaliar a variabilidade conjunta de duas variáveis aleatórias, representando o grau de associação entre elas. Valores positivos indicam uma relação direta, valores negativos indicam uma relação inversa, e uma covariância de zero indica independência linear entre as variáveis. A Figura 4.2 ilustra exemplos visuais dessas possíveis relações (RICE, 2006).

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] \tag{4.1}$$

A fórmula para calcular a covariância entre duas variáveis X e Y é dada pela Equação 4.1. Nessa fórmula, o processo consiste em calcular a esperança das diferenças entre os valores de X e seu valor esperado μ_x, multiplicada pelas diferenças entre os valores de Y e seu valor esperado μ_y. Essa abordagem permite medir o quanto os desvios conjuntos das variáveis se afastam ou se aproximam dos valores esperados de cada variável.

De maneira geral, determinar o grau de dependência entre duas variáveis pode ser uma tarefa desafiadora, especialmente quando os valores para calcular a covariância são pequenos em termos absolutos. A escala da medida por si pode não ser suficiente para determinar a intensidade da relação. Isso acontece porque a covariância depende das unidades das variáveis envolvidas, o que pode obscurecer a interpretação direta dos resultados. (RICE, 2006)

Devido a essas características, optamos por usar a função de covariância entre o atributo sensível original e o atributo sensível reconstruído pela rede como penalidade na função de custo do *fair autoencoder*. Assim, uma vez que o atributo



(a) (b) (c)

Figura 4.2: Relações lineares possíveis que duas variáveis podem ser determinadas pela função de covariância. (a) Relação direta (b) Relação inversa (c) Independente .

sensível reconstruído depende dos parâmetros da rede, a ideia da proposta é que, ao incluir essa penalidade na função de custo a ser minimizada, o *autoencoder* seja incentivado a aprender uma representação que assegure a independência entre as variáveis latentes extraídas do atributo sensível original.

$$loss = \frac{1}{n}\sum_{i=1}^{n}||x_i - \hat{x}_i||^2 + \lambda \cdot |\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)(\hat{x}_{ij} - \bar{\hat{x}}_j)| \tag{4.2}$$

Na Equação 4.2, apresentamos a função de custo adotada pelo *fair autoencoder*. Nessa equação, n denota o número de instâncias no conjunto de dados, x_i e $\hat{x}_i$ representam os vetores de entrada e de saída do *autoencoder*, respectivamente. Além disso, m representa a quantidade de atributos sensíveis a serem considerados na penalidade, x_{ij} e $\hat{x}_{ij}$ correspondem aos atributos sensíveis originais e os reconstruídos pelos *autoencoder*, respectivamente, enquanto $\bar{x}_j$ e $\bar{\hat{s}}_j$ denotam os valores esperados do atributo sensível j original e reconstruído pelo *autoencoder*, respectivamente. O parâmetro λ tem como função controlar o nível de justiça desejada.

A primeira parte da Equação 4.2 consiste no erro quadrático médio (MSE), amplamente utilizado em redes neurais e *autoencoders* durante a etapa de treinamento. O segundo termo representa a covariância entre os atributos sensíveis originais e os atributos sensíveis reconstruído pelo *autoencoder*. Conforme discutido no início desta seção, a função de covariância é capaz de determinar a associação linear entre duas variáveis. Outro ponto importante a ser destacado é que para que essa desassociação seja eficiente os valores na equação devem ser normalizados previamente.

Portanto, ao incluir o módulo dessa penalidade na função de custo, estamos induzindo os parâmetros do *fair autoencoder* a reconstruir uma representação mais independente dos atributos sensíveis originais. Como foi visto no Capítulo 3, as

redes neurais aprendem propagando o erro da função de custo por todas as conexões para corrigir os parâmetros. Dessa forma, a correção dos parâmetros da rede será no sentido de extrair variáveis latentes com menor grau de dependência linear da informação sensível original.

Após a extração das variáveis latentes do *fair autoencoder*, as mesmas podem ser utilizadas por um modelo de aprendizado de forma que se consiga extrair conhecimento dos dados, sem que grupos protegidos sejam prejudicados. Neste trabalho, foi utilizado um SVM após a aplicação do método proposto e medimos os resultados em relação ao *statistical parity*, acurácia, F1 *score* e Coeficiente de Correlação de Matthews (MCC) e comparamos com outros implementações.

Vale destacar o Coeficiente de Correlação de Matthews, que será o balizador da qualidade dos resultados obtidos neste trabalho. Esta métrica representa a qualidade de classificações binárias. Ela leva em consideração verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, e é uma alternativa ao F1 *score* quando os conjuntos de dados apresentam classes desequilibradas.

Na Equação 4.3, é apresentado como o MCC é calculado. Onde TP representa a quantidade de verdadeiros positivos, TN a quantidade de verdadeiros negativos, FP a quantidade de falsos positivos e FN a quantidade de falsos negativos. O resultado dessa métrica varia entre -1 e 1: 1 representa uma predição perfeita, 0 indica um resultado aleatório e -1 denota uma discordância total entre as predições e os verdadeiros valores.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.3}$$

Uma das principais vantagens do MCC em relação a outras métricas é que ele é um indicador equilibrado, mesmo quando as classes estão em proporções muito diferentes. Por exemplo, em situações em que uma classe é muito mais frequente que a outra, a acurácia pode não ser uma boa métrica, pois um classificador que sempre prevê a classe majoritária terá alta acurácia, mas não será necessariamente útil. O MCC, por outro lado, considera ambas as classes de forma equilibrada.

4.3 Trabalhos Relacionados

Fairness é um tema amplamente discutido na literatura, especialmente devido ao aumento significativo de sistemas de tomada de decisão baseados em modelos de aprendizado de máquina. Esses modelos têm o potencial de reproduzir vieses presentes nos dados de treinamento se não forem adequadamente tratados. Quando esse viés afeta um grupo historicamente desfavorecido, por exemplo, quando uma base de dados possuí dados que desfavorece mulheres em decisões de crédito, o modelo

pode perpetuar esse preconceito para decisões futuras.

Diante desse desafio, várias abordagens foram desenvolvidas, algumas buscando evitar a utilização de atributos sensíveis no processo de tomada de decisão, enquanto outras procuram garantir que o resultado da classificação ou regressão não seja influenciado por esses atributos, tornando a classe positiva independente dos dados sensíveis. No entanto, independentemente da abordagem adotada, todos os métodos têm como objetivo reduzir a discriminação ou garantir a igualdade estatística (*statistical parity*).

No trabalho proposto por ZAFAR *et al.* (2015), é apresentado uma abordagem chamada *Decision Boundary Covariance*, na qual os autores introduzem uma restrição na função de custo. Essa restrição tem como objetivo limitar a covariância entre o atributo sensível e a classificação feita pelo modelo por um parâmetro c, permitindo assim controlar o nível de tolerância ao viés desejado.

A restrição proposta neste estudo é modelada utilizando a covariância do atributo sensível e o sinal da decisão dos modelos empregados. Ao incorporar a restrição diretamente nos modelos utilizados, este trabalho se enquadra no método de *in-processing* de *fairness*. O uso da função de covariância inspirou a proposta desta dissertação em apresentar uma abordagem de pré-processamento, utilizando a covariância, para gerar um novo conjunto de dados.

ZEMEL *et al.* (2013) apresentam um dos primeiros estudos sobre *fair representation*, denotado *Learn Fair Representation* (LFR). Nesse contexto, os dados originais são transformados em um novo conjunto de dados com características semelhantes, porém sem o viés negativo em relação aos grupos desfavorecidos presentes no conjunto original. Para alcançar esse objetivo, os autores mapeiam as entradas originais em um novo espaço de representação utilizando uma distribuição de probabilidade que busca remover as informações relacionadas ao atributo sensível dos indivíduos. Essa nova distribuição utiliza o conceito de *statistical parity* para garantir que, nesse novo espaço, os dados sejam independentes em relação ao atributo sensível.

