Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

A utilização de técnicas de classificação aplicadas ao perfilamento de trabalhadores do Sistema Nacional de Emprego: uma abordagem de aprendizado de máquina

Amilton Lobo Mendes Júnior

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

LM538u

Lobo Mendes Júnior, Amilton
A utilização de técnicas de classificação aplicadas ao perfilamento de trabalhadores do Sistema Nacional de Emprego: uma abordagem de aprendizado de máquina / Amilton Lobo Mendes Júnior; orientador Gladston Luiz da Silva. -- Brasília, 2023.
59 p.

Dissertação(Mestrado Profissional em Computação Aplicada) - Universidade de Brasília, 2023.

1. Aprendizado de máquina. 2. Sistema público de emprego. 3. Perfilamento de trabalhadores. 4. Aprendizado supervisionado. I. Luiz da Silva, Gladston, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

A utilização de técnicas de classificação aplicadas ao perfilamento de trabalhadores do Sistema Nacional de Emprego: uma abordagem de aprendizado de máquina

Amilton Lobo Mendes Júnior

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Gladston Luiz da Silva (Orientador)
CIC/UnB

Prof.a Dr.a Maristela Terto de Holanda
CIC/UnB

Prof. Dr. Francisco José da Silva e Silva
Universidade Federal do Maranhão

Prof. Dr. Marcelo Ladeira
CIC/UnB

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 20 de julho de 2023

Dedicatória

Dedico este trabalho primeiramente à minha mãe, Flor de Maria Pires Mendes, que sempre será, direta ou indiretamente, minha luz guia em relação a tudo que sou.

Em segundo lugar, dedico aos meus filhos, Gabriel e Ana Laura, que me motivam a querer ser sempre melhor, e à minha eterna companheira, Patrícia, que sempre me mostra que juntos podemos ser mais.

Por fim, dedico ao meu pai, que, à sua maneira, sempre lutou por nós, e a minhas irmãs, sobrinho, sobrinhas e cunhados. Vocês me inspiram e fazem com que eu queira ser inspiração para vocês.

Agradecimentos

Agradeço primeiramente a Deus, por mais esta bênção em minha vida.

Agradeço a minha família. Em vários momentos em que as dificuldades pareciam ser maiores que minha capacidade, palavras carinhosas de cobrança ou comentários implícitos sobre o orgulho que sentiam por mim não permitiram que eu desistisse.

Não tenho palavras para expressar o agradecimento ao Prof. Dr. Gladston Luiz da Silva, cuja dedicação e experiência fizeram com que eu, entre outras coisas, voltasse para a realidade, me ajudando a delimitar o escopo do trabalho. Só assim pude alcançar a esse objetivo do qual tenho tanto orgulho.

Agradeço a todos os professores do Programa de Pós-graduação em Computação Aplicada (PPCA). A superação de todas as turbulências pelas quais passamos durante o período de pandemia do Covid-19 não seria possível sem a participação de vocês.

Agradeço aos meus colegas servidores públicos e colaboradores os quais mantêm viva a vocação de servir. Vocês são uma das minhas principais inspirações.

Por fim, agradeço aos colegas e funcionários do Programa de Pós-graduação em Computação Aplicada.

Resumo

O presente trabalho propõe a utilização de técnicas de aprendizado de máquina na atividade do perfilamento de trabalhadores do sistema público de emprego brasileiro, o Sistema Nacional de Emprego - Sine. A utilização de um mecanismo automatizado de perfilamento de trabalhadores permitirá que esforços sejam direcionados para o tratamento preventivo de trabalhadores mais propensos a permanecerem por mais tempo fora do mercado formal de trabalho. Esse tratamento antecipado poderá contribuir com a antecipação do retorno dos trabalhadores ao mercado de trabalho formal, com o potencial resultado de reduzir os gastos com o seguro-desemprego, os quais foram em 2022 em torno de 35 bilhões de reais. A área sob a curva (AUC) foi escolhida como métrica para a avaliação dos modelos *Logistic Regression* (LR), *Gradient Boosting Machines* (GBM), *Extreme Gradient Boosting* (XGBoost) e uma Rede Neural com a utilização de um *embedding* de códigos da Classificação Brasileira de Ocupações(CBO). No experimento, o modelo com os melhores resultados apresentados foi o XGBoost. Melhorias futuras podem incluir a adição de variáveis relacionadas ao mercado local de trabalho e a transições de setores da economia, de ocupações e de residência.

Palavras-chave: Aprendizado de máquina, Sistema público de emprego, Perfilamento de trabalhadores, Aprendizado supervisionado

Abstract

The present work proposes the application of machine learning techniques in the profiling of workers of brazilian public employment system, the National Employment System - NES (Sine in portuguese). The use of an automated worker profiling mechanism will allow efforts to be directed towards the preventive treatment of workers who are more likely to remain longer outside the formal labor market. This early treatment could contribute to bringing workers back to the formal job market earlier, with the potential result of reducing unemployment insurance expenses, which in 2022 were around 35 billion reais. The area under the curve (AUC) was chosen as a metric for evaluating the models *Logistic Regression* (LR), *Gradient Boosting Machines* (GBM), *Extreme Gradient Boosting* (XGBoost) and a Neural Network using an *embedding* of codes from the Brazilian Classification of Occupations (CBO). In the experiment, the model with the best results was XGBoost. Future improvements may include the addition of variables related to the local labor market and transitions in sectors of the economy, occupations and residence.

Keywords: Machine learning, Public employment system, Worker profiling, Supervised learning

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AUC *Area Under the Curve.*

Caged Cadastro Geral de Empregados e Desempregados.

CBO Classsificação Brasileira de Ocupações.

Cnae Código Nacional de Atividade Econômica.

CPF Cadastro de Pessoa Física.

CRISP-DM *Cross-Industry Standard Process for Data Mining.*

eSocial Sistema de Escrituração Digital das Obrigações Fiscais, Previdenciárias e Trabalhista.

GBM *Gradient Boosting Machines.*

KDD *Knowledge Discovery in Databases.*

Lime *Local Interpretable Model-agnostic Explanations.*

LR *Logistic Regression.*

MTP Ministério do Trabalho e Previdência.

OCDE Organização para a Cooperação e Desenvolvimento Econômico.

Rais Relação Anual de Informações Sociais.

ROC *Receiver Operating Characteristic.*

Shap *SHapley Additive exPlanations.*

Sine Sistema Nacional de Emprego.

SPE Serviços Públicos de Emprego.

WPRS *Worker Profiling and Reemployment Services.*

XGBoost *eXtreme Gradient Boosting.*

Capítulo 1

Introdução

1.1 Contextualização

O mercado de trabalho pode ser definido como o conjunto de processos por meio do qual vagas de emprego disponíveis nas empresas são preenchidas por pessoas que estão em busca de emprego a partir de um conjunto complexo de redes de informações e arranjos organizacionais [9] . Em função dessa complexidade, se torna mais barato para um indivíduo contratar serviços de intermediários para a busca de vagas ou de trabalhadores mais próximos ao perfil desejado. Neste contexto, para aqueles trabalhadores ou empresas que não possuem recursos para pagar pelos serviços de intermediação, são disponibilizados os Serviços Públicos de Emprego (SPE).

Além dos serviços de intermediação de mão de obra, os SPE têm ainda a responsabilidade de auxiliar os trabalhadores desempregados a reduzir ou eliminar as barreiras que os impedem de serem reinseridos no mercado de trabalho [9]. Outra atividade desempenhada pelos SPE é o suporte às populações vulneráveis no enfrentamento aos desafios do mercado de trabalho [10].

Entre os principais desafios a serem enfrentados pelos SPE, destacam-se os relacionados às mudanças demográficas, à globalização, às inovações tecnológicas e a desalinhamentos do mercado de trabalho [11].

Em relação às mudanças demográficas, como o envelhecimento da população, é desejado que os SPE estimulem a participação no mercado de trabalho dos trabalhadores inativos, a partir do desenvolvimento de habilidades individuais. Aqui devem ser considerados os desafios inerentes a cada faixa etária.

No que se refere à globalização, a internacionalização das cadeias produtivas implica na necessidade de maior investimento em inovação e desenvolvimento de habilidades, de forma a fazer frente à competitividade internacional, assim como reintegrar os trabalhadores que foram desempregados em decorrência do rearranjo dos processos produtivos.

Além disso, o desenvolvimento tecnológico tem sido responsável por significantes mudanças no mercado de trabalho, com a redução de postos de trabalho nos setores primários e o aumento da demanda por profissionais a serem empregados em serviços e atividades baseados em conhecimento.

Neste contexto, os SPE podem desempenhar um importante papel na redução do desalinhamento entre a demanda das empresas por profissionais mais qualificados e os trabalhadores disponíveis que não possuem as competências requeridas. Isso pode ser feito a partir da oferta de cursos cujo objetivo é desenvolver nos trabalhadores as competências mais demandadas pelos empregadores, visando à requalificação profissional daqueles que tiveram seus empregos impactados pelas mudanças tecnológicas. Isso pode ser feito tanto reduzindo o tempo para a reinserção no mercado de trabalho daqueles que foram desligados, quanto evitando o desligamento dos trabalhadores requalificados.

No Brasil, a gestão das políticas públicas com foco na empregabilidade é de competência do Ministério do Trabalho e Emprego (MTP). Os serviços relacionados à inserção de trabalhadores no mercado de trabalho são ofertados diretamente à população por meio do Sistema Nacional de Emprego (Sine). Entre os serviços disponibilizados pelo Sine, encontram-se o seguro desemprego, a intermediação de mão de obra, a qualificação profissional e o fomento à geração de renda por meio do microcrédito.

A fim de otimizar o trabalho dos SPE, Gazier [12] sugere que há ganhos resultantes do estabelecimento de critérios objetivos para a separação de trabalhadores em grupos de acordo com o grau de empregabilidade de cada indivíduo. Com essa segregação, é possível, por exemplo, que sejam antecipadas as medidas para a redução do tempo em desemprego prioritariamente para os indivíduos dos grupos de maior risco.

Entre as regras de diferenciação de trabalhadores mais comuns, destacam-se as regras de elegibilidade, o critério do assistente social, o *screening* e o perfilamento estatístico [9]. Nas regras de elegibilidade, os serviços a serem oferecidos aos trabalhadores estão relacionados a características individuais, como a seleção baseada no gênero. No caso do critério do assistente social, o responsável pelo atendimento do usuário estabelece uma pontuação de priorização do trabalhador baseado na percepção do atendente. Em relação ao *screening*, os grupos de trabalhadores são definidos a partir de testes realizados com os usuários, de forma que seja atribuída uma nota a cada indivíduo de maneira objetiva. Por fim, no perfilamento estatístico são utilizadas técnicas estatísticas para identificar trabalhadores com uma maior probabilidade de permanecerem em uma situação indesejada, como por exemplo, fora do mercado de trabalho por um longo período.

As medidas atualmente utilizadas para a identificação de grupos prioritários no Sine são baseadas: 1) no critério do atendente do posto e 2) na convocação de trabalhadores, baseada em regras de elegibilidade. No primeiro caso, ao realizar o atendimento presen-

cial, o atendente utiliza a sua percepção para avaliar e orientar os usuários dos serviços do Sine. Já a convocação é o mecanismo por meio do qual os atendentes dos postos convocam trabalhadores para participar de processos seletivos para vagas cuja gestão é feita pelo posto de atendimento. Nesse processo, a lista dos convocados é preenchida automaticamente por trabalhadores pertencentes a grupos prioritários, definidos a partir de parâmetros previamente cadastrados no sistema, como por exemplo, os beneficiários do seguro de desemprego.

Mecanismos de perfilamento que possuam alta dependência da atuação humana tendem a ser cada vez mais questionados, em função da crescente pressão sobre gastos orçamentários. Isso ocorre porque o aumento da capacidade de atendimento desses serviços está diretamente relacionado à ampliação do número de atendentes, o que implica em um maior gasto de recursos [12].

Neste contexto, Rudolph [13] afirma que modelos estatísticos podem ser usados como apoio à tomada de decisão de avaliadores humanos no processo de identificação de trabalhadores com maior risco de permanecerem desempregados.

Adicione-se a esse fato que a crise sanitária ocasionada pela pandemia da COVID-19 impactou sobremaneira o mercado de trabalho brasileiro, em especial os setores de serviços relacionados ao turismo e restaurantes [14].

O contexto epidemiológico foi também um importante fator para a aceleração da transformação digital de uma série de setores privados e da administração pública [15]. Dessa forma, pode-se concluir que, nos setores mais impactados pela transformação digital, haverá a substituição, por recursos humanos mais capacitados, daqueles que exercem atividades repetitivas e que possuem menor escolaridade.

Atualmente, o MTP gerencia a plataforma com dados de trabalhadores e empregadores que utilizam os serviços do Sine, assim a base de dados do Sistema de Escrituração Digital das Obrigações Fiscais, Previdenciárias e Trabalhistas, conhecido como o eSocial, que contém informações de contratações e desligamentos de todos os vínculos formais de empresas que atuam no território nacional.

O eSocial, instituído pelo Decreto nº 8373/2014 [16], foi uma iniciativa de simplificação, do Governo Federal, por meio do qual os empregadores passaram a comunicar, de forma unificada, as informações relativas a trabalhadores [17]. Neste sentido, passaram a ser feitas pelo referido sistema as obrigações que anteriormente eram realizadas por meio de outros 15 sistemas, como exemplo a Relação Anual de Informações Sociais(Rais) [18] e o Cadastro Geral de Empregados e Desempregados (Caged) [19] .

Embora armazenem informações referentes a um mesmo contexto, que é o das contratações e das demissões realizadas por empresas sediadas em território nacional, há diferenças entre a Rais e o Caged. Entre as diferenças existentes, registre-se que aquele

sistema armazena informações de todos os vínculos trabalhistas que estiveram ativos em determinado ano de referência, enquanto este registra eventos de admissões ou desligamentos ocorridos em determinado mês de referência. Em resumo, a Rais registra o estoque de vínculos empregatícios anuais enquanto o Caged traz informações do fluxo de contratações e demissões ocorridas mensalmente.