Os autores SATTIGERI *et al.* (2019) apresentam uma proposta de *Generative Adversarial Network* (GAN), onde duas redes neurais competem, uma como geradora e outra como discriminadora. O objetivo da primeira rede é gerar dados similares aos dados originais, enquanto o discriminador busca distinguir os dados reais dos dados gerados. Utilizando o *statistical parity* na rede geradora, o estudo busca gerar dados livres do viés associado ao atributo sensível. Outro ponto que vale destacar deste trabalho foi a utilização de bases de dados de imagens, o que não é tão comum nos trabalhos da área. No entanto, os autores obtiveram bons resultados na redução do preconceito nessas bases.

No artigo apresentado por LOUIZOS *et al.* (2016), os autores exploram o desafio de aprender uma representação que seja invariante ao atributo sensíveis nos dados,

enquanto retêm o máximo possível das informações restantes. Sua abordagem é baseada em uma arquitetura de *variational autoencoder*, com prioridades que promovem a independência entre fatores de variação sensíveis e latentes. Isso garante que qualquer processamento subsequente, como classificação, seja realizado em uma representação livre da informações sensíveis. Para aprimorar ainda mais a remoção de dependências, eles introduzem um termo de penalidade adicional baseado na medida *Maximum Mean Discrepancy*. Através de experimentos, os autores demonstram que seu método supera técnicas anteriores na eliminação de fontes indesejadas de variação, preservando representações latentes informativas.

Outro trabalho notável na área de *fair representation* foi proposto por MADRAS *et al.* (2018). Neste estudo, os autores introduzem o conceito de *Adversarially Fair and Transferable Representations*. Eles fornecem uma fundamentação teórica para integrar redes neurais adversárias uma estratégia que busca representações latentes dos dados, assegurando simultaneamente utilidade e justiça. Além disso, o método proposto demonstrou capacidade de realizar *transfer learning*, transferindo as representações aprendidas para outras aplicações, mantendo métricas de *fairness* satisfatórias.

Capítulo 5

Resultados e Discussões

Neste capítulo serão apresentado os experimentos realizados nesta dissertação, assim como as características das base de dados escolhidas para os experimentos e por fim, discutiremos os resultados e apresentando gráficos e tabelas com os resultados.

5.1 Base de Dados

Para realizar os experimentos, foram utilizadas duas bases de dados distintas, cada uma com características e atributos diferentes. Essas bases de dados são amplamente empregadas na literatura e contêm informações de pessoas reais. Os conjuntos de dados utilizados e suas características são os seguintes:

Adult Income é uma base de dados proveniente de um Censo dos Estados Unidos. Ela é composta por 48.842 indivíduos e possui 15 atributos: 6 são numéricos, 7 são categóricos e 2 são binários. A classificação é binária e indica se o indivíduo possui ou não uma renda superior a $50.000 dólares. No *Adult Income*, os atributos gênero, raça e idade são considerados sensíveis e apresentam viés de *fairness*. Neste trabalho, o atributo sensível escolhido foi o gênero, por ser o mais frequentemente utilizado e, portanto, oferecer uma maior possibilidade de comparação com outros trabalhos (LE QUY *et al.*, 2022). Esta base de dados contém 3.620 registros com algum atributo de valor nulo. Em diversos trabalhos, como os apresentados por CHOI *et al.* (2020); IOSIFIDIS e NTOUTSI (2018); ZAFAR *et al.* (2017), os autores optaram por remover essas instâncias para evitar problemas na análise. Nesta dissertação, adotou-se a mesma abordagem, resultando em 45.222 instâncias restantes.

A Figura 5.1 apresenta a distribuição da quantidade de indivíduos de cada gênero nas classes positiva e negativa (MCDONOUGH *et al.*, 1997). Como é comum em conjuntos de dados com esse tipo de problema, a quantidade de indivíduos do grupo protegido é desproporcionalmente menor se comparado ao grupo não protegido, tanto em números absolutos quanto em relação a classe positiva.

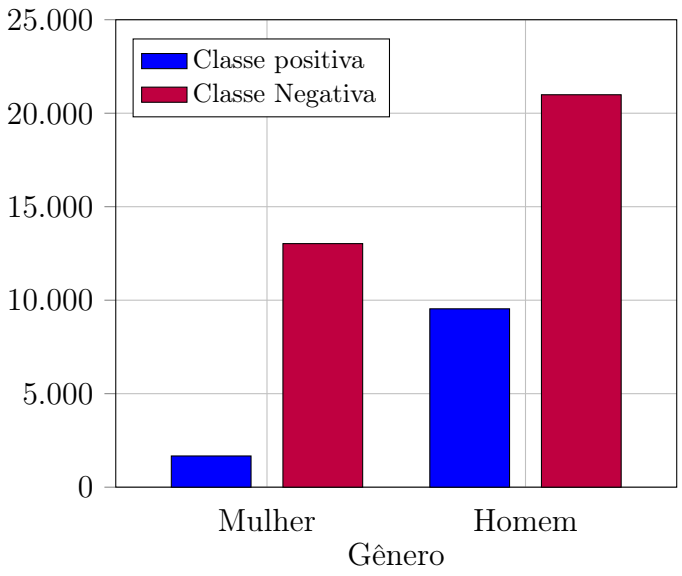
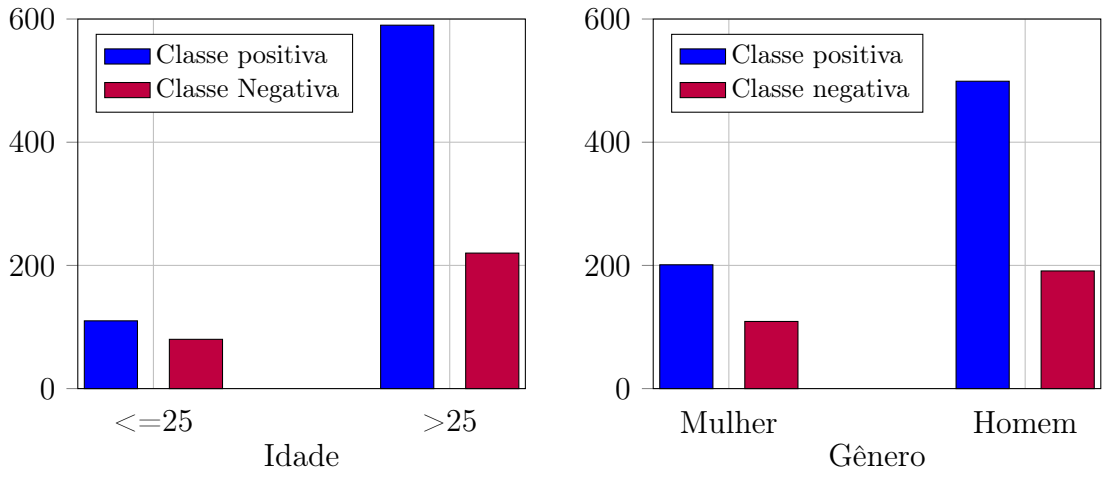


Figura 5.1: Quantidade de indivíduos do grupo protegido, mulheres, e não protegido, homens, pertencentes a classe positiva e negativa da Base de dados *Adult Income*.

German Credit consiste em uma base de dados de análise de crédito de contas bancárias, onde a classe determina se existe ou não um risco de crédito. A base possui 1.000 instâncias, sem nenhum valor faltante, e 21 atributos, sendo 13 categóricos, 7 numéricos e 1 binário. Dois atributos sensíveis foram considerados em diferentes execuções: idade e gênero, onde indivíduos com idade menor do que 25 anos e mulheres compõe os grupos protegidos respectivamente. As Figuras 5.2(a) e 5.2(b) apresentam informações sobre os atributos sensíveis da base. É importante ressaltar que os atributos *personal-status* e *gender* aparecem juntos no conjunto original, e, por isso, foi realizado um processo de pré-processamento para separar a informação do gênero como atributo sensível (LE QUY *et al.*, 2022).

Essa base de dados, embora tenha poucas instâncias e apresente instabilidade nas métricas de desempenho quando aplicamos modelos de aprendizado, é amplamente utilizada em outros trabalhos da área. Observando a Figura 5.1, nota-se que, embora os grupos protegidos sejam menos representativos, a classe positiva possui mais instâncias do que a classe negativa.



(a) (b)

Figura 5.2: (a) Quantidade de indivíduos do grupo protegido, idade menor ou igual a 25, e não protegido, idade maior do que 25, pertencentes a classe positiva e negativa da base *German Credit* (b) Quantidade de indivíduos do grupo protegido, Mulheres, e não protegido, Homens, pertencentes a classe positiva e negativa da base *German Credit*.

5.2 Organização dos Experimentos

Para avaliar a proposta apresentada no Capítulo 4, alguns experimentos foram conduzidos utilizando os conjuntos de dados apresentados na seção anterior, com o objetivo de avaliar o desempenho da proposta e encontrar os parâmetros necessários para se atingir determinado nível de justiça em relação aos atributos sensíveis, considerando as métricas escolhidas para este trabalho.

Nos primeiros experimentos, foram utilizados hiper-parâmetros *default* dos modelos empregados, bem como testes manuais com diferentes valores e tipos de arquitetura de rede, com o objetivo de validar a proposta. Posteriormente, uma nova versão desse experimento foi desenvolvida, aplicando métodos de *bayesian search* e *random search* com *cross validation* para encontrar hiper-parâmetros que otimizassem as métricas desejadas em cada etapa do processo. Os resultados e a configuração dessa nova versão serão detalhados nas seções subsequentes.

Após o ajuste dos parâmetros, realizamos uma comparação entre o método proposto, denominado *Covariance Fair Autoencoder* (CFA), e outras duas variações desse método que desconsideram a penalidade na *fair loss function*. A primeira variação, chamada de 'Regular', não aplica nenhuma técnica de *fairness*, ela ainda faz uso da extração de variáveis latentes, porém com o parâmetro $\lambda = 0$, permitindo-nos avaliar o impacto real da nossa proposta. A segunda variação é similar à primeira, mas com a remoção do atributo sensível. Dessa forma, ao comparar o método pro-

posto com implementações que fazem uso do mesmo fluxo de transformações de dados, mas sem a penalidade na função de custo do *autoencoder*, que remove parte da informação sensível, nós podemos avaliar a eficácia do nosso método em gerar um conjunto mais justo de dados.

Em seguida, apresentamos uma comparação entre os resultados obtidos aqui, com outras implementações de trabalhos de *fairness* da literatura. Foram selecionados trabalhos que utilizaram os mesmos conjuntos de dados e consideraram os mesmos atributos sensíveis, assim como métricas similares para avaliar o desempenho da proposta.

Por último, utilizamos uma técnica de *data visualization* chamada t-SNE (*t-Distributed Stochastic Neighbor Embedding*), inicialmente proposta por VAN DER MAATEN e HINTON (2008) e adotada por LOUIZOS *et al.* (2016) para visualizar dados de múltiplas dimensões em 2D ou 3D, mantendo relações de grupos e proximidades entre instâncias. Com isso em mente, realizamos uma comparação entre os dados originais das bases com a saída do *fair autoencoder*, com o objetivo de demonstrar que a distribuição da representação mais justa é mais heterogênea do que a original em relação ao atributo sensível.

Para aplicar as métricas e comparar os experimentos da proposta, treinamos o modelo SVM usando as variáveis latentes extraídas da *Covariance Fair Autoencoder*, em seguida, avaliamos o conjunto de teste das bases para obter as métricas de *statistical parity*, acurácia, F1 *score* e mcc. Escolhemos esse modelo devido ao seu amplo uso na literatura em tarefas de classificação, assim como as bases de dados selecionadas para este estudo. Além disso, o SVM é considerado um modelo robusto e eficiente.

5.2.1 Configuração dos Hiper-parâmetros

Para o primeiro experimento, em cada conjunto de dados, dividiu-se os dados em treino e teste. Utilizou-se uma divisão de 80% para treino e 20% para teste no *Adult Income* e 70% para treino e 30% para teste no *German Credit*. Essa diferença se deu por conta da quantidade de instâncias em cada conjunto. Como o *German Credit* possui apenas 1.000 instâncias, optamos por utilizar 30% para teste para que esse conjunto fosse o mais representativo possível.

Antes da aplicação do método proposto, foram realizados alguns pré-processamentos nas bases de dados, como a transformação de atributos categóricos para o formato de *one hot encoding*. Essa técnica foi escolhida pelo seu amplo uso na literatura para transformar dados categóricos em dados numéricos sem gerar ambiguidade, o que acabou aumentando o número de colunas em relação à configuração original dos conjuntos de dados. Além disso, na etapa de autoencoder, uma norma-

lização dos valores foi aplicada para que todos os valores ficassem na mesma escala, variando entre 0 e 1. Isso facilita a convergência do treinamento do *autoencoder*.

Já na etapa de classificação das variáveis latentes do *autoencoder* pelo SVM, foi realizado uma nova normalização dos dados, desta vez para uma escala entre -1 e 1. Como as variáveis latentes extraídas pelo autoencoder podem variar em relação à escala de valores, fazer uma nova normalização antes de utilizar esses dados se tornou necessário.

Tabela 5.1: Configuração do *fair autoencoder* para cada base de dados, mostrando o intervalo de valores utilizado pelo *bayesian search* para se determinar a taxa de aprendizado (*learning rate*) que melhor reduzissem o erro entre a saída da rede e os dados originais. Bem como outros valores utilizados como tamanho dos *batches* e quantidade de neurônios em cada camada.

Configuração do Autoencoder					
Base de dados	*Learning rate*	Batch	Épocas	encoder/ decoder	Camada escondida
Adult Income	$(10^{-2}, 1)$	200	30	80	50
German Credit (Gênero)	$(10^{-3}, 1)$	10	40	60	40
German Credit (Idade)	$(10^{-3}, 1)$	10	40	60	40

A Tabela 5.1 apresenta as configurações do *autoencoder* utilizadas para cada base de dados. No primeiro experimento, foi utilizado o método de *bayesian search*. Esse método emprega uma busca probabilística para encontrar os parâmetros que melhor otimizem uma métrica específica para modelos de aprendizagem. No caso do *autoencoder* proposto, o objetivo foi encontrar a melhor taxa de aprendizado (*learning rate*) que otimiza-se o erro entre a saída da rede com e os dados originais. O intervalo de *learning rate* presente na Tabela 5.1 representa um intervalo contínuo utilizado pelo *bayesian search*.

Os parâmetros fixos, como a quantidade de épocas, o número de neurônios nas camadas internas do *autoencoder* e o tamanho do *batch*, foram baseados no trabalho de LOUIZOS *et al.* (2016), que também utilizou as bases *Adult Income* e *German Credit* no contexto de *fairness* e um modelo de *autoencoder*, assim como alguns testes manuais de valores e aqueles que obtiveram melhores resultados foram escolhidos.

Para *cross validation*, utilizamos o método *K-Fold* com $k = 3$. Optamos por um k pequeno devido ao desbalanceamento na quantidade de indivíduos dos diferentes grupos sensíveis em ambas as bases de dados. Dessa forma, aumentamos as chances

de que cada conjunto de validação seja mais representativo do conjunto original em relação aos atributos sensíveis.

Tabela 5.2: Configurações do SVM utilizado após o processo de extração de variáveis latentes pelo *fair autoencoder*, mostrando o *kernel* utilizado, e os intervalos para escolhidos para os hiper-parâmetros C e *gamma* que melhor otimizassem o MCC final.