As informações do Sine, Rais e Caged, embora pudessem ser utilizadas em um modelo de perfilamento estatístico, não são usadas atualmente para essa tarefa no Brasil, impactando de modo substancial e negativamente nas ações do governo em relação à empregabilidade e consequente ao mercado de trabalho.

Pelo exposto e em função da grande quantidade de dados armazenados nas bases da Rais, Caged e Sine, este trabalho propõe o desenvolvimento de um algoritmo baseado em aprendizado de máquina, capaz de manipular grande quantidade de dados na tarefa de identificar antecipadamente os trabalhadores com maiores dificuldades para a reintegração no mercado de trabalho formal.

1.2 Questão de Pesquisa

A partir da contextualização apresentada anteriormente, este trabalho propõe-se a responder se é possível identificar antecipadamente, a partir dos dados disponíveis, quais e onde estão os trabalhadores com maiores dificuldades para a reintegração no mercado de trabalho a fim de indicar para os gestores locais quais trabalhadores devem ser atendidos de maneira prioritária e quais características possuem maior impacto no tempo do trabalhador em desemprego.

1.3 Justificativa

O mecanismo de agrupar trabalhadores desempregados segundo as dificuldades individuais para a reinserção no mercado de trabalho é conhecida como perfilamento (*profiling* em inglês, tradução própria) [13].

De acordo com [20], a inexistência de um mecanismo de perfilamento leva à utilização de estratégias menos eficientes para a seleção de trabalhadores que precisam de maior assistência, como: 1) assumir que todos os usuários dos serviços públicos de emprego possuem as mesmas dificuldades para a inserção no mercado de trabalho; ou 2) priorizar o atendimento a grupos específicos, como indígenas ou portadores de deficiências. A primeira abordagem falha ao resultar no gasto de recursos para uma assistência intensa e ampla oferecida a trabalhadores que não precisariam de qualquer auxílio para a reintegração no mercado de trabalho. Já a segunda abordagem ignora que características de

indivíduos distintos e pertencentes a um mesmo grupo podem impactar nas chances deles se reintegrarem no mercado de trabalho, por não terem sido consideradas como critério de segmentação

O'Connell, Mcguinness e Kelly [21] acrescentam que a antecipação de intervenções de assistência sobre os trabalhadores com maiores dificuldades em se recolocarem no mercado de trabalho resultam em maior efetividade e eficiência da utilização de recursos públicos.

Acrescente-se ainda que [14] e [22] apresentam uma série de impactos negativos impostos pela pandemia do COVID-19 no mercado de trabalho, no mundo inteiro e inclusive no Brasil, como a redução da massa salarial e o aumento da informalidade e do desemprego, tornando ainda mais necessário que medidas técnicas administrativas sejam tomadas com o objetivo de minimizarem os transtornos e impactos negativos que essas mazelas imprimiram na vida de trabalhadores e empregadores.

Nesse contexto, a existência de um mecanismo que permita a identificação antecipada de trabalhadores que necessitem de um maior auxílio para serem reintegrados ao mercado de trabalho possibilitará que os serviços de orientação profissional atualmente oferecidos pelos gestores regionais do Sine sejam direcionados preferencialmente aos grupos de maior risco, o que resultará na maior efetividade e eficiência da política pública de emprego brasileira, que parece ser fundamental para qualquer mercado de trabalho e em especial em tempos de recuperação econômica, como por exemplo, na fase pós COVID-19.

1.4 Hipótese de Pesquisa

Diante do exposto, apresenta-se a hipótese de que é possível identificar, a partir de técnicas de classificação aplicadas sobre os dados referentes às características pessoais e curriculares, além de informações do mercado de trabalho, o risco dos trabalhadores registrados no Sine, após a demissão, necessitarem de mais de doze meses para a reintegração profissional. O período de 12 meses foi escolhido em função desse ser o parâmetro utilizado pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) para definir desempregados de longa duração[23].

1.5 Objetivos

1.5.1 Objetivo Geral

O objetivo geral deste trabalho é construir um modelo de perfilamento fundamentado em técnicas de aprendizado de máquina que seja capaz de prever se um trabalhador registrado

no Sine necessitará de um período superior a doze meses para a reinserção no mercado de trabalho após a sua demissão.

1.5.2 Objetivos Específicos

A identificação dos trabalhadores registrados no Sine que necessitarão de mais de doze meses para a reintegração no mercado de trabalho tem como objetivos específicos:

1. Detectar os trabalhadores cujas estimativas de tempo em desemprego sejam maiores que o parâmetro considerado;
2. Identificar as características desses trabalhadores que mais impactam na duração do tempo em desemprego; e,
3. Indicar fatores mais relevantes para o tempo em desemprego em uma determinada região.

1.6 Metodologia

Esta seção descreve a metodologia utilizada nesta pesquisa, em duas etapas: caracterização e estruturação, apresentadas a seguir.

Caracterização da pesquisa

De acordo com definção de [24], esta investigação pode ser classificada quanto ao nível de profundidade como pesquisa descritiva, pois nela são estudadas as correlações entre diferentes variáveis, dos trabalhadores e do mercado de trabalho, com o tempo em desemprego.

Segundo ponderações de [25], esta pesquisa pode ser considerada como trabalho original quanto à natureza por apresentar conhecimento novo derivado dos dados utilizados no estudo. Reforça-se a isso o fato dela possuir implicação prática na sua realização, com reflexos diretos na política pública de emprego brasileira.

Em relação aos procedimentos técnicos, o presente trabalho classifica-se como documental seguindo a definição de [26], tendo em vista que as bases da Rais, Caged e Sine, embora já tenham sido objeto de estudos diversos, terão um tratamento analítico reelaborado de acordo com os objetivos da presente pesquisa.

No que tange à forma de abordagem, este estudo é classificado como quantitativo posto que ele testa hipóteses e as avalia a partir da utilização de técnicas estatísticas[27].

Estruturação da Pesquisa

Esta pesquisa foi estruturada a partir do *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [28], processo iterativo pelas seguintes fases: Entendimento do negócio, Entendimento dos dados, Preparação dos dados, Modelagem e Avaliação do modelo, organizadas conforme Figura 1.1.

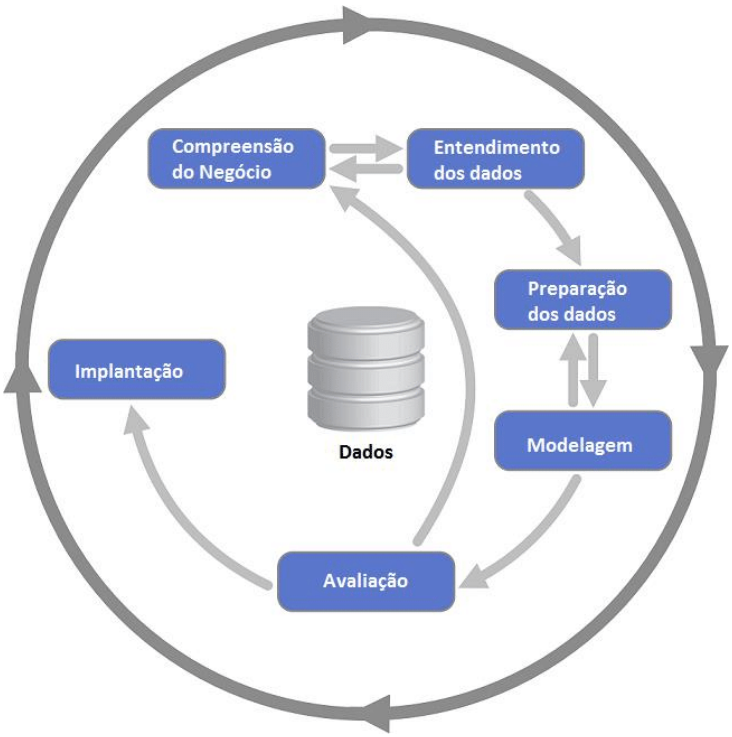

Figura 1.1: CRISP-DM : Ciclo do processo de análise de dados, como citado por Ramos [1].

Conforme ilustrado na Figura 1.1, o CRISP-DM é composto por seis fases iterativas organizadas de forma que os resultados de uma etapa podem implicar na revisão de etapas antecedentes. Isso leva a um processo de refinamento contínuo dos modelos resultantes do trabalho de análise de dados. As etapas que compõem o são descritas a seguir:

1. Entendimento do negócio: conjunto de atividades que visam ao entendimento dos objetivos de negócio a serem impactados pela tarefa de Mineração de Dados;
2. Entendimento dos dados: contempla procedimentos para a coleta, organização e documentação de dados para a melhor compreensão das informações disponíveis relacionadas à tarefa de Mineração de Dados. Nesta etapa também são identificados problemas de qualidade nos dados e análise exploratória dos dados;

3. Preparação dos dados: composto por ações relacionadas à preparação do conjunto de dados final, ou seja, a base a ser utilizada no treinamento e teste de modelos das etapas posteriores;

4. Modelagem: processo de treinamento, teste e calibração de modelos candidatos que possuem maior probabilidade de apresentar resultados satisfatórios na solução do problema;

5. Avaliação: etapa em que os modelos mais promissores possuem seus desempenhos comparados e analisados; e,

6. Implantação: fase em que o modelo é adaptado para a disponibilização em ambiente de produção.

No presente estudo, a etapa de **Entendimento do Negócio**, foi tratada no Capítulo 1, onde foi definido que o trabalho objetivará a construção de um modelo de identificação, a partir das bases disponíveis no MTP, de trabalhadores que possuem maior risco de permanecerem desempregados por um período superior a 12 meses após o desligamento.

Em seguida, na fase de **Entendimento e Preparação dos Dados** foram realizados os procedimentos de análise exploratória e verificação de qualidade nas bases do do Rais, referentes ao ano de 2019 e do Sine e do Caged referentes aos anos 2020 e 2021. O período foi escolhido para que fosse possível identificar os grupos que possuíssem pelo menos um vínculo de emprego formal vigente durante os anos de 2019 a 2021, e que apresentaram maiores dificuldades de reinserção no mercado de trabalho durante a retomada da economia após o início da pandemia de COVID-19, em março de 2020.

A base do Sine, obtida a partir de extração de dados realizada sobre as plataformas digitais que suportam a política pública de emprego brasileira, possui, além de informações demográficas, dados curriculares dos trabalhadores que utilizam os serviços de intermediação de mão de obra. Essas informações são de fundamental importância para a construção do perfil profissional do trabalhador. Ademais, também estão disponíveis na base do Sine informações a respeito de pretensões profissionais e o resultado de processos seletivos realizados por meio da plataforma do Sine.

No que se refere às bases da Rais e do Caged, os dados são armazenados em formato texto, assim como no banco de dados Teradata®. Essas bases, disponíveis para estudos realizados na Secretária de Trabalho do MTP, contêm dados relativos, entre outras informações, à atividade exercida e ao salário recebido pelo trabalhador em determinado contrato de trabalho, à área da atividade econômica da empresa contratante, além de características demográficas dos empregados. Em resumo, essas bases contêm informações sobre empregos formais, das quais o registro é mandatório para as empresas sediadas em

território nacional. Há algumas diferenças entre a Rais e o Caged, como o alcance e a frequência de atualização.

Na Rais são armazenadas informações de todos os vínculos empregatícios formais que se encontraram ativos em algum momento no ano de referência. Já no Caged, são registradas informações de admissões e demissões apenas dos vínculos celetistas ocorridos em determinado mês de referência.

Informações a respeito das atividades exercidas e pretensões profissionais são armazenadas nas bases supracitadas utilizando-se códigos de ocupações da Classificação Brasileira de Ocupações (CBO). Uma ocupação corresponde a um conjunto de atividades potencialmente exercidas no contexto de uma dada relação de emprego. Esse código é composto por seis dígitos, de forma que, o primeiro dígito identifica grupos, os três, quatro e cinco primeiros dígitos representam, respectivamente, os subgrupos principais, os subgrupos e as famílias. Por fim, os seis dígitos em conjunto formam as cerca de 2.400 ocupações previstas na CBO. Essa estrutura é baseada na classificação internacional, dada a público em 1998[29].

Na fase de **Preparação dos dados**, foi construída a versão final da base que utilizada no treinamento e validação dos modelos de aprendizado de máquina. Neste contexto, foram realizados os procedimentos de mitigação dos problemas de qualidade dos dados identificados na etapa de Entendimento dos Dados.

Nesta fase também foi criada a variável alvo "Desempregado de Longa Duração". Essa variável foi construída a partir do cálculo de tempo entre os eventos de desligamento e de admissão subsequente de um mesmo trabalhador, registradas no Caged. Quando esse intervalo foi superior ou igual a 12 meses, atribuiu-se "Sim"a essa variável ou "Não", caso contrário.

Adicionalmente, foram definidos os critérios para a seleção de atributos a serem utilizados no estudo, assim como para a separação das amostras em treinamento e testes.

Por se tratar de uma tarefa de classificação, utilizou-se como métrica para a avaliação dos modelos, que foram construídos na etapa de modelagem, o AUC.

Ato contínuo, na **Modelagem de dados** foram treinados os modelos de classificação utilizando-se os algoritmos de Regressão Logística, *Random Forest*, GBM, XGBoost e Redes Neurais.

A Regressão Logística foi escolhida como algoritmo base por ser um modelo mais simples. Outra vantagem é que os coeficientes dos atributos aprendidos durante o treinamento possuem aplicação direta na política pública, conforme exposto no Capítulo 2. Os modelos *Random Forest*, *Gradiente Boosting Machines*, *Extreme Gradient Boosting* e Rede Neural com *embedding* de código da Classificação Brasileira de Ocupações terão seus desempenhos comparados entre si e em relação ao algoritmo base de Regressão Logística.

A performance preditiva dos algoritmos também foi armazenada nesta fase para avaliação posterior.

Na fase de **Avaliação de modelos** foram conduzidas a avaliação e a comparação dos modelos, assim como foi feita a consolidação dos resultados.