Configuração do SVM			
Base de dados	*Kernel*	C	gamma
Adult Income	rbf	$(10^{-3}, 1)$	$(10^{-3}, 1)$
German Credit - Gênero	rbf	$(10^{-3}, 1)$	$(10^{-3}, 1)$
German Credit - Idade	rbf	$(10^{-3}, 1)$	$(10^{-3}, 1)$

Na Tabela 5.2, são apresentadas as configurações do SVM para cada base de dados, incluindo o *kernel* e os parâmetros C e *gamma*. Para o SVM, foi utilizado o método *random search* de busca de parâmetros. A função escolhida para ser otimizada foi o MCC, a ideia por trás dessa escolha foi buscar aprender parâmetros que conseguissem determinar a relação entre o conjunto mais justo gerado pelo *fair autoencoder* e um bom desempenho de precisão na classificação. Assim como nos hiper-parâmetros ajustados no *autoencoder*, os intervalos para os parâmetros C e *gamma* na Tabela 5.2 representam valores contínuos que foram utilizados pelo método de busca.

O *kernel* RBF (*Radial Basis Function*) foi escolhido para realizar uma transformação não linear dos dados que se deseja classificar. Como a função de covariância, empregado na função de custo escolhida pelo *autoencoder*, é uma função que determina a relação linear entre duas variáveis, utilizar um *kernel* não linear é um bom teste para o método proposto, no que diz respeito à remoção de parte da informação sensível.

Para as duas bases de dados, a função de ativação utilizada nos *encoders* e *decoders* foi a ReLU (*Rectified Linear Unit*). Essa função é definida como $f(x) = \max(0, x)$ e é amplamente utilizada em redes neurais com mais de uma camada interna, devido à sua simplicidade, cálculo eficiente de gradientes e prevenção de saturação dos valores. Após a normalização dos dados originais para o autoencoder, os valores da entrata ficaram entre 0 e 1, por esse motivo a função de ativação escolhida para a saída da rede foi a *sigmoid*. A função *sigmoid* é definida como $S(x) = \frac{1}{1+e^{-x}}$, e produz valores entre 0 e 1. Com isso, ela se encaixa bem para reconstruir os dados na mesma escala.

Os intervalos escolhidos para os hiper-parâmetros do *fair autoencoder* e do SVM basearam-se em outros trabalhos que utilizam modelos similares, como os apresen-

tados por LOUIZOS *et al.* (2016); ZEMEL *et al.* (2013); MAKHZANI e FREY (2013); PAN *et al.* (2020); ZHANG (2018), e foram ajustados conforme as experimentações realizadas. Limites superiores maiores foram testados; no entanto, com esses valores, os modelos tendiam a classificar todo o conjunto de teste em uma única classe. Por isso, valores menores foram testados até que se obtivessem resultados mais equilibrados.

Com a utilização dos métodos de *Bayesian Search* e *Random Search*, que buscaram otimizar a performance dos modelos nas diferentes etapas do processo, encontrar valores adequados para o parâmetro λ foi um desafio, tanto pelo custo computacional de testar diferentes parâmetros quanto pela contradição em relação ao efeito que alguns valores de λ teriam no resultado. Afinal, quanto maior o valor de λ, maior será a perda nas outras métricas por consequência. O desafio, portanto, é encontrar valores de λ que ainda possam extrair conhecimento das bases de dados e ao mesmo tempo tornem o conjunto mais justo.

5.3 Resultados

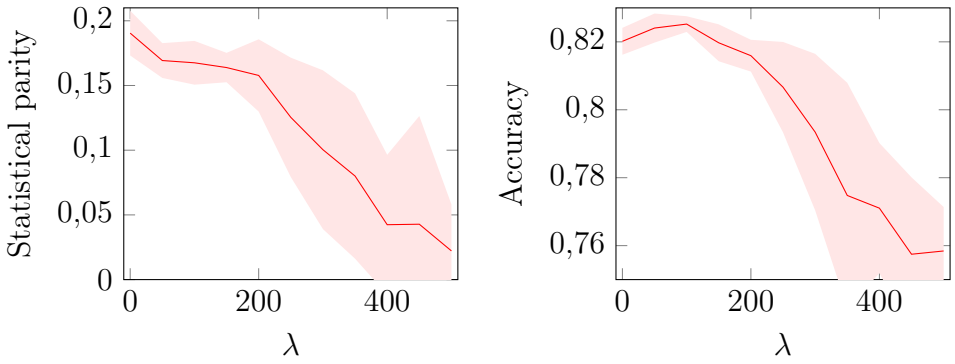
Neste seção é apresentado os resultados dos experimentos realizar para o método proposto, bem como a conclusão e análise de cada um deles.

Para o primeiro experimento, buscamos determinar a relação entre o parâmetro λ da *Covariance Loss Function* e as métricas escolhidas para validar a proposta, assim como determinar os melhores hiper-parâmetros para cada valor de λ em cada etapa. Para isso, testamos um conjunto de valores discretos do parâmetro λ, que podem ser vistos na Tabela 5.3. Uma vez encontrados esses parâmetros com os métodos de busca citados na seção anterior, para cada valor de λ, foram realizadas 10 repetições do processo de transformação da base e, posteriormente, a classificação pelo SVM. Em cada repetição, uma nova amostra dos conjuntos de treino e teste foi gerada. Após as 10 repetições, foi calculada a média das métricas de *statistical parity*, acurácia, F1 e MCC para o conjunto de teste.

Tabela 5.3: Configuração dos intervalos de λ utilizados por cada base de dados e os respectivos atributos sensível considerados

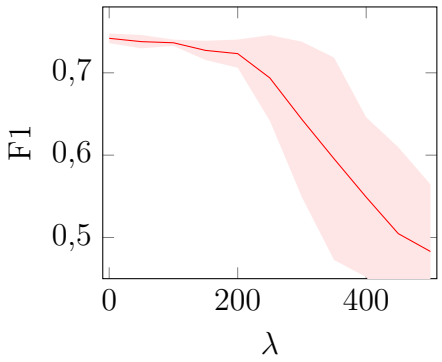
Configuração dos intervalos do λ	
Bases de dados	Intervalos de λ
Adult Income	[0, 50, 100, ..., 500]
German Credit - Gênero	[0, 25, 50, ..., 200]
German Credit - Idade	[0, 25, 50, ..., 300]

Na Figura 5.3, é apresentada a relação entre o parâmetro λ e as métricas do SVM do conjunto de teste da base de dados *Adult Income*. É perceptível que, quanto maior

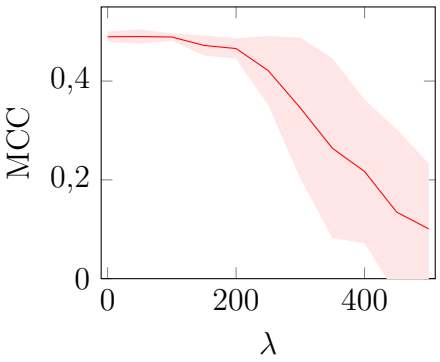


(a) Adult Income - Statistical parity

(b) Adult Income - Acurácia



(c) Adult Income - F1 score



(d) Adult Income - MCC

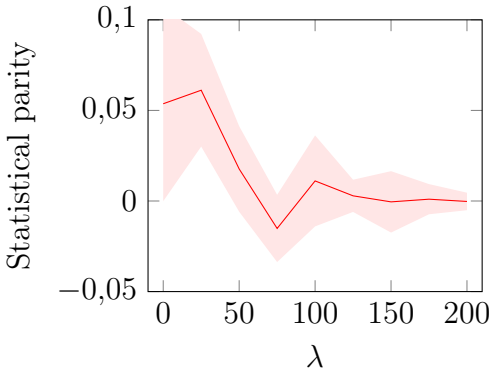
Figura 5.3: (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC, para diferentes valores de λ no conjunto de teste da base *Adult Income* após a classificação pelo SVM.

o valor de λ, todas as métricas sofrem uma queda. No caso do *statistical parity*, essa queda indica uma maior equidade entre os grupos protegidos e não protegidos. No entanto, a partir de um determinado valor, as métricas, principalmente F1 e MCC, começam a declinar a ponto de tornar os dados menos úteis para outros modelos.