Na fase de **Implantação**, foram discutidos aspectos quanto a estratégia de utilização dos resultados apresentados, na operação dos postos de atendimento do Sine.

Por fim, as etapas de Entendimento dos Dados, Preparação dos Dados e Modelagem são apresentadas no Capítulo, enquanto as fases de Avaliação dos Modelos e Implantação são apresentadas no capítulo de Conclusão deste trabalho.

1.7 Estrutura do Trabalho

No Capítulo 1 é apresentado o contexto relacionado a este estudo, assim como a justificativa, a hipótese e os objetivos de pesquisa, além da metodologia deste trabalho. O Capítulo 2 apresenta o referencial teórico utilizado para a elaboração desta dissertação, com os conceitos utilizados no restante deste documento, além de expor experiências na aplicação do perfilamento estatístico em políticas sociais diversas e, especificamente, em políticas públicas de emprego. No Capítulo 3 é apresentado o estudo de caso conduzido nesta pesquisa, com os resultados dos procedimentos de entendimento e preparação dos dados, como a análise descritiva e a limpeza das bases, assim como os resultados das etapas de modelagem e validação dos modelos, em que são expostos os resultados do estudo. Por fim, na seção Conclusões e Trabalhos Futuros são apresentadas as conclusões do trabalho, os resultados alcançados, considerando os objetivos geral e específicos estabelecidos, assim como as possibilidades de trabalhos futuros.

Capítulo 2

Referencial Teórico

2.1 Descoberta de conhecimento em bases de dados

Segundo [3], a descoberta de conhecimento em bases de dados (KDD em inglês) é um processo composto por etapas de pré-processamento, mineração e pós-processamento de dados, conforme esquema apresentado na Figura 2.1.

Figura 2.1: O processo de descoberta de conhecimento em bases de dados (KDD)[2].

2.1.1 Pré-processamento

A etapa de pré-processamento de dados é a fase em que são realizadas, entre outras atividades, a limpeza e o pré-processamento dos dados. O objetivo dessa limpeza de dados é a garantia da qualidade dos dados. Já a finalidade da preparação dos dados é ajustar as informações ao formato esperado pelos modelos a serem utilizados na mineração desses dados[3].

Limpeza de Dados

[3] afirmam que a qualidade de dados pode ser prejudicada por erros humanos, de medição ou de coleta de dados. O primeiro caso compreende as situações em que ocorre a inserção

de informações incorretas nas bases de dados utilizadas em determinada atividade, como a introdução de um número de identificação errado no cadastro de um trabalhador. Os erros de medição são causados por problemas no dispositivo de captação, o que acontece, por exemplo, quando realizamos uma ligação para uma pessoa incorreta devido a um erro no processamento de voz de um assistente virtual de um telefone . Por fim, erros de coleta de dados acontecem em situações em que registros são omitidos ou incluídos, parcial ou totalmente, de maneira indevida, como a situação em que a imagem de uma espécie diferente da pesquisada é introduzida em uma base em função de suas características, que são similares às da espécie alvo.

Entre os principais problemas de qualidade de dados, [3] destacam a existência de valores anômalos, valores faltantes e valores inconsistentes.

Valores anômalos são aqueles registros ou atributos que apresentam valores diferentes do esperado. É possível que dados anômalos sejam legítimos em alguns cenários, além de serem de interesse da tarefa de análise de dados, como eventos ilegítimos em um sistema de detecção de fraude.

O problema de valores faltantes ocorre quando um ou mais atributos de uma observação não estão disponíveis para a análise. A fim de lidar com esse problema, diferentes estratégias de limpeza de dados podem ser utilizadas, como estimar valores a serem utilizados no lugar dos campos faltantes, eliminar atributos ou registros incompletos, ou ainda, ignorar que existam os dados faltantes durante a análise de dados.

Dados também podem conter valores inconsistentes em seus atributos. Por exemplo, considerando uma situação em que um dado trabalhador esteja registrado ao mesmo tempo com gêneros diferentes em bases de governo distintas, se faz necessário corrigir e harmonizar esses dados caso essa informação seja utilizada em algum trabalho de análise de dados, como por exemplo, quando se pretende avaliar a importância do gênero no tempo em desemprego.

Pré-processamento de Dados

Entre as técnicas aplicáveis no pré-processamento de dados, encontram-se a agregação, a amostragem de dados, a redução de dimensionalidade, a seleção de atributos, e a discretização ou binarização [3].

A agregação de dados é o processo por meio do qual duas ou mais observações são combinadas a fim de criar um único registro, como por exemplo, agrupar operações diárias em registros mensais.

A amostragem de dados é um processo de seleção de um subconjunto dos dados a serem analisados, comumente utilizado na mineração de dados com o objetivo de reduzir o custo ou o tempo para processamento de toda a massa de dados. De forma a garantir

que as observações da amostra possuam características semelhantes às da população, é necessário que a amostra selecionada seja representativa.

Em algumas situações, os conjuntos de dados possuem um número muito grande de atributos, o que pode implicar em uma série de problemas, como a perda de performance de alguns algoritmos de aprendizagem de máquina ou redução da interpretabilidade dos modelos resultantes. Nestes casos, técnicas de redução de dimensionalidade ou de seleção de atributos podem ser aplicadas. A redução de dimensionalidade é obtida a partir da criação de novos atributos por meio da combinação dos atributos antigos. Já a seleção de atributos é o processo por meio do qual a redução do número de atributos é obtida a partir da remoção de informações redundantes ou irrelevantes. Técnicas de Análise de Sobrevivência podem ser instrumentos interessantes na identificação de atributos significativos para a análise do tempo de desemprego. Essa técnica será melhor detalhada na subseção Análise de Sobrevivência.

A discretização e a binarização são atividades de pré-processamento necessárias para, respectivamente, transformar valores contínuos em discretos, ou discretos e contínuos em um ou mais atributos binários. Um exemplo de binarização a ser utilizado no contexto deste trabalho é transformar o atributo com a ocupação de um vínculo de trabalho em n atributos binários, onde n é o número de possíveis códigos de CBO. Neste caso, um código de CBO específico será representado com o número 1 na coluna referente ao seu código CBO e 0's em todas as outras $n-1$ colunas referentes aos demais códigos de CBO.

Ainda no contexto da binarização do código CBO, considerando: 1) a grande quantidade de possibilidades de códigos e; 2) que há grupos de CBO's que possuem maior proximidade entre si do que em relação a outros grupos. A estratégia escolhida para a representação desses dados deve, idealmente, manter essas características na nova representação sob pena da perda de informações. Tome-se como exemplo os códigos CBO de Analista de Desenvolvimento de Sistemas; Analista de Redes e; Pedreiro. É fácil observar que os dois primeiros códigos de CBO possuem mais similaridade entre si do que em relação ao terceiro. Desconsiderar essa característica no processo de discretização dos valores levará a modelos que tratam os códigos acima como equidistantes entre si.

[30] sugere que em situações semelhantes à apresentada acima, a utilização de *embeddings* pode ser mais adequada. [31] afirmam que utilizar a abordagem de mapear variáveis categóricas em *embeddings* permite que valores similares sejam mapeados próximos uns dos outros no espaço de *embedding*, o que possibilita a manutenção das características intrínsecas das variáveis originais em suas respectivas representações.

Como exemplo do mapeamento de atributos para *embeddings*, no processamento de linguagem natural, palavras e frases têm sido mapeadas para o espaço semântico, de forma que a representação vetorial nesse espaço mantém palavras similares próximas e a

distância entre palavras e a direção dos vetores de suas diferenças também possuem valor semântico[32]. Desta forma, os vetores de representação das palavras mantêm relações como:

$$Rei - Homem \approx Rainha - Mulher \tag{2.1}$$

$$Paris - França \approx Roma - Itália \tag{2.2}$$

Análise de Sobrevivência

Análise de Sobrevivência é uma coleção de procedimentos estatísticos para análise de dados em que se deseja obter o tempo até que determinado evento de interesse ocorra [33]. O tempo pode ser medido em qualquer unidade de tempo, como minutos, meses ou anos. Um exemplo de evento de interesse é o tempo até a recolocação profissional, após a demissão ou mesmo a morte de um grupo de pacientes em tratamento médico.

A principal motivação para utilizar Análise de Sobrevivência é lidar adequadamente com dados censurados, que são situações em que não há a certeza de que o evento em análise ocorreu ou não. Um caso de dados censurados é verificado, por exemplo, em um estudo em que se avalie o tempo entre a demissão e a reintegração profissional de um grupo de trabalhadores e que, durante esse processo, algum trabalhador não tenha sido reinserido no mercado de trabalho até o término do estudo. Nessa situação, não se pode precisar quanto tempo esse trabalhador levou para a sua reinserção no mercado de trabalho. Nesse caso, considera-se uma censura à direita.

Há dois conceitos principais relacionados à Análise de Sobrevivência, a função de sobrevivência, denotada por $S(t)$, e a função de risco, indicada como $h(t)$. A função de sobrevivência $S(t)$ calcula a probabilidade de o evento analisado não ocorrer até determinado tempo t. Em relação à função de risco, ela calcula a probabilidade de que o evento ocorra no momento t, considerando a não incidência anterior a t.

O modelo *Weibull* é o modelo paramétrico de sobrevivência mais amplamente usado. A sua função de risco é dada por $h(t) = \lambda p t^{p-1}$, onde p e $\lambda > 0$. A variável p determina o formato da função de risco, de forma que, se $p > 1$, $p = 0$ ou $p < 0$, o risco é cresce, permanece constante ou decresce com o aumento do tempo, respectivamente [33].

[34] utilizou técnicas de Análise de Sobrevivência nos dados do mercado de trabalho da Romênia, a fim de identificar fatores que mais impactam no tempo para a recolocação profissional.

2.1.2 Mineração de Dados

A Mineração de Dados é o processo automático por meio do qual se descobre informação útil em grandes repositórios de dados[3]. [35] definem que a etapa de mineração de dados está comumente relacionada a quatro diferentes estilos de aprendizado. Na Classificação, um conjunto de exemplos previamente identificados são apresentados ao esquema de aprendizado a fim de conseguir identificar a classe mais adequada a novas observações. Na Associação, são identificadas quaisquer relações entre os atributos das observações, mesmo que não sejam referentes à classificação das observações. No Agrupamento, registros com características semelhantes são agrupados de acordo com a proximidade em relação às outras observações do grupo. Já na Predição Numérica, a relação entre retorno e atributos que se deseja inferir não está relacionada a uma categoria, mas sim a uma quantidade numérica.

Conforme definido acima, a classificação é a tarefa de atribuir rótulos a observações ainda não rotuladas a partir de um classificador. O processo de construção de um classificador envolve a utilização de uma amostra de dados rotulados previamente, chamado de conjunto de treinamento. Após a separação do conjunto de treinamento, é realizado o processo de aprendizado estatístico, também chamado de indução, em que são estabelecidas relações entre os rótulos observados, os hiperparâmetros dos modelos e os valores dos atributos de cada observação do referido conjunto. Após isso, utiliza-se o modelo treinado para rotular instâncias não observadas anteriormente, em um processo conhecido como dedução. Essas etapas estão ilustradas na Figura 2.2.

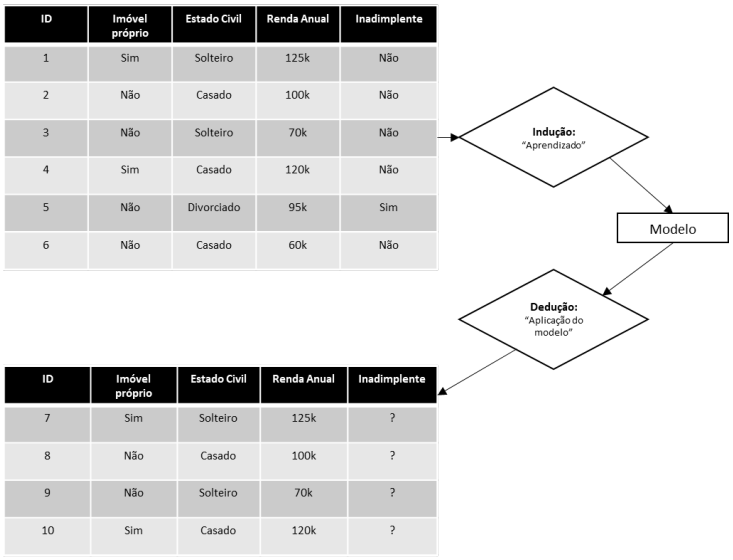

Figura 2.2: Processo de aprendizado estatístico para a construção de um classificador. Adaptado de [3].

Hiperparâmetros são parâmetros dos algoritmos de aprendizagem que não são afetados pelo processo de aprendizagem, ou seja, devem ser selecionados antes do treinamento e se mantêm constantes durante todo esse processo. Por esse motivo, o ajuste de hiperparâmetros, que é a atividade de testar diferentes combinações de valores de hiperparâmetros a fim conseguir melhores resultados no processo de aprendizagem, é uma importante parte do processo de construção de um modelo de aprendizado de máquina. [30] Para fins de simplificação, o termo parâmetro será utilizado como sinônimo de hiperparâmetro no restante deste texto.

Em situações em que o modelo de aprendizado de máquina selecionado possui grandes quantidades de possíveis combinações de valores de parâmetros, o ajuste desse modelo de forma manual se torna proibitivo em função do tempo para se avaliar todas as combinações. Para esses casos as bibliotecas de aprendizagem de máquina oferecem ferramentas para testar todas as combinações de uma lista de parâmetros no treinamento de determinado modelo preditivo. Após a identificação da combinação de parâmetros utilizados no treinamento do melhor modelo, ou seja, a combinação que resultou no modelo com a melhor performance, a ferramenta retorna, além dos valores dos parâmetros, informações sobre os resultados do modelo selecionado. Como exemplo, a biblioteca *scikit-learn*[1] disponibiliza a classe *GridSearchCV* para essa atividade.

Métricas

A fim de avaliar a performance dos modelos, é fundamental que sejam utilizadas técnicas que permitam analisar e comparar a capacidade preditiva dos modelos de classificação. Nesse sentido, a matriz de confusão, apresentada na Tabela 2.1, é a forma mais básica para se avaliar a performance de um algoritmo de Classificação [3]. Por meio dessa matriz é possível analisar o número de observações classificadas corretamente ou incorretamente a partir dos seguintes quantificadores:

Tabela 2.1: Uma matriz de confusão para um problema de classificação binária

		Classe estimada	
		+	-
Classe real	+	VP	FN
	-	FP	VN

1. Verdadeiro Positivo (VP) : número de exemplos positivos classificados corretamente;
2. Falso Positivo (FP) : número de exemplos negativos classificados erroneamente como positivos (também chamados de erro do Tipo I);

[1] https://scikit-learn.org/

3. Falso Negativo (FN) : número de exemplos positivos classificados erroneamente como negativos (também chamados de erro do Tipo II); e,
4. Verdadeiro Negativo (VN) : número de exemplos negativos classificados corretamente.

Embora a matriz confusão apresente a informação da performance de classificação de maneira concisa, a sua utilização não é tão adequada quando se deseja comparar a performance de diferentes classificadores. Por esse motivo, usualmente as informações da matriz de confusão são sumarizadas em métricas de avaliação, como a acurácia, a precisão, a revocação e a F_1 *measure* [3].

Uma das métricas mais simples, a acurácia mede a taxa de acerto de determinado modelo a partir da Fórmula 2.3.

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.3}$$

Outra métrica de avaliação comumente utilizada como medida de performance de algoritmos, a área abaixo da curva *Receiver Operating Characteristic* (ROC) foi investigada por [36]. Esse autor concluiu que a utilização dessa técnica, conhecida como *Area Under the Curve* (AUC), apresenta uma série de vantagens em relação à acurácia, motivo pelo qual recomendou a sua utilização como métrica para avaliação de algoritmos. A Figura 2.3 ilustra o comportamento da curva ROC, a partir da qual é calculada a medida AUC.

Em situações em que há o desbalanceamento entre classes, ou seja, um grande número de observações em uma classe, associado a um número baixo em outra, não é aconselhável utilizar-se a acurácia como métrica, posto que ela tende a favorecer classificadores que conseguem ter alta taxa de acerto apenas nas classes majoritárias. Nesta situação é desejável utilizar outras métricas, como a precisão.

A precisão é uma métrica por meio da qual se avalia a capacidade de um classificador acertar ao identificar uma observação como pertencente a um grupo de interesse. A precisão é calculada a partir da Fórmula 2.4.

$$Precisão = \frac{VP}{VP + FP} \tag{2.4}$$

Quando há o interesse de avaliar a capacidade de um classificador em identificar corretamente as instâncias positivas de uma amostra, utiliza-se a revocação, medida a partir da Fórmula 2.5.

$$Revocação = \frac{VP}{VP + FN} \tag{2.5}$$

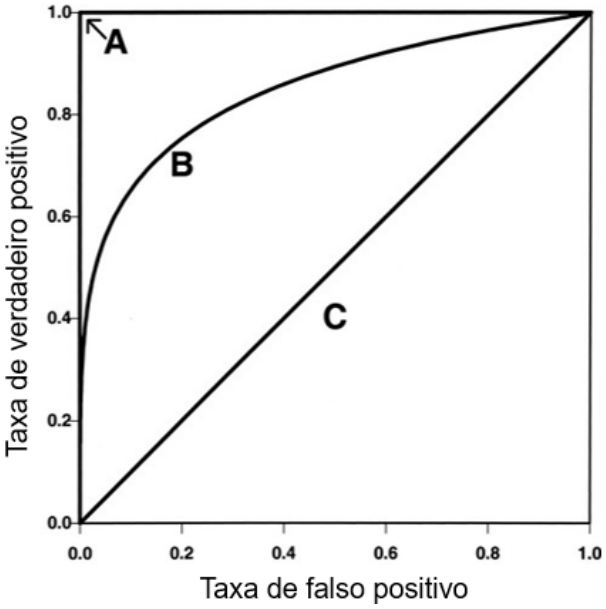

Figura 2.3: Ilustração de três curvas hipotéticas representando: (A) um classificador com acurácia perfeita, (B) uma típica curva ROC e (C) uma linha diagonal, representando um classificador aleatório. Adaptado de [4].

A fim de avaliar a capacidade de um modelo classificar com alta taxa de sucesso o maior número de observações positivas, é possível combinar a revogação e a precisão em uma única métrica, o F_1 *score*, dado pela Fórmula 2.6.

$$F_1 score = \frac{2 * VP}{2 * VP + FP + FN} \tag{2.6}$$

Apesar de geralmente ser desejado que os algoritmos selecionados sejam aqueles que apresentam melhor performance de acordo com as métricas utilizadas, isso não é verdade para todas as situações. Como exemplo, há casos em que modelos mais flexíveis, ou seja, aqueles capazes de produzir um número maior de possíveis formatos de curva para se ajustar aos dados, são preteridos por outros modelos menos complexos quando o contexto prioriza a interpretabilidade à capacidade preditiva. Isso porque os modelos mais flexíveis, também chamados de complexos, apresentam uma maior dificuldade no entendimento das relações entre os atributos e a variável alvo, o que é indesejado em algumas situações. Além disso, os modelos mais simples apresentam a vantagem de serem menos suscetíveis ao sobreajuste, isto é, quando o modelo apresenta bom desempenho ao classificar as observações no treinamento, mas possui baixa capacidade preditiva durante a etapa de teste [5]. A Figura 2.4 ilustra a compensação entre essas características em diferentes modelos de classificação.

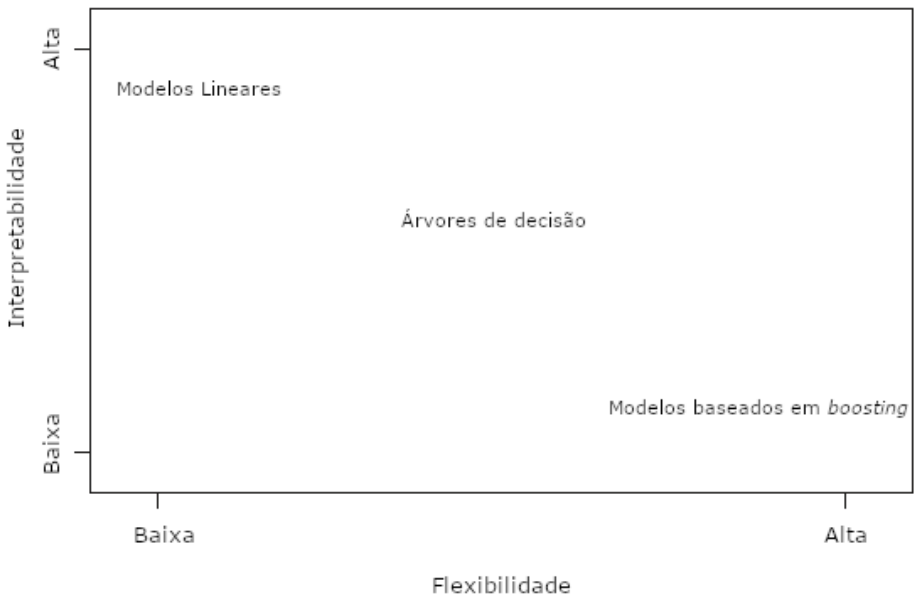

Figura 2.4: Ilustração da compensação entre interpretabilidade e flexibilidade, usando diferentes métodos de aprendizado estatístico. Adaptado de James, Witten, Hastie e Tibshirani. [5].

Entre as técnicas de classificação existentes, há a Regressão Logística, o *Random Forest*, o *Gradient Boosting Machines* (GBM), o *eXtreme Gradient Boosting* (XGBoost) e as redes neurais.

Regressão Logística

A Regressão Logística é uma técnica de aprendizado estatístico modelada a partir de uma função $p(X)$, que possui resultado contínuo entre 0 e 1 para qualquer valor de X, a partir da Fórmula 2.7:

$$p(X) = \frac{\epsilon^{\beta_0+\beta_1 X}}{1+\epsilon^{\beta_0+\beta_1 X}} \tag{2.7}$$

O valor $p(X)$ se refere à probabilidade de uma observação com atributo X pertencer a uma classe de interesse em função dos valores dos parâmetros β_1 e do intercepto β_0.

Isto posto, a referida função sempre resultará em uma curva em forma de S, conforme Figura 2.5, adaptada de [5] a partir de dados simulados de inadimplemento no pagamento do cartão de crédito.

Neste contexto, a tarefa de aprendizado consiste em encontrar os valores para os parâmetros do preditor, β_1, e da variável livre, β_0, de forma que o retorno da função corresponda aos rótulos previamente atribuídos, ou seja, a função $p(X)$ deve retornar valores mais próximos de 1 quando a classificação for positiva, e valores mais próximos de

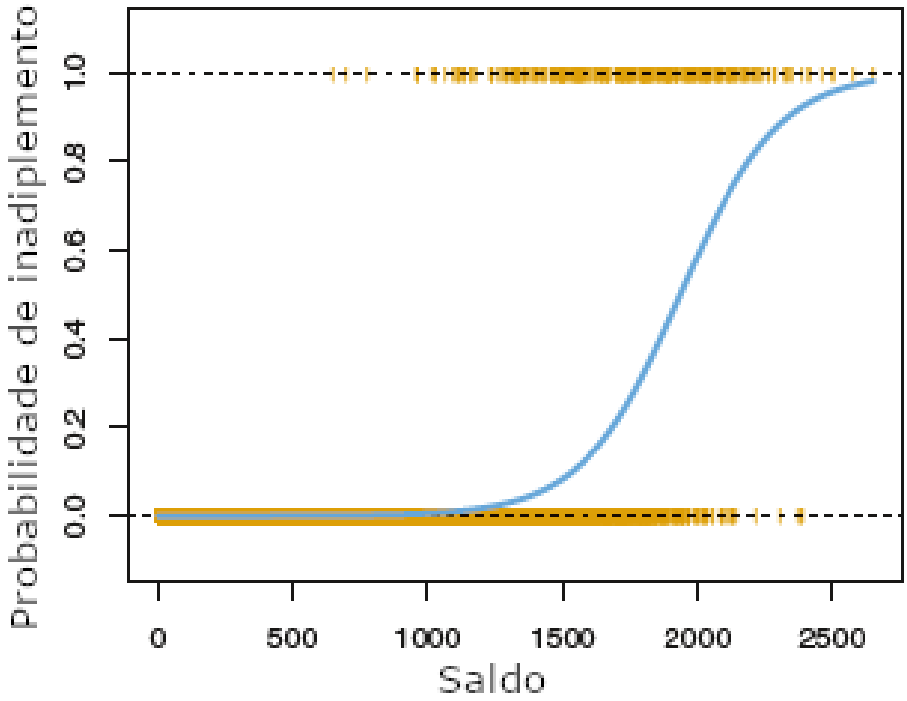

Figura 2.5: Probabilidade estimada de inadimplemento em função do saldo usando regressão logística. Adaptado de [5].

0 em caso contrário. Isso é obtido a partir da função de máxima verossimilhança, dada pela Fórmula 2.8:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \tag{2.8}$$

Para problemas que envolvam múltiplas variáveis preditoras, é possível generalizar a Fórmula 2.7 para qualquer quantidade de preditores, conforme a Fórmula 2.9.

$$p(X) = \frac{\epsilon^{\beta_0+\beta_1 X_1 \ldots \beta_n X_n}}{1 + \epsilon^{\beta_0+\beta_1 X_1 \ldots \beta_n X_n}} \tag{2.9}$$

Como resultado da Regressão Logística, obtém-se, além da probabilidade das observações pertencerem a determinada classe, os coeficientes de cada variável preditora e do intercepto, assim como a relevância estatística de cada coeficiente. As informações a respeito da relevância estatística são de grande importância na execução de políticas públicas, posto que, ao se identificar a influência de cada fator para que uma observação seja classificada em um grupo, é possível realizar intervenções que intensifiquem ou reduzam esses fatores, de forma que se aumente a probabilidade de alcançar um objetivo.

Como exemplo, em uma situação hipotética em que a Regressão Logística identificou

com alta relevância estatística que o ensino profissionalizante reduz o tempo em desemprego, é possível concluir que a oferta de cursos profissionalizantes para os trabalhadores desempregados tenderia a reduzir o tempo para a reinserção no mercado de trabalho dos indivíduos que realizassem esses cursos.

Random Forest

Random Forest é um modelo de aprendizado de máquina construído a partir da combinação de classificadores baseados em Árvores de Decisão[37]. De acordo com [6], modelos baseados em árvore particionam o espaço de características em um processo recursivo até que um critério de parada seja alcançado, conforme exemplo da Figura 2.6a.

(a) (b)

Figura 2.6: Representação do processo de construção de uma árvore de classificação [6].(a) Partições resultantes da divisão do espaço de características bidimensionais. (b) Árvore de classificação correspondente ao espaço de características apresentado à esquerda.

Ao final desse processo, obtém-se uma árvore que representa o modelo, conforme Figura 2.6b. Neste sentido, o processo de classificação ocorre ao se percorrer a referida árvore a partir da seleção sucessiva das sub árvores à esquerda ou à direita de cada nó, de acordo com os valores dos atributos de determinada observação. Por fim, atribui-se à observação a classe associada à região de cada folha da árvore.

Segundo [6], modelos baseados em árvores de decisão puros, apresentam, entre outros problemas, instabilidade e sobreajuste. A instabilidade se refere à natureza hierárquica do processo de construção da árvore: um erro no nó superior se propaga para todos os nós abaixo. De forma semelhante, o sobreajuste pode ocorrer quando existem um número grande de variáveis preditoras, fazendo com que o treinamento resulte em um modelo com baixa taxa de erro no treinamento, mas pouco adaptado a observações fora do grupo treinamento.