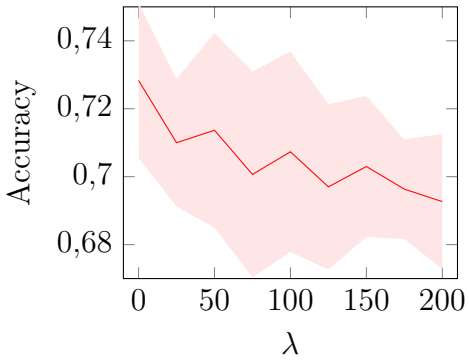
Já na Figura 5.4, apresentamos a relação entre o parâmetro λ e as métricas do conjunto de teste da base *German Credit*, considerando o gênero como atributo sensível, após aplicado a classificação pelo SVM. Embora essa base seja bem menor e portanto mais instável em relação as métricas, ainda assim, o método é capaz de controlar em algum grau o *statistical parity* em detrimento de uma queda nas outras métricas.

Por outro lado, na Figura 5.5, onde variamos o valor de λ para a base *German Credit* considerando a idade como atributo sensível, também observamos uma queda em todas as métricas. No entanto, essa queda é menos significativa se comparada à mesma base quando o gênero é utilizado como *feature* sensível. Isso pode reforçar a ideia de que o atributo idade está menos correlacionado com os outros atributos e, portanto, é mais tolerante a variações para valores maiores de λ.

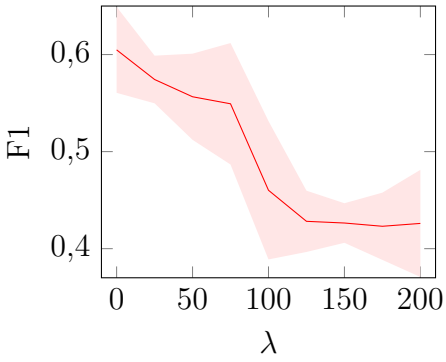
A partir dos resultados apresentados nas Figuras 5.3, 5.4 e 5.5, a Tabela 5.4



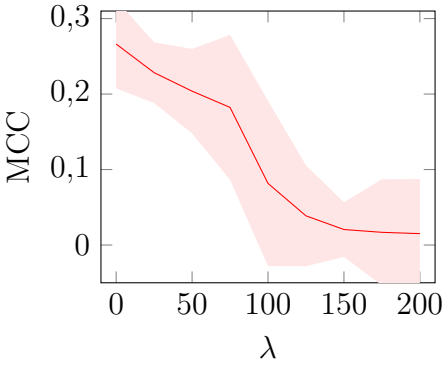
(a) German Credit (Gênero) - Statistical parity



(b) German Credit (Gênero) - Acurácia



(c) German Credit (Gênero) - F1 score



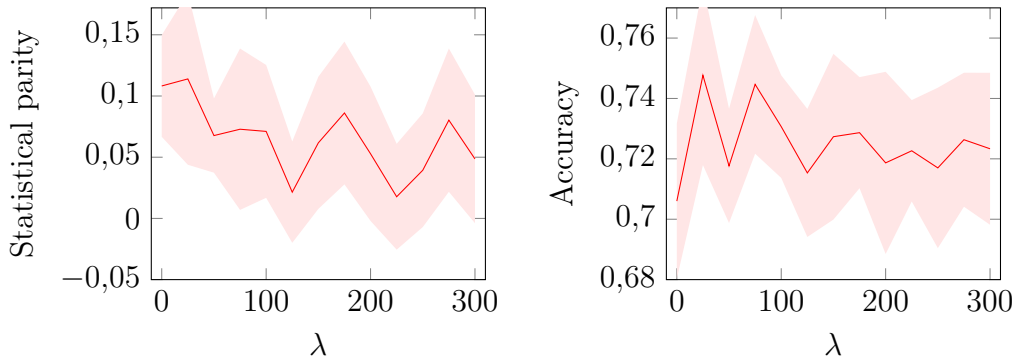
(d) German Credit (Gênero) - MCC

Figura 5.4: (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC para diferentes valores de λ no conjunto de teste da base *German Credit* considerando gênero como atributo sensível, após a classificação pelo SVM.

mostra os resultados para os conjuntos de teste das bases de dados com valores fixos do parâmetro λ. Para comparar os resultados entre o método proposto (CFA) e um modelo *base line* (Regular) que ignora a penalidade da *covariance loss function* do *fair autoencoder*, ou seja, $\lambda = 0$, incluímos também os resultados de uma abordagem que faz uma simples remoção do atributo sensível, de forma a comparar esses resultados com o método proposto.

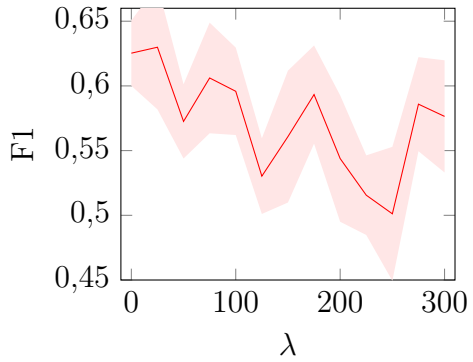
Não encontramos trabalhos que lidem diretamente com o *trade off* entre o *statistical parity* e outras métricas relacionadas à taxa de acerto e precisão de maneira a ter um parâmetro controlando essa relação tão diretamente. Para encontrar um equilíbrio entre a redução do *statistical parity* e a manutenção das métricas, escolhemos valores para o parâmetro λ que tolerassem uma perda de até 30% do MCC em relação ao *base line* chamado "Regular"na Tabela 5.4. Como o MCC é a principal métrica relacionada à qualidade da classificação, decidimos limitar o nível de redução por essa métrica nos demais experimentos.

Para todas as implementações da Tabela 5.4, utilizamos o processo de transformação pelo autoencoder e, posteriormente, utilizamos o SVM para fazer a classifi-

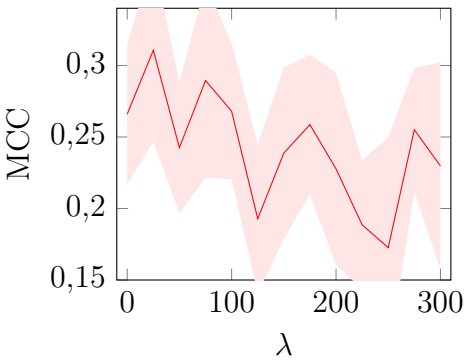


(a) German Credit (Idade) - Statistical parity

(b) German Credit (Idade) - Acurácia



(c) German Credit (Idade) - F1 score



(d) German Credit (Idade) - MCC

Figura 5.5: (a) *Statistical parity*, (b) acurácia, (c) F1 e (d) MCC para diferentes valores de λ para o conjunto de teste da base *German Credit* considerando idade como atributo sensível, após a classificação pelo SVM.

cação. Dessa forma, a única diferença entre as implementações está na aplicação da *covariance loss function* através do parâmetro λ. Em todos os casos, o método proposto foi capaz de diminuir o *statistical parity*, deixando esta métrica mais próxima de zero, ao custo de uma redução em acurácia, F1 e MCC, que é um *trade-off* conhecido na área.