[37] mitigou os problemas supracitados das árvores de classificação puras a partir da geração de múltiplas árvores de decisão no modelo de *Random Forest*. Ao final, o resultado da classificação das múltiplas árvores é agregado em uma única predição do modelo.

É importante destacar que caso as árvores que compõem o modelo sejam correlacionadas, há pouco benefício na aplicação da técnica. Em função disso, a fim de evitar a correlação entre as árvores, o algoritmo proposto pelo autor prevê a amostragem de dados e de atributos antes da construção de cada uma das subárvores que fazem parte do modelo. No que tange aos dados, é utilizada uma técnica chamada de *Bootstrap*, em que os dados são amostrados com reposição. Já em relação aos atributos, é realizada a amostragem usando apenas um subgrupo de variáveis preditoras, de forma que cada árvore possua no máximo $\sqrt{p}$ atributos, onde p é o numero total de atributos[5].

Gradient Boosting Machines

O modelo *Gradient Boosting Machines* (GBM), proposto por [38], é formado a partir de um agregado de classificadores baseados em árvores de decisão em um processo por meio do qual novos classificadores são adicionados de forma a aprimorar a capacidade preditiva dos classificadores inseridos anteriormente. Isso é obtido a partir da minimização da função de erro ao se adicionar novas árvores por meio do gradiente descendente.

Ademais, a fim de garantir que as árvores com menor erro tenham maior relevância no cálculo final do modelo, a taxa de aprendizado decresce a cada iteração, o que resulta em um número maior de árvores com menor erro em comparação com aquelas que possuem maior erro. Com isso, as classificações retornadas pelas últimas árvores são mais relevantes para o modelo. Destaque-se ainda a possibilidade da utilização de uma função de erro ajustada ao contexto em que a tarefa de aprendizado de máquina está inserida. Esse maior controle permite que o resultado do modelo possua uma melhor correspondência com o mundo real.

eXtreme Gradient Boosting

[39] propuseram o algoritmo eXtreme Gradient Boosting (XGBoost) , baseado no *Gradient Tree Boosting*, também conhecido como *Gradient Boosting Machines*.

Entre as otimizações existentes no XGBoost em relação ao modelo de [38], destacam-se a utilização de um algoritmo aproximadamente guloso, aprendizado paralelo e o *Weighted Quantile Sketch* para o treinamento de grandes bases de dados.

Em relação ao algoritmo aproximadamente guloso, a estratégia gulosa usualmente utilizada em outras técnicas baseadas em GBM seleciona os valores para as divisões em cada nó a partir do melhor ganho de informação ao se realizar o corte. Desta forma, há um crescimento exponencial de quantidade de combinações a serem testadas quando se aumenta o número de atributos e de observações do conjunto de treinamento. O algoritmo aproximadamente guloso implementado no XGBoost realiza a avaliação apenas de alguns quantis do conjunto de treinamento, o que reduz o número de testes a serem feitos.

A fim de permitir a paralelização do processo de divisão dos dados em quantis, é aplicado o aprendizado paralelo. Por meio desta técnica, os dados são divididos em grupos menores, que são processados em diferentes núcleos (em um mesmo processador ou diferentes nós da rede). A partir disso, os valores são então agrupados em um histograma, que é utilizado para a o cálculo do valor aproximado dos quantis.

Por fim, de forma a evitar que o processo de seleção dos quantis resulte na perda de informação relevante, a técnica de *Weighted Quantile Sketch* reduz as quantidades de observações em quantis em que há um maior grau de incerteza de classificação, o que na prática aumenta a relevância de observações em que o algoritmo possui mais dificuldade em classificar.

Redes Neurais

Redes Neurais são técnicas de aprendizado de máquina que simulam o mecanismo de aprendizado de organismos biológicos. Da mesma forma que os neurônios que compõem o sistema nervoso humano e são conectados uns aos outros por meio de dendritos e axônios, as redes neurais artificiais são compostas por unidades computacionais conectadas umas as outras. Cada entrada de um neurônio é multiplicada por um peso antes de ser aplicada uma função a esses valores. O peso aplicado a cada entrada exerce um papel semelhante à força das ligações sinápticas que ocorre em organismos vivos[7]. Essa arquitetura é apresentada na Figura 2.7.

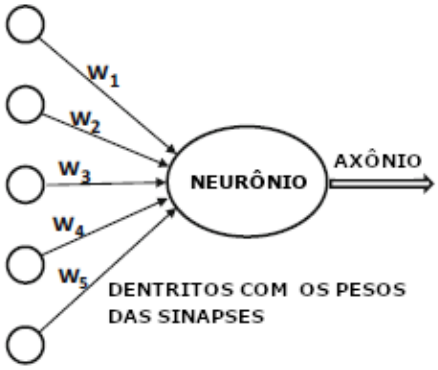

Figura 2.7: Rede neural artificial. Adaptado de Aggarwal [7].

O processo de aprendizado funciona a partir da mudança dos pesos que conectam os neurônios de acordo com retorno de quão correta está a predição em relação aos dados de treinamento.

2.1.3 Pós-processamento

Na etapa de pós-processamento, os resultados do trabalho são incorporados ao processo de trabalho. Neste sentido, no contexto da utilização de modelos de classificação em políticas públicas, é fundamental que, além da capacidade preditiva dos algoritmos, seja possível a indicação de quais fatores mais influenciam na escolha das categorias às quais as observações serão associadas. Isso porque modelos caixa preta, que são aqueles em que as relações entre os dados e as categorias são muito complexas, pouco ajudam no desenho de ações para mitigação de problemas relacionados aos atributos mais relevantes para o processo de classificação.

Considerando que alguns algoritmos apresentam alta performance na tarefa de predição, mas que não são facilmente explicáveis, foram desenvolvidas técnicas que possibilitaram a identificação dos atributos mais relevantes na tarefa de classificação para um determinado modelo.

Modelos de Explicabilidade

Sobre explicabilidade de modelos, Ribeiro[8] afirma que indivíduos tendem a não utilizar algoritmos de aprendizado em que não confiam. O autor define que a confiança em um modelo está relacionada a predições individuais e a modelos em geral. O primeiro caso corresponde a indivíduos confiarem o suficiente em um modelo a ponto de tomarem decisões a partir de suas predições. Já o segundo caso diz respeito a indivíduos confiarem que o modelo operará de maneira razoável em produção, ou seja, operando em condições reais. De forma a solucionar problemas relacionados à confiança nas predições e nos modelos, o autor apresenta um algoritmo capaz de explicar predições, o *Local Interpretable Model-agnostic Explanations* (Lime).

Ribeiro[8] define ainda que a tarefa de explicação de modelos como a utilização de recursos gráficos ou textuais para o entendimento do relacionamento entre as observações e as predições de um modelo. Esse processo pode ser melhor ilustrado a partir da Figura 2.8, em que se ilustra a utilização de um modelo para a identificação de doenças por um profissional de saúde é representada. No exemplo, além da indicação de que a doença é a gripe, o sistema apresenta quais sintomas corroboram com o diagnóstico, e quais deles o enfraquece.

[40] acrescentam que, embora modelos simples possuam informações que os permite explicar sem a utilização de processamentos adicionais, modelos mais complexos, como as redes neurais profundas, são de difícil interpretação. Em função dessa necessidade, os autores propuseram o *SHapley Additive exPlanations* (Shap), que unifica seis métodos

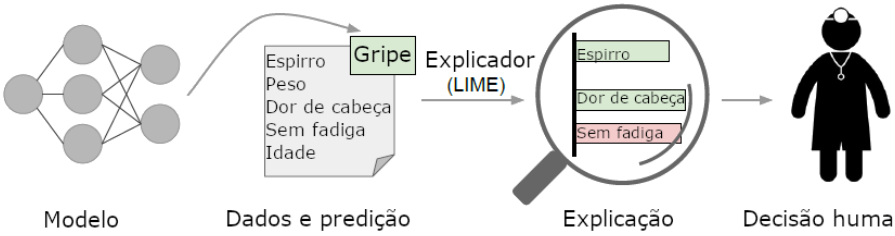

Figura 2.8: Explicação de predição em um processo de identificação de doença a partir dos sintomas. Palavras marcadas em verde contribuem para a classificação, enquanto as em vermelho são evidências contrárias, adaptado [8].

aditivos de atribuição de características e que apresenta vantagens em relação ao quanto à proximidade com a intuição humana e à eficiência computacional.

O Shap é um framework para a explicabilidade de modelos baseado na teoria dos jogos. A ideia central desse framework é identificar a contribuição marginal de cada atributo do conjunto de dados na tarefa de classificação. Isso é obtido a partir da medição da performance do modelo ao se utilizar diferentes combinações de atributos[41].

2.2 Perfilamento

Segundo [42], a habilidade de identificar com precisão os membros de um grupo alvo em um ambiente digital desorganizado é um desafio enfrentado adequadamente a partir da utilização de técnicas de classificação. Nesse sentido, no contexto do direcionamento de peças publicitárias para grupos alvo, [43] sugere haver ganhos ao se utilizar abordagens de aprendizado de máquina aplicadas a dados de usuários e conteúdos de mensagens em comparação com estratégias tradicionais de segmentação de mercado que utilizam apenas informações demográficas ou geográficas.

A Figura 2.9 exemplifica um processo em que estão inseridas as atividades de identificação de públicos alvos a partir de uma população. Após a utilização de mecanismos de identificação do público alvo, são selecionados indivíduos sobre os quais serão realizadas intervenções ou sobre os quais serão aplicados os tratamentos previstos. A etapa de seleção se faz necessária por uma série de motivos que dependem do contexto em que o perfilamento está inserido, como garantir que intervenções sejam realizadas prioritariamente sobre os indivíduos com maior probabilidade obterem impacto positivo resultante delas. Em seguida, os tratamentos são aplicados sobre os indivíduos selecionados, para, por fim, os resultados serem são avaliados a fim de verificar o impacto das intervenções sobre os indivíduos selecionados.

Figura 2.9: Processo de perfilamento. Fonte: autor.

2.3 Trabalhos relacionados

Esta seção apresenta algumas pesquisas científicas que possuem ligação com o presente trabalho e analisa as possíveis contribuições desses estudos na solução proposta.

2.3.1 Perfilamento em outras áreas

Modelos de prevenção de risco baseado em aprendizado de máquina têm sido utilizados na saúde para a identificação, por exemplo, de fatores relevantes para a mortalidade de pacientes com COVID-19 [44] ou de pacientes submetidos à intervenção coronária percutânea [45].

[46] utilizaram um modelo construído a partir de técnicas de Aprendizado de Máquina para solucionar um problema de retenção de usuários de jogos móveis. A partir do resultado, é possível subsidiar a tomada de decisão de quais intervenções possuem maior possibilidade de aumentar a retenção de jogadores.

No contexto de políticas públicas sociais, *et al.* [47] propuseram a utilização de um modelo baseado em *Random Forest* para a identificação do risco de exclusão social permanente na região de Castela e Leão, na Espanha.

2.3.2 Experiências de perfilamento no mundo

A fim de identificar as técnicas de perfilamento estatístico utilizadas nos sistemas públicos de emprego, foi utilizado como base o artigo da OCDE em que foram comparadas as experiências de utilização de técnicas de perfilamento estatístico em sistemas públicos de emprego de diferentes países [48].

Em seguida, foram realizadas consultas das bases Web of Science® e Scopus® com os termos "*worker AND profiling*", "*risk AND long-term AND unemployment*" e "*jobseeker profiling*".

Por fim, considerando a baixa quantidade de resultados de artigos indexados nas bases de artigos científicos que tratam desse tema, foram feitas consultas adicionais ao Google

Scholar®com os termos *"Predicting"*, *"risk"*, *"long-term"* e *"unemployment"*. O resultado da consolidação e seleção de artigos é apresentado abaixo.

Em operação desde 1993 nos Estados Unidos, o *Worker Profiling and Reemployment Services* (WPRS) é composto por dois componentes: 1) um mecanismo de perfilamento de trabalhadores, que identifica no momento da requisição do seguro desemprego os trabalhadores que possuem uma maior probabilidade de exaurir as parcelas do seguro a que teriam direito, que pode ser feito a partir do perfilamento estatístico ou *screening*; e, 2) um conjunto de serviços que auxiliam o grupo assim identificado na recolocação profissional [49]. A indicação de qual serviço deve ser oferecido para o trabalhador identificado com maior chance de exaurir as parcelas de seguro-desemprego depende da expertise do conselheiro local.

Segundo [50], nos Estados Unidos os serviços de reemprego são obrigatórios dependendo do grau de risco atribuído ao trabalhador pelo mecanismo de perfilamento, sem qualquer discricionariedade por parte do assistente social que acompanha o trabalhador. Além disso, em função de preocupações relacionadas aos direitos civis, o WPRS não utiliza como variáveis independentes informações relacionadas a raça, grupos étnicos, idade e gênero, o que compromete a capacidade preditiva do modelo[51], diferente do modelo utilizado no presente trabalho.

O modelo australiano, nomeado como *Job Seeker Classification Instrument* (JSCI) e em execução desde 1998, foi desenvolvido para fins de identificação antecipada de trabalhadores que, desempregados possuem maior probabilidade de permanecerem nessa situação por mais de 12 meses [20]. Esse modelo australiano foi desenvolvido a partir das seguintes etapas: 1) a realização de uma pesquisa com os trabalhadores que se encontravam em busca da recolocação profissional, momento em que, a partir da análise das respostas em conjunção com a avaliação de registros administrativos, chegaram à identificação dos fatores de risco, associados ao desemprego prolongado e ao efeito médio de cada um desses fatores; 2) a avaliação de especialistas, etapa em que foram feitas recomendações de fatores adicionais não observáveis por meio da pesquisa, mas que influenciavam a desvantagem em relação ao mercado de trabalho; e, por fim, 3) consulta ampla a demais atores que atuam no mercado de trabalho. Diferente do modelo australiano, a proposta apresentada neste estudo atribui o peso das variáveis apenas a partir do treinamento dos modelos, ou seja, baseado apenas na utilização de técnicas de aprendizado de máquina, sem a necessidade de intervenção humana.

[51] sugeriram um modelo estatístico a ser utilizado na Dinamarca. O modelo apresentado se divide em duas etapas. Inicialmente, estima-se o tempo em que o trabalhador permanece desempregado para, em seguida, com esse resultado calcular a probabilidade de permanência fora do mercado de trabalho formal, por um período adicional de 6 meses

após o momento em que a observação foi realizada.

Já na Irlanda, [21] analisaram a utilização de um modelo desenvolvido a partir de dados de registros administrativos e de questionários aplicados a todos os indivíduos que solicitaram seguro desemprego, por um período superior a 13 semanas, entre setembro e dezembro de 2006. Após essa etapa, os indivíduos foram acompanhados pelas 78 semanas seguintes a fim de serem desenvolvidos modelos de perfilamento com o objetivo de prever a probabilidade da permanência do trabalhador em desemprego por um período adicional de seis, doze e quinze meses.

Nesse estudo, [21] apresentaram ainda algumas desvantagens dos sistemas de perfilamento estatístico, quais sejam: a possibilidade de o modelo atribuir incorretamente o risco de determinado indivíduo, a necessidade de atualização contínua dos modelos a fim de ajustá-los a mudanças ocorridas no mercado de trabalho e o alto custo inicial para a sua implementação. Entretanto, esses autores destacaram que os benefícios relacionados à utilização de um modelo de perfilamento estatístico superam as desvantagens, tendo em vista: 1) a redução do número agregado de intervenções, que passam a ser feitas de acordo com os recursos disponíveis apenas sobre aqueles trabalhadores com maior risco de permanecerem desempregados; e 2) em razão das intervenções poderem ser realizadas de maneira individualizada, posto que o modelo de perfilamento traz informações sobre quais características do indivíduo mais impactam no tempo em desemprego.

No âmbito da aplicação de técnicas de aprendizado de máquina utilizadas em experiências de perfilamento de trabalhadores em outros países, [21] afirmam que não é comum a publicação de detalhes acerca da performance ou da implementação dos modelos utilizados, o que foi observado nas experiências da Dinamarca [51] e da Irlanda [21]. Neste contexto, o presente trabalho contribui com a apresentação dos resultados de performance e de treinamento dos modelos.

Apesar disso, a OCDE [11] registra que a maioria dos países que optaram por utilizar algoritmos menos complexos, como os modelos baseados em Regressão Logística em função de sua maior explicabilidade, ou seja, maior facilidade para se utilizar as informações resultantes dos modelos nas políticas públicas. Isso porque há algoritmos que, embora possuam maior capacidade preditiva, pouco informam como os atributos individuais influenciaram na tarefa de classificação, o que reduz a capacidade de se atuar sobre as causas que influenciam a ampliação do tempo em desemprego.

[52] observaram ganhos de performance na utilização de algoritmos mais complexos que o de Regressão Logística, que, como dito anteriormente, é comumente utilizado na tarefa de perfilamento. Neste contexto, esses autores comparam o desempenho do XGBoost com o de outros algoritmos, em um experimento realizado com os dados do Instituto Nacional de Emprego e Desenvolvimento Profissional Português. Como resultado, os

autores demonstraram que o XGBoost apresentou melhor desempenho que os demais algoritmos comparados no experimento.

De forma a mitigar os problemas relacionados à interpretabilidade do XGBoost, os autores [52] avaliaram também a utilização do Shap, técnica utilizada para a explicabilidade de modelos, semelhante ao Lime, na construção de um protótipo a ser utilizado durante o atendimento à população nos escritórios da política pública de emprego em Portugal. Este estudo difere do trabalho de [52] por avaliar o desempenho de um classificador baseado em rede Neural com códigos CBO mapeados para um *embedding*, além da utilização de dados de diferentes bases brasileiras.

O presente trabalho inova ao utilizar técnicas de Aprendizado de Máquina sobre os dados das plataformas digitais do Sistema Nacional de Emprego, além de registros administrativos do Rais e do Caged, com a finalidade de criar um mecanismo de perfilamento para identificar grupos de trabalhadores com a maior propensão de permanecer fora do mercado formal de trabalho. Ademais, esse trabalho propõe uma ferramenta para a identificação das características dos trabalhadores que mais contribuem para a majoração do tempo em desemprego, de forma que essa informação permita subsidiar os gestores brasileiros, dos serviços públicos de emprego, na seleção de intervenções mais adequadas ao contexto em que os trabalhadores estão inseridos.

Capítulo 3

Estudo de Caso

Este capítulo apresenta a análise exploratória de dados das bases do Sine, Rais e Caged, como parte do trabalho da fase de Entendimento dos dados, e expões a etapa de Preparação de dados, a qual detalha o trabalho de construção da base dos dados que será utilizada no treinamento e teste dos modelos de aprendizado de máquina.

3.1 Entendimento e Preparação dos Dados

3.1.1 Entendimento dos dados

Registro Anual de Informações Sociais (Rais)

Conforme exposto anteriormente, foi utilizada a base do Rais do ano de 2019, a fim de identificar trabalhadores que possuíam algum vínculo empregatício naquele ano. Foram encontrados 66.667.417 registros de empregos formais, dos quais 55.545.344 possuíam números de identificação únicos, ou seja, o número de Cadastro de Pessoa Física (CPF).

Ao se analisar os atributos referentes aos tipos de admissão e desligamento, percebeu-se que mais de 46 milhões de registros armazenavam o valor '0', não descrito na documentação como pertencente ao domínio desse campo. Essas situações ocorrem quando não houve admissão ou desligamento no ano de referência, ou seja, quando a contratação ocorreu em ano anterior e não houve demissão no ano em que os dados da Rais foram coletados.

No que se refere ao atributo raça e cor, que é auto declarado pelo trabalhador, foram identificadas 9.956.515 observações com valores nulos, ou seja, sem informação.

Em relação às pessoas com deficiência, observou-se que 66.003.641 de vínculos registrados não apresentaram informação relacionada a trabalhador com deficiência, ou seja, mais de 99% da população analisada não possuía qualquer das deficiências previstas na Rais, quais sejam: física, auditiva, visual, reabilitado, intelectual (mental), múltipla.

Considerando a necessidade de identificar inequivocamente as características dos trabalhadores com maior probabilidade de necessitarem de mais tempo para serem recolocados no mercado de trabalho, faz-se necessário que seja realizada a remoção de observações de trabalhadores que possuem múltiplos registros na Rais, quando há informações distintas de raça, data de nascimento ou sexo em pelo menos um desses registros.

A partir dos registros resultantes do processo de limpeza citado no parágrafo anterior, foi analisada a distribuição numérica de vínculos no ano de 2019 segundo as regiões geográficas, a manutenção dos vínculos ao final do ano, o gênero e a faixa etária.

A Figura 3.1 apresenta a distribuição de vínculos empregatícios que em algum momento estiveram ativos no ano de 2019. Nela é possível observar a maior concentração de vínculos de empregos formais na região Sudeste, com quase três vezes mais vínculos que a segunda região com a maior quantidade de vínculos. Em seguida, encontram-se respectivamente o Sul e o Nordeste com uma maior proximidade do número de vínculos, seguidos do Centro-Oeste e, por fim, o Norte. A partir do gráfico é possível observar também os vínculos que permaneceram ativos ao final do ano ou em que houve o desligamento.

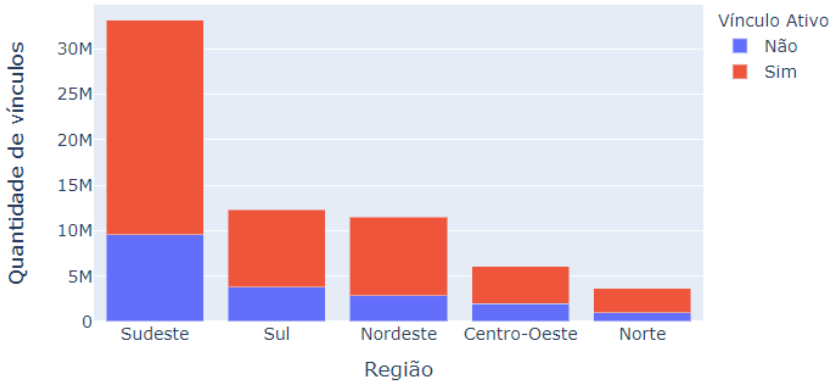

Figura 3.1: Número de vínculos da Rais por região no ano de 2019 mantidos ativos ou não.

Ao se analisar a distribuição por gênero, é possível verificar que há um maior número de vínculos formais de homens em relação a de mulheres, respectivamente cerca de 32 e 25 milhões de observações, conforme Figura 3.2.

Em relação à idade, a Figura 3.3 permite verificar que as diferentes regiões apresentam uma distribuição bem aproximada de percentual de vínculos de emprego formal por faixa etária. É possível observar ainda que a maior parte dos vínculos se concentram nas faixas de 30 a 49 anos. De outro modo, as faixas até 17 ou acima de 65 possuem o menor número de observações em comparação com as demais faixas etárias.

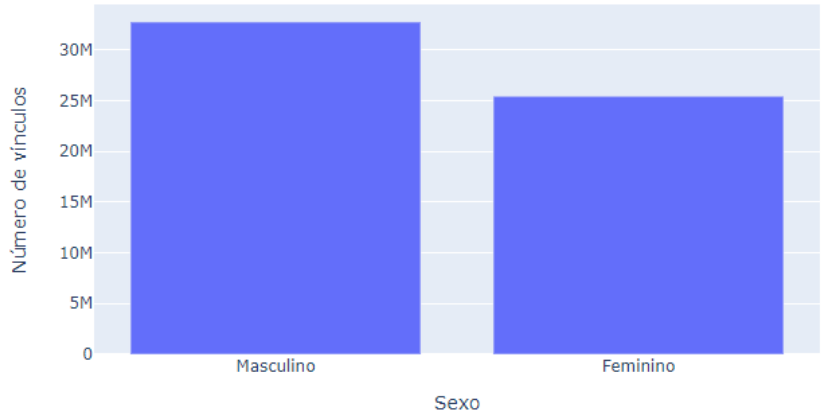

Figura 3.2: Número de vínculos da Rais no ano de 2019 por gênero.

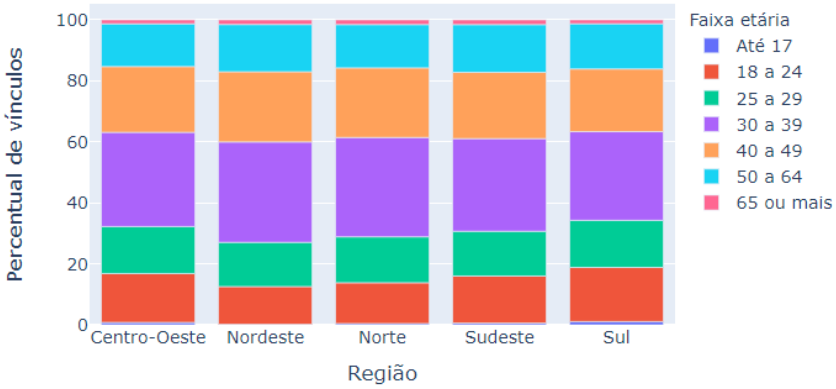

Figura 3.3: Número de vínculos da Rais no ano de 2019 por faixa etária e região.

Cadastro Geral de Empregados e Desempregados (Caged)

O presente estudo utilizou também o Caged a fim de identificar os desligamentos ocorridos durante os anos de 2020 e 2021. Diferente do Rais, o Caged registra apenas os eventos de desligamentos e contratações ocorridos em um mês de referência, ou seja, o fluxo de contratações e desligamentos mensal. Durante os períodos em análise a base registrou 65.517.779 eventos de contratações ou desligamentos.

Após a realização de uma atividade de limpeza de dados semelhante à descrita na subseção anterior, em que foram removidas observações de trabalhadores registrados com múltiplas datas de nascimento, raça ou gênero, retirou-se da população a ser analisada também os trabalhadores que tiveram registros com pelo menos uma observação com data de movimentação inválida, além das observações relacionadas a transferências, morte ou aposentadoria. As movimentações podem ser definidas como os eventos de entrada ou de saída do mercado de trabalho formal.

No caso dos registros com data de movimentação inválida, fez-se necessária a sua remoção porque a partir dessa variável, quando válida, é realizado o cálculo do tempo em que o trabalhador permanece fora do mercado de trabalho, procedimento descrito na Preparação dos dados (seção 3.1.2). Dessa forma, optou-se por remover todas as observações dos trabalhadores que apresentaram datas inválidas de movimentação.

No tange às observações relacionadas a transferências, é importante informar que no Caged são registradas movimentações entre filiais de um mesmo estabelecimento como um desligamento e uma admissão subsequente nas filiais de origem e destino, respectivamente. Dessa forma, o tempo entre dois eventos de transferência não possui relevância para o estudo aqui realizado. Também foram removidas da análise os registros relacionados a morte e a aposentadoria por não serem objeto do estudo.

Ao final desse processo de limpeza dos dados, foram analisados o saldo anual de contratações, que é a diferença entre o número de admissões e de desligamentos, considerando a região, o sexo, a raça e cor, a faixa etária, a atividade econômica e a escolaridade, conforme detalhado abaixo.

A Figura 3.4 apresenta o comportamento no saldo de contratações nos anos de 2020 e 2021, com o aumento de mais de 600% na região Sudeste. Esse aumento se observa também em relação às demais regiões, o que indica um movimento de recuperação do mercado de trabalho em comparação ao auge da pandemia, que foi no ano de 2020.