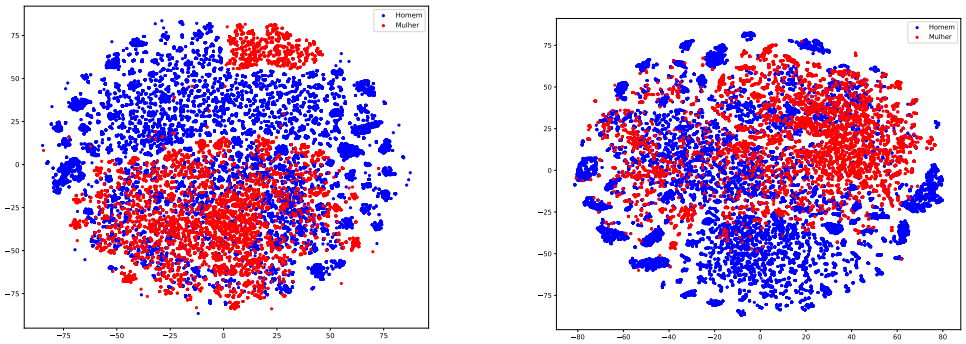
É interessante observar, na Tabela 5.4, que a simples remoção do atributo sensível não gerou impacto significativo nas métricas na maioria dos cenários, reforçando a ideia do *redlining effect*, em que a informação do atributo sensível está implicitamente presente nos outros atributos. Apenas na base *German Credit*, considerando a idade como atributo sensível, a simples remoção do atributo idade provocou uma diminuição do *statistical parity*. Isso indica que o *redlining effect* neste caso não se faz tão presente para este atributo. Mesmo assim, o método proposto foi capaz de atingir um nível próximo de *statistical parity* na média das repetições.

Outro comparativo realizado utilizou o t-SNE para visualizar a distribuição dos dados em duas dimensões. Esta comparação foi utilizada em outros trabalhos sobre *fairness*, como o proposto por LOUIZOS *et al.* (2016). Dado que o t-SNE é uma técnica que garante a proximidade espacial entre indivíduos similares, ao comparar

Tabela 5.4: *Statistical parity*, acurácia, F1 score e MCC para as base de dados: Adult Income e German Credit usando gênero; gênero e idade respectivamente, para diferentes abordagens

Adult Income			
Métricas	Regular ($\lambda = 0$)	Sem Atributo Sensível	CFA $\lambda = 3 \times 10^2$
Statistical Parity	0,19	0,17	**0,1**
Acurácia	**0,82**	0,81	0,79
F1	**0,74**	0,73	0,64
MCC	**0,49**	0,48	0,34
German Credit - Gênero			
Métricas	Regular ($\lambda = 0$)	Sem Atributo Sensível	CFA $\lambda = 50$
Statistical Parity	0,05	0,05	**0,01**
Acurácia	0,73	**0,74**	0,71
F1	0,6	**0,62**	0,55
MCC	0,26	**0,3**	0,2
German Credit - Idade			
Métricas	Regular ($\lambda = 0$)	Sem Atributo Sensível	CFA $\lambda = 50$
Statistical Parity	0,1	**0,05**	0,06
Acurácia	0,7	**0,74**	0,71
F1	**0,62**	**0,62**	0,57
MCC	0,26	**0,3**	0,24

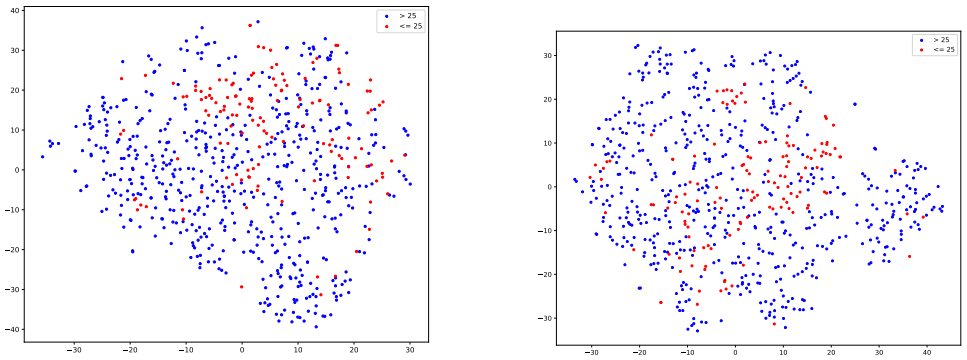
os dados originais com o *output* do *fair autoencoder* proposto destacando o atributo sensível, é possível avaliar o impacto desse atributo na distribuição dos dados.



(a) (b)

Figura 5.6: (a) t-SNE dos dados originais da base *Adult Income* considerando gênero como atributo sensível (b) t-SNE dos dados reconstruídos da base *Adult Income* considerando gênero como atributo sensível e $\lambda = 3 \times 10^2$.

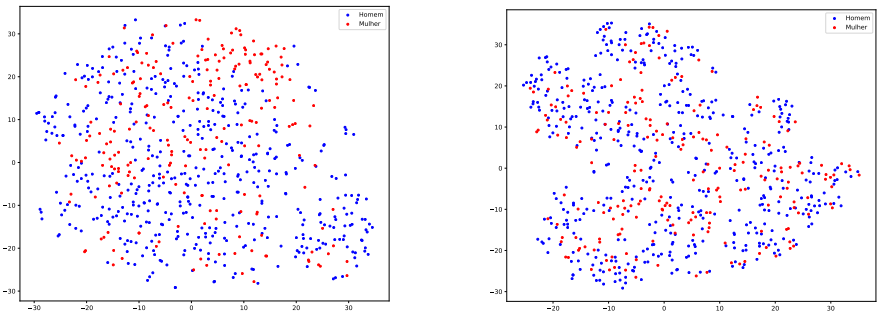
A Figura 5.6 apresenta um comparativo utilizando o t-SNE, dos dados de treino, entre o conjunto original e a saída do *fair autoencoder* da base *Adult Income*. É perceptível que, nos dados originais, as mulheres (representadas em vermelho) estão agrupadas em áreas específicas da visualização. Em contrapartida, nos dados transformados, há um espalhamento maior de ambos os grupos, indicando uma heterogeneidade na distribuição em relação ao atributo sensível. Isso sugere que essa informação influencia menos na similaridade entre homens e mulheres na base.



(a) (b)

Figura 5.7: (a) t-SNE dos dados originais da base *German Credit* considerando idade como atributo sensível (b) t-SNE dos dados reconstruídos da base *German Credit* considerando idade como atributo sensível e $\lambda = 50$.

Já na Figura 5.7, apresentamos o comparativo do conjunto de treino usando o t-SNE da base de dados *German Credit*, onde a idade é o atributo sensível. Nesse cenário, a distribuição do grupo protegido é menos perceptível por conta da menor quantidade de indivíduos na base, outro sinal disso pode estar relacionado com a menor interdependência da idade com as outras *features*, característica destacada no primeiro experimento com a simples remoção da coluna idade desta base.



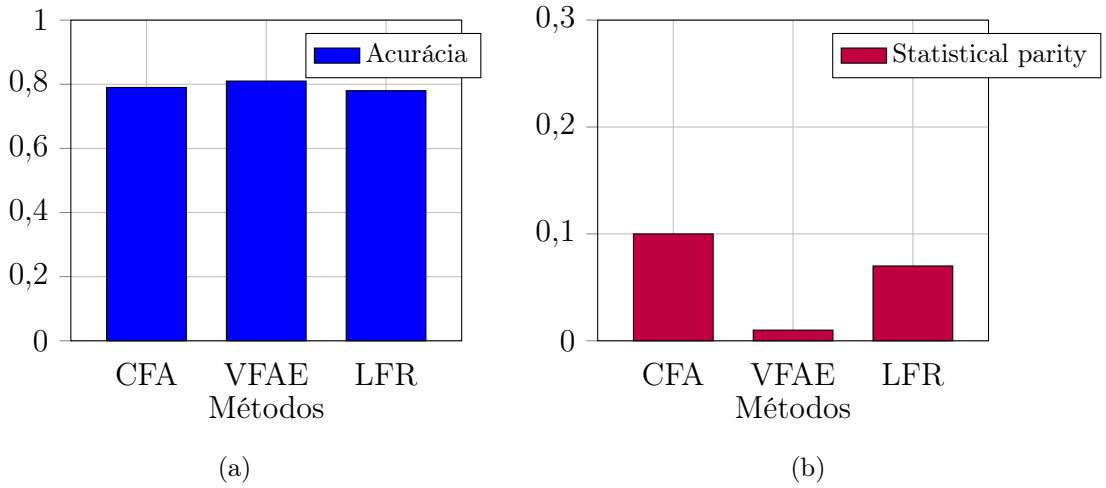
(a) (b)

Figura 5.8: (a) t-SNE dos dados originais da base *German Credit* considerando gênero como atributo sensível (b) t-SNE dos dados reconstruídos da base *German Credit* considerando gênero como atributo sensível e $\lambda = 50$.