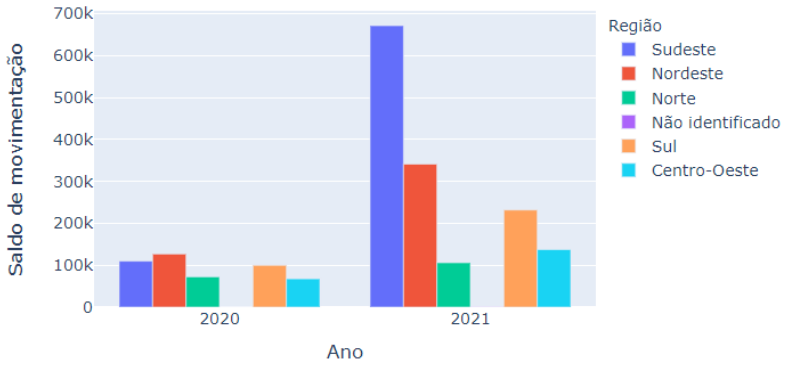

Figura 3.4: Saldo de movimentações nos anos de 2020 e 2021 por região.

No que se refere ao sexo, é possível verificar uma grande diferença no saldo de movimentações de mulheres nos anos de 2020 e 2021, conforme exposto na Figura 3.5. Embora haja uma diferença também no saldo de movimentações de trabalhadores homens, o que correspondeu a um aumento de 75% no ano de 2021 em relação ao ano de 2020, as movimentações de trabalhadoras tiveram um aumento no saldo de movimentações no mesmo período superior a 10 vezes, o que sugere que no auge da pandemia e, nesse contexto, as mulheres sofreram um impacto bem superior.

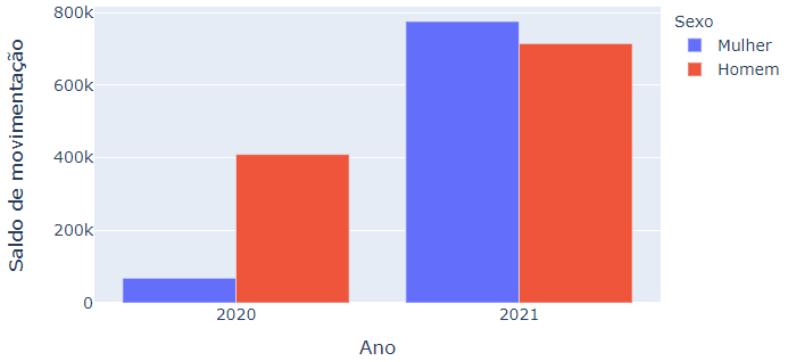

Figura 3.5: Saldo de movimentações nos anos de 2020 e 2021 por sexo.

Já em relação a raça, foi verificado que de 2020 a 2021 a variação absoluta no saldo de movimentações foi superior nos trabalhadores identificados como brancos, seguidos dos pardos e aqueles em que a raça não foi informada, conforme pode ser observado na Figura 3.6. É importante destacar que, embora esse campo seja preenchido pelo empregador, a informação de raça e cor é declarada pelo empregado, o que é impactada pela forma como o trabalhador se identifica.

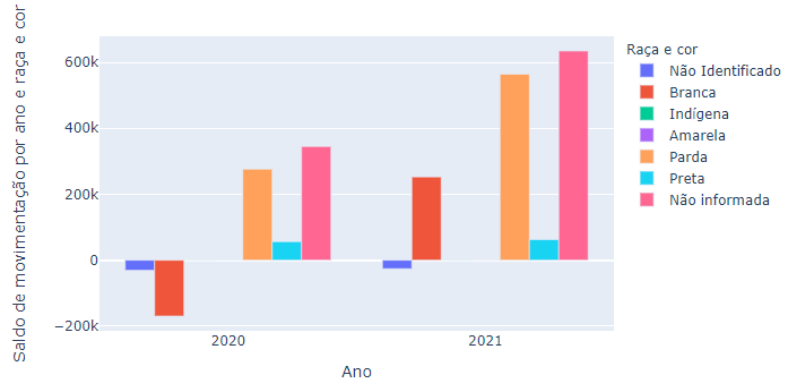

Figura 3.6: Saldo de movimentações nos anos de 2020 e 2021 por raça e cor.

Adicionalmente, os dados do Caged permitiram observar o impacto da pandemia por faixa etária. Nesse sentido, a Figura 3.7 mostra que, apesar da melhora observada no ano de 2021, as faixas etárias acima de 50 anos mantêm saldo negativo de movimentações, ainda que desconsiderados os eventos de aposentadoria ou morte.

No âmbito das atividades econômicas, registradas a partir do Código Nacional de Atividade Econômica (Cnae), os setores com maior impacto foram o de Alojamento e Alimentação, assim como o de Transporte, Armazenamento e Correio, como apresentado na Figura 3.8. Neste contexto, é importante destacar que: 1) uma empresa pode atuar em diferentes atividades econômicas; 2) a atividade econômica registrada no Caged diz

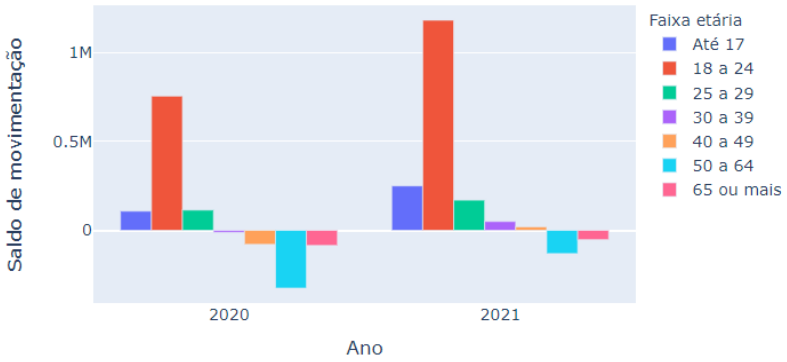

Figura 3.7: Saldo de movimentações nos anos de 2020 e 2021 por faixa etária.

respeito à atuação do trabalhador em determinado vínculo empregatício, sem que haja relação obrigatória com a atividade econômica principal do empregador.

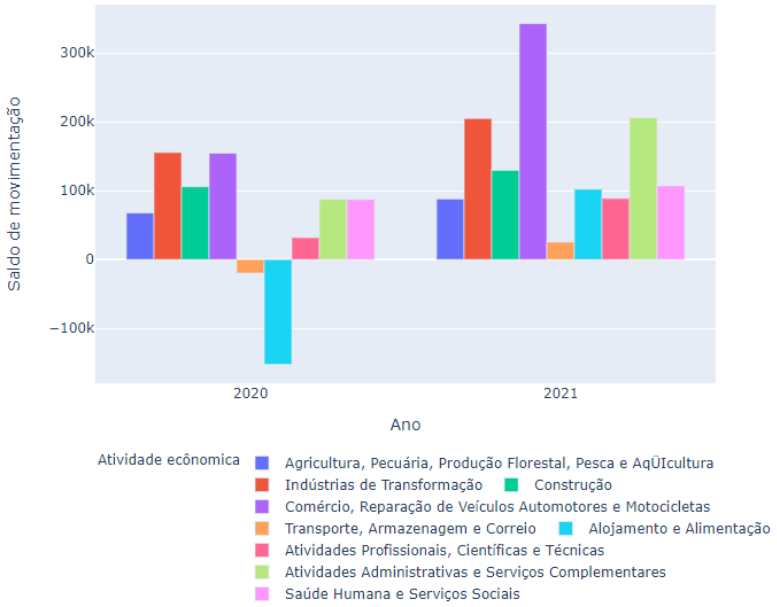

Figura 3.8: Saldo de movimentações nos anos de 2020 e 2021 por setor de atividade econômica.

Por fim, em relação à escolaridade, é possível observar que houve saldo negativo de movimentações em relação aos vínculos em que o empregador indicou como necessário escolaridade inferior ao médio incompleto, possivelmente por envolver atividades menos propensas ao trabalho remoto.

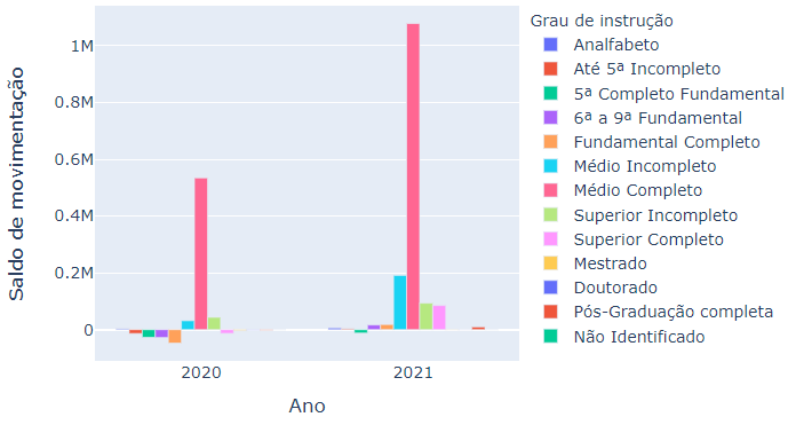

Figura 3.9: Saldo de movimentações nos anos de 2020 e 2021 por escolaridade.

Sistema Nacional de Emprego (Sine)

Os registros do Sine utilizados no presente trabalho se referem aos dados dos trabalhadores que ativaram o seu cadastro na plataforma de intermediação de mão de obra durante os anos de 2020 e 2021. Ao se fazer uma extração da base de dados utilizando esse critério, foram obtidas 5.095.854 observações. Ao se remover os registros com duplicidade de gênero, raça ou data de nascimento, em um procedimento semelhante ao realizado com as bases do Caged e do Rais, o total restante foi de 5.081.543 observações.

Embora o cadastro de um trabalhador tenha sido atualizado na plataforma do Sine, há a possibilidade de que esse trabalhador não tenha realizado nenhuma ação de intermediação de mão de obra e, por uma série de fatores, como por exemplo, quando o sistema não identificou nenhuma vaga de trabalho adequada às experiências e ao perfil profissional do trabalhador, ou quando o trabalhador atualizou as suas informações apenas para requisitar o seguro desemprego, sem interesse em utilizar os demais serviços oferecidos pelo Sine. A Figura 3.10 mostra que o número de trabalhadores que participaram de algum processo seletivo intermediado pelo Sine é oito vezes menor que o daqueles que não utilizaram os serviços de intermediação de mão de obra.

3.1.2 Preparação dos dados

Considerando que apenas cerca de 10% do público do Sine utiliza os serviços de intermediação, conforme observado na Figura 3.10, e tendo em vista que mecanismos de perfilamento estatístico apenas são eficientes na medida em que a informação resultante desses mecanismos é utilizada pelos gestores dos sistemas públicos de emprego na construção de um plano de atendimento individual[12], optou-se neste estudo por utilizar os dados da unidade da federação que apresenta a maior taxa de encaminhamentos para entrevistas.

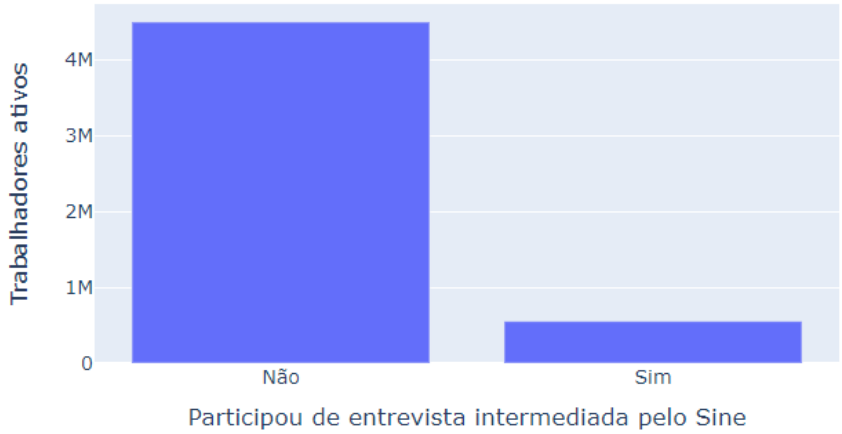

Figura 3.10: Participação dos trabalhadores ativos no Sine em atividades de intermediação de mão de obra mediadas pelo Sine.

A Tabela 3.1, com as informações do percentual por estado dos trabalhadores registrados no Sine que foram encaminhados ou não para entrevistas no período de 2020 a 2021, podemos ver que o Paraná representa mais de 21% dos encaminhamentos do Sistema Público de Emprego brasileiro, o que motivou a utilização dos dados desse estado nas etapas seguintes.

Além da seleção dos dados do Paraná, a preparação da base de dados final envolveu a realização das atividades listadas a seguir:

1. Remoção dos trabalhadores registrados na base do Sine que apresentaram inconsistências de data de nascimento, raça ou gênero retirados das bases do Rais e do Caged subseção anterior;
2. Filtrar as bases do Rais e do Caged de forma a manter apenas observações referentes a pessoas também registradas no Sine;
3. Junção das bases do Sine, do Rais e do Caged e criação do campo de prioridade e preenchimento dele de acordo com as seguintes regras:
   (a) Alta prioridade : aqueles trabalhadores presentes no Sine, mas fora do Rais e do Caged durante o período em análise, ou seja, sem vínculo formal de emprego registrado;
   (b) Média prioridade : trabalhadores que durante o período de análise:
      i. foram contratados após o desligamento do único vínculo empregatício;
      ii. não foram contratados após o registro de desligamento, que ocorreu há um período superior a 12 meses;

Tabela 3.1: Distribuição da participação dos trabalhadores em encaminhamentos por Estado

Estado	Encaminhados	
	Não	Sim
Paraná	5,59%	**21,43%**
São Paulo	27,89%	17,18%
Rio Grande do Sul	5,86%	9,05%
Minas Gerais	10,18%	8,68%
Ceará	2,78%	7,58%
Santa Catarina	4,82%	5,98%
Mato Grosso do Sul	1,43%	4,52%
Mato Grosso	2,06%	4,12%
Bahia	5,16%	3,79%
Goiás	3,62%	2,77%
Rio de Janeiro	8,87%	2,59%
Pernambuco	3,21%	2,19%
Espirito Santo	2,21%	1,77%
Pará	3,13%	1,62%
Distrito Federal	2,17%	1,31%
Paraíba	1,33%	1,28%
Amazonas	1,63%	0,99%
Rondônia	0,81%	0,74%
Tocantins	0,51%	0,54%
Maranhão	1,82%	0,53%
Rio Grande do Norte	1,35%	0,4%
Alagoas	0,89%	0,26%
Roraima	0,28%	0,25%
Sergipe	0,81%	0,21%
Piauí	1,05%	0,18%
Amapá	0,25%	0,04%
Acre	0,28%	0,02%

iii. não foram contratados após o registro de desligamento, que ocorreu há um período inferior a 12 meses;

(c) Baixa prioridade : trabalhadores que durante o período de análise :

i. não apresentaram desligamentos posteriores às admissões;

ii. apresentaram desligamentos em número inferior ao de admissões, ou seja, saldo positivo de vínculos;

Após a preparação da base, foram separados dois conjuntos de dados, o primeiro com todas as observações rotuladas como Média prioridade, e o segundo apenas com as

observações rotuladas como Média prioridade enquadradas nas situações i) e ii) descritas acima.