Na Figura 5.8, ao analisar o gênero como atributo sensível no t-SNE da base *German Credit*, percebe-se uma distribuição equilibrada dos indivíduos entre os grupos protegido e não protegido. Esse resultado sugere que o método proposto conseguiu minimizar o impacto dessa informação, mesmo com uma quantidade limitada de entradas.

As bases *Adult Income* e *German Credit* são frequentemente adotadas em estudos sobre *fairness*. Isso se deve à presença de diferentes atributos sensíveis, à variedade de tipos de *features* e ao fato de serem bases com informações de pessoas reais. Essas características fazem dessas bases opções ideais para a exploração de técnicas e métricas de *fairness*.

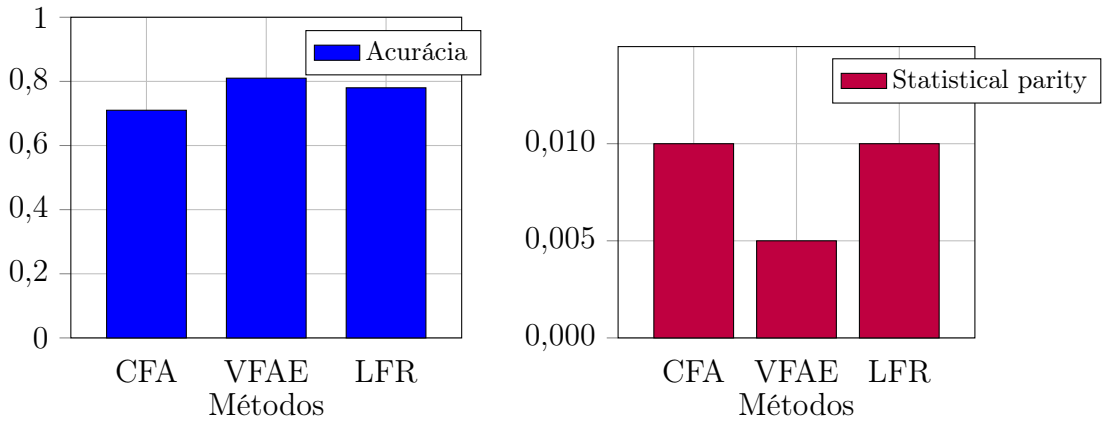
Para contrapor nossa proposta em relação a outros trabalhos, realizamos comparações com duas abordagens notáveis. A primeira, denominada *Variational Fair Autoencoder* (VFAE), foi apresentada por LOUIZOS *et al.* (2016) e utiliza um *variational autoencoder* para obter uma representação justa dos dados. A segunda, chamada *Learning Fair Representation* (LFR), busca uma representação de dados mais justa, empregando uma distribuição de probabilidade que remapeia os dados para um conjunto mais equitativo. Para essa comparação, o CFA foi configurado com o λ dos melhores resultados obtidos para cada base presentes na Tabela 5.4.



(a) (b)

Figura 5.9: (a) Comparativo de acurácia entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). (b) Comparativo de *statisitcal parity* entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (Learning Fair Representation).

Na Figura 5.9, apresentamos um comparativo da base *Adult Income* com outros dois métodos em relação à acurácia e *statistical parity*. O comparativo mostra que nosso método conseguiu se equiparar em termos de acurácia e obteve um bom resultado na diminuição do *statistical parity*. Vale ressaltar que os valores dos outros métodos foram estimados com base nos gráficos apresentados nos respectivos artigos.



(a) (b)

Figura 5.10: (a) Comparativo de acurácia entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). (b) Comparativo de *statisitcal parity* entre o modelo proposto CFA com outros dois modelos da literatura VFAE (*Variational Fair Autoencoder*) e LFR (*Learning Fair Representation*). Considerando o gênero como atributo sensível.

Já na Figura 5.10, apresentamos um comparativo da base *German Credit*, utilizando o gênero como atributo sensível, entre o CFA com outros dois métodos da literatura. Nesse cenário, nosso método também se mostrou equiparável em termos de acurácia e *statistical parity*. Assim como no primeiro comparativo, os valores dos métodos VFAE e LFR foram estimados.

Com os experimentos realizados, conseguimos demonstrar que o nosso método é capaz de diminuir a influência do atributo sensível, reduzindo o *statistical parity* em detrimento de uma redução nas métrica de acurácia, F1 *score* e MCC, mantendo um controle sobre o grau de justiça desejado com o parâmetro λ. Outro resultado importante é a capacidade do método proposto de atenuar o *redlining effect* do atributo sensível em dados que apresentam viés.

Capítulo 6

Conclusões

Neste capítulo será apresentado as conclusões deste trabalho, bom como suas limitações, contribuições e possíveis trabalhos futuros.

6.1 Objetivo do trabalho

Esta dissertação tem como objetivo apresentar um método de *fair representation* capaz de lidar com bases de dados que contenham viés de *fairness* relacionado a um grupo de indivíduos historicamente desfavorecido na sociedade.

Considerando o problema de *fairness*, no qual membros de um grupo específico têm menor probabilidade de pertencer à classe positiva devido a discriminações e preconceitos da sociedade refletidos nos dados, a abordagem proposta visa transformar os dados originais em uma representação mais justa. Isso é feito mediante a remoção parcial da informação do atributo sensível na nova representação.

Para tanto, o presente trabalho emprega um *autoencoder* com uma função de custo contendo um termo que visa promover a justiça dos dados. Dado que o *autoencoder* é capaz de extrair características latentes do conjunto de entrada, a inclusão desse termo resulta na geração de uma representação com menor viés em relação ao atributo sensível, quando comparada aos dados originais.

A desvinculação do atributo sensível ocorre por meio da inclusão, na função de custo, da covariância entre a informação sensível original e a representação gerada pelo *autoencoder*. Devido à natureza do *autoencoder*, essas informações tendem a estar correlacionadas, e a introdução desse termo na função de custo a ser minimizada permite reduzir a influência dessa *feature*.

6.2 Contribuições

A principal contribuição deste trabalho reside na apresentação de uma abordagem de *fair representation* por meio de aprendizado não supervisionado, capaz de aprender uma representação mais equitativa dos dados. Com a hipótese de que é possível tornar os dados transformados linearmente independente da base original, uma série de experimentos foram conduzidos para validar e mensurar tal conjectura.

Os resultados experimentais corroboraram a viabilidade de transformar uma base de dados enviesada em um conjunto mais justo com o método proposto, ao mesmo tempo que preserva informações suficientes para viabilizar a extração de conhecimento por outros modelos de *machine learning*. As principais métricas utilizadas englobaram *statistical parity*, acurácia, F1 e MCC. Observou-se que a redução de *statistical parity* foi alcançada ao empregar os dados mais justos no modelo SVM, entretanto, tal ganho veio acompanhado de uma diminuição nas métricas de desempenho.

Além disso, pela inclusão de um parâmetro λ no termo da função de custo do *fair autoencoder*, que busca desassociar as variáveis latentes do atributo sensível, o método foi capaz de controlar o nível de *fairness* desejado. Assim, para valores maiores desse parâmetro, conseguimos atingir níveis menores de injustiça.

6.3 Limitações e possíveis trabalhos futuros

Uma das limitações intrínsecas à proposta deste estudo é a adoção da função de covariância para efetuar a supressão do atributo sensível da base original. Essa medida estatística, se restringe a determinar a dependência linear entre duas variáveis, por tanto, no que tange ao tratamento de relações não lineares entre os atributos, o presente trabalho não é capaz de garantir uma justiça aprimorada no processamento dos dados.

Diante dessa circunstância, uma perspectiva promissora para trabalhos futuros consiste em incorporar uma função de custo que considere relações não lineares, o que pode alavancar significativamente os resultados. Alternativamente, contemplar o emprego de funções de correlação, como o cosseno ou o índice de *Pearson*, emerge como uma estratégia viável para a mitigação do atributo sensível. Ressalta-se que tais abordagens podem revelar-se mais eficazes ao tratar de bases de dados que possuam características específicas.

Outra característica do método proposto é a capacidade natural de lidar com mais de um atributo sensível ao mesmo tempo, sem precisar de modificações estruturais na proposta. Portanto, tentar remover mais de um atributo sensível de uma mesma base de dados mostra-se como uma boa abordagem a ser testada e validada

em sequência. Além disso, realizar experimentos com outras bases de dados, possuindo diferentes características, também se apresenta como uma boa oportunidade para validar a capacidade e limitações do *covariance fair autoencoder*.

Referências Bibliográficas

FOULDS, J. R., ISLAM, R., KEYA, K. N., et al. "An intersectional definition of fairness", *Proceedings - International Conference on Data Engineering*, v. 2020-April, pp. 1918–1921, 2020. ISSN: 10844627. doi: 10.1109/ICDE48307.2020.00203.

LE QUY, T., ROY, A., IOSIFIDIS, V., et al. "A survey on datasets for fairness-aware machine learning", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 12, n. 3, pp. e1452, 2022.

YUCER, S., AKÇAY, S., AL-MOUBAYED, N., et al. "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–19, 2020.

MEHRABI, N., MORSTATTER, F., SAXENA, N., et al. "A survey on bias and fairness in machine learning", *ACM computing surveys (CSUR)*, v. 54, n. 6, pp. 1–35, 2021.

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G., et al. "Fairness Constraints: Mechanisms for Fair Classification", *20th International Conference on Artificial Intelligence and Statistics*, v. 54, 2015. ISSN: 15523098. doi: 10.1109/TRO.2009.2019886. Disponível em: <`http://arxiv.org/abs/1507.05259`>.

PESSACH, D., SHMUELI, E. "Algorithmic fairness", *arXiv preprint arXiv:2001.09784*, 2020.

VERMA, S., RUBIN, J. "Fairness definitions explained". In: *Proceedings of the international workshop on software fairness*, pp. 1–7, 2018.

KAMISHIMA, T., AKAHO, S., ASOH, H., et al. "Fairness-aware classifier with prejudice remover regularizer". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pp. 35–50. Springer, 2012.

KAMISHIMA, T., AKAHO, S., SAKUMA, J. "Fairness-aware Learning through Regularization Approach". In: *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650, 2011. doi: 10.1109/ICDMW.2011.83.

PEDRESHI, D., RUGGIERI, S., TURINI, F. "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568, 2008.

LOUIZOS, C., SWERSKY, K., LI, Y., et al. "The variational fair autoencoder", *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pp. 1–11, 2016.

KENFACK, P. J., KHAN, A. M., HUSSAIN, R., et al. "Adversarial Stacked Auto-Encoders for Fair Representation Learning", 2021. Disponível em: <`http://arxiv.org/abs/2107.12826`>.

CUNNINGHAM, P., CORD, M., DELANY, S. J. "Supervised learning". In: *Machine learning techniques for multimedia: case studies on organization and retrieval*, Springer, pp. 21–49, 2008.

GHAHRAMANI, Z. "Unsupervised learning". In: *Summer school on machine learning*, Springer, pp. 72–112, 2003.

LI, Y. "Deep reinforcement learning: An overview", *arXiv preprint arXiv:1701.07274*, 2017.

KRIESEL, D. "Neural Networks", p. 286, 2013. Disponível em: <`http://www.dkriesel.com/en/science/neural_networks`>.

NIELSEN, M. "Using neural nets to recognize handwritten digits", *Neural networks and deep learning*, pp. 1–75, 2015.

GÜNTHER, F., FRITSCH, S. "Neuralnet: training of neural networks." *R J.*, v. 2, n. 1, pp. 30, 2010.

BALDI, P. "Autoencoders, unsupervised learning, and deep architectures". In: *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49. JMLR Workshop and Conference Proceedings, 2012.

ZHANG, G., LIU, Y., JIN, X. "A survey of autoencoder-based recommender systems", *Frontiers of Computer Science*, v. 14, pp. 430–450, 2020.

JING, L., ZBONTAR, J., OTHERS. "Implicit rank-minimizing autoencoder", *Advances in Neural Information Processing Systems*, v. 33, pp. 14736–14746, 2020.

YU, W., ZENG, G., LUO, P., et al. "Embedding with autoencoder regularization". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 208–223. Springer, 2013.

ZHAO, J., KIM, Y., ZHANG, K., et al. "Adversarially regularized autoencoders". In: *International conference on machine learning*, pp. 5902–5911. PMLR, 2018.

BURGES, C. J. C. "A Tutorial on Support Vector Machines for Pattern Recognition", *Data mining and knowledge discovery*, v. 2, pp. 121–167, 1998. ISSN: 13845810. doi: 10.1023/A:1009715923555. Disponível em: <http://link.springer.com/10.1023/A:1009715923555%5Cnpapers3://publication/doi/10.1023/A:1009715923555>.

DIETTERICH, T. G., KONG, E. B. "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms", 1995.

HELLSTRÖM, T., DIGNUM, V., BENSCH, S. "Bias in Machine Learning–What is it Good for?" *arXiv preprint arXiv:2004.00686*, 2020.

ZEMEL, R., WU, Y., SWERSKY, K., et al. "Learning fair representations", *30th International Conference on Machine Learning, ICML 2013*, v. 28, n. PART 2, pp. 1362–1370, 2013.

BISHOP, C. M. "Latent variable models". In: *Learning in graphical models*, Springer, pp. 371–403, 1998.

SATTIGERI, P., HOFFMAN, S. C., CHENTHAMARAKSHAN, V., et al. "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network", *IBM Journal of Research and Development*, v. 63, n. 4/5, pp. 3:1–3:9, 2019. doi: 10.1147/JRD.2019.2945519.

MADRAS, D., CREAGER, E., PITASSI, T., et al. "Learning adversarially fair and transferable representations". In: *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

RICE, J. A. *Mathematical statistics and data analysis*. Belmont, CA, Cengage Learning, 2006.

CHOI, Y., FARNADI, G., BABAKI, B., et al. "Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, pp. 10077–10084, 2020.

IOSIFIDIS, V., NTOUTSI, E. "Dealing with bias via data augmentation in supervised learning scenarios", *Jo Bates Paul D. Clough Robert Jäschke*, v. 24, pp. 11, 2018.

ZAFAR, M. B., VALERA, I., ROGRIGUEZ, M. G., et al. "Fairness constraints: Mechanisms for fair classification". In: *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

MCDONOUGH, P., DUNCAN, G. J., WILLIAMS, D., et al. "Income dynamics and adult mortality in the United States, 1972 through 1989." *American journal of public health*, v. 87, n. 9, pp. 1476–1483, 1997.

VAN DER MAATEN, L., HINTON, G. E. "Visualizing Data using t-SNE", *Journal of Machine Learning Research 9*, v. 219, pp. 187–202, 2008. ISSN: 15729338. doi: 10.1007/s10479-011-0841-3.

MAKHZANI, A., FREY, B. "K-sparse autoencoders", *arXiv preprint arXiv:1312.5663*, 2013.

PAN, H., TANG, W., XU, J.-J., et al. "Rolling bearing fault diagnosis based on stacked autoencoder network with dynamic learning rate", *Advances in Materials Science and Engineering*, v. 2020, pp. 1–12, 2020.

ZHANG, Y. "A better autoencoder for image: Convolutional autoencoder". In: *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. pdf (accessed on 23 March 2017)*, 2018.