O primeiro grupo foi utilizado no estudo de análise de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine do Paraná, apresentado a seguir.

Já o segundo grupo foi utilizado no treinamento dos modelos de classificação, etapa apresentada na seção seguinte. A situação (iii) das observações rotuladas como Média prioridade não foram utilizadas no treinamento dos referidos modelos por não ser possível precisar se os trabalhadores precisariam de mais de 12 meses ou de menos tempo para serem recolocados no mercado de trabalho nessas situações, parâmetro objeto do estudo.

Em seguida, foram removidas todas as observações do segundo grupo que apresentaram escolaridade inválida. Além disso, após a seleção de variáveis restaram somente as seguintes variáveis na base do segundo grupo :

1. Faixa etária: categórico;
2. Indicador se o trabalhador possui deficiência: boleano;
3. Indicador se o trabalhador é estrangeiro: boleano;
4. Indicador se o trabalhador é estudante: booleano;
5. Indicador se o trabalhador fala algum idioma além do português: boleano;
6. Raça e cor: categórico;
7. Sexo: categórico;
8. Escolaridade: categórico;
9. CBO do último vínculo de emprego: categórico;
10. Tempo de emprego no último vínculo (meses): inteiro;
11. Seção da Cnae: categórico;

Após a seleção de atributos foi criada a variável boleana "Risco", utilizada como atributo alvo do modelo de classificação. O preenchimento dessa variável foi feito de forma a armazenar o valor "1"quando o trabalhador estiver há mais de 12 meses fora do mercado de trabalho e "0"em caso contrário. Ao final desse processo, restaram 101.346 observações, dentre as quais 50.706 apresentaram "1"na variável risco e 50.640 o valor "0".

Ato contínuo, foi realizada a cópia da base resultante dos passos anteriores seguida do pré-processamento de ambas. A primeira base, utilizada no treinamento dos algoritmos *Logistic Regression*, *Random Forest*, *Gradient Boosting* e *XGBoost*, teve como atividade de

pré-processamento a discretização das variáveis categóricas. Já a segunda base, utilizada no modelo preditivo baseado em Redes Neurais, também teve seus valores categóricos discretizados, com exceção da variável CBO, que passou por uma etapa de transformação em representação vetorial, ou seja, da criação de um *embedding* de códigos de CBO, procedimento descrito abaixo.

Criação da Representação Vetorial do Código de CBO

A criação do *embedding* da variável CBO foi feita a fim de verificar se a capacidade preditiva do modelo objeto deste trabalho seria aprimorada ao se utilizar informações semânticas da última ocupação exercida pelo trabalhador antes do desligamento no treinamento daquele modelo. A fim de obter a representação do *embedding* da CBO, a rede neural representada pela Figura 3.11 foi utilizada.

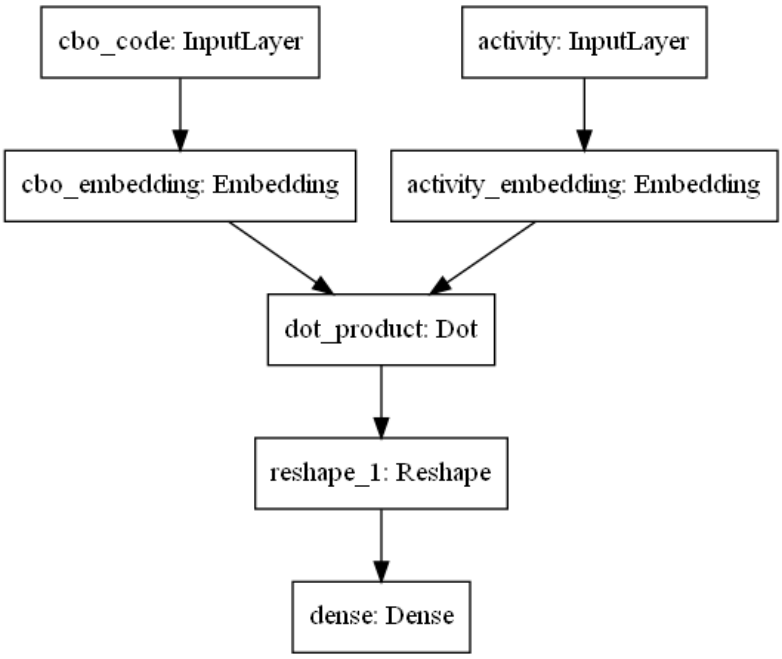

Figura 3.11: Arquitetura da Rede Neural utilizada para a geração do *embedding* da CBO.

Para isso foi criada uma matriz cujo número de linhas é igual ao número de códigos CBO e cujo número de colunas é igual ao número de atividades únicas descritas no perfil ocupacional da CBO. Em seguida, essa matriz foi preenchida da seguinte forma:

- 0 : nas células x, y, quando o código CBO x não prevê a atividade y seu perfil ocupacional e;
- 1 : nas células x, y, quando o código CBO x prevê a atividade y seu perfil ocupacional.

O treinamento da Rede Neural descrita na Figura 3.11 foi então realizado percorrendo todas as células da matriz de forma que as entradas do algoritmo são o código de CBO e o código da atividade, representando respectivamente as linhas e as colunas da matriz. Já a saída foi o valor armazenado na célula, ou seja, 0 ou 1. Ao final desse processo, nas situações em que dois códigos possuíam muitas atividades em comum, a distância dos vetores que representam esses códigos na camada *cbo_embedding* códigos foi pequena. De forma contrária, a distância foi grande quando houve poucas ou nenhuma atividade em comum.

Análise de Sobrevivência do tempo para a reinserção no mercado formal de trabalho do público do Sine-Paraná

A fim de analisar isoladamente o efeito das variáveis da base resultante da preparação de dados no tempo para a recolocação no mercado de trabalho, foram utilizadas técnicas de Análise de Sobrevivência sobre alguns dos atributos, conforme apresentado abaixo.

Inicialmente, foi analisado o comportamento do tempo em desemprego dos trabalhadores após o desligamento nas diferentes mesorregiões do Paraná. A Figura 3.12 mostra que é esperado um menor tempo para a recolocação no mercado de trabalho no Sudoeste Paranaense em comparação com as outras mesorregiões do Paraná. Já a messorregião que apresentou o maior tempo de recolocação foi o Noroeste Paranaense.

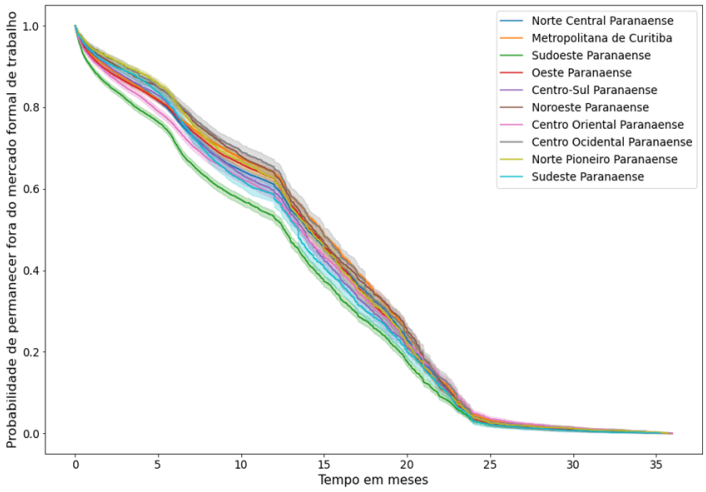

Figura 3.12: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por mesorregião.

A fim de garantir maior clareza para a visualização das informações, as informações do tempo de desemprego no Paraná por nível de escolaridade foram divididas distribuídas nas duas figuras abaixo.

Nesse sentido, a Figura 3.13 apresenta as curvas relacionadas aos primeiros anos de escolaridade, ou seja, com formação até o nível fundamental completo. Além disso, situações em que o cadastro do trabalhador apresentou dados inválidos de escolaridade também foram plotados no gráfico. Pela análise dessa informação, ao descartar as situações com escolaridade inválida, pode-se verificar que os trabalhadores com ensino fundamental incompleto e o fundamental completo conseguem se recolocar no mercado de trabalho em tempo inferior aos analfabetos e com escolaridade até o 5º ano incompleto.

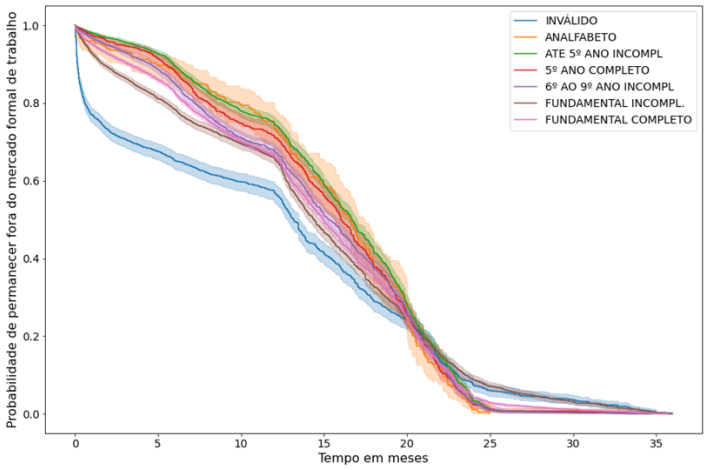

Figura 3.13: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por escolaridade nos primeiros anos de ensino.

Já em relação a trabalhadores com nível de escolaridade igual ou superior ao nível médio incompleto, cujas curvas são apresentadas na Figura 3.14, aqueles com nível médio completo ou superior incompleto são os que necessitam de menor tempo de recolocação, enquanto os com superior completo, mestrado e doutorado são os que precisam de mais tempo para a recolocação.

É importante destacar que, em função do pouco número de observações, os trabalhadores com mestrado e doutorado são os que apresentam um maior intervalo de confiança para o tempo até a recolocação, o que é representado pelas faixas com grande amplitude em torno das curvas. Isso significa que para esses grupos há uma menor precisão em relação aos valores esperados para o tempo em desemprego. Uma possível explicação para o baixo número de observações para esses grupos é que o Sine possui uma concen-

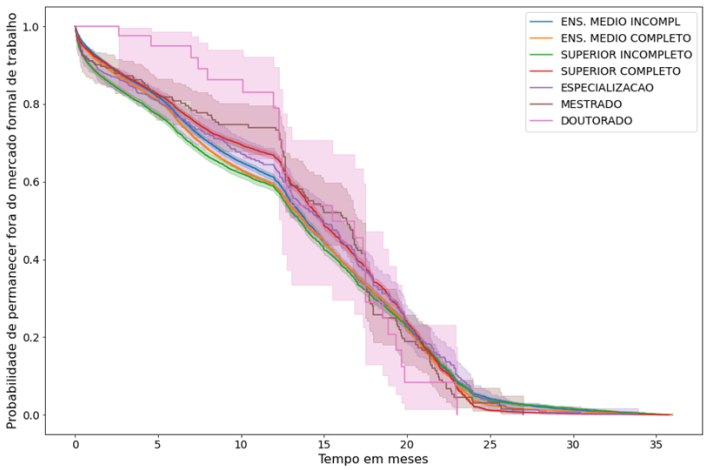

Figura 3.14: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por escolaridade a partir do ensino médio.

tração de vagas que exigem menor escolaridade, o que resulta em um menor interesse de trabalhadores com maior qualificação.

A Figura 3.15 mostra que os trabalhadores registrados no Sine no estado do Paraná do gênero masculino conseguem se recolocar no mercado de trabalho em menos tempo que aqueles identificados como do gênero feminino.

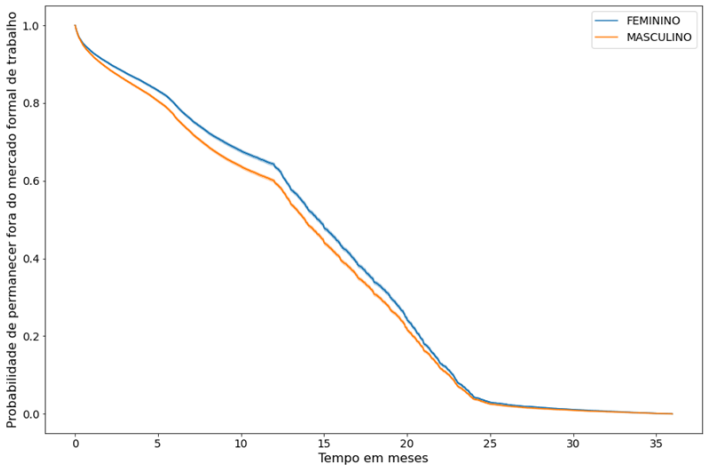

Figura 3.15: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por gênero.

Ao se analisar a dimensão faixa etária, a Figura 3.16 mostra que, com exceção da faixa

de 15 a 17 anos, o tempo para a recolocação profissional dos trabalhadores registrados no Sine do Paraná é ordenado de forma que as menores idades apresentem menor tempo para a recolocação profissional. A amplitude do intervalo de confiança na faixa acima de 65 anos pode ser explicado pela proximidade dessa faixa com a idade de aposentadoria, o que resulta em um menor interesse na busca por empregos formais.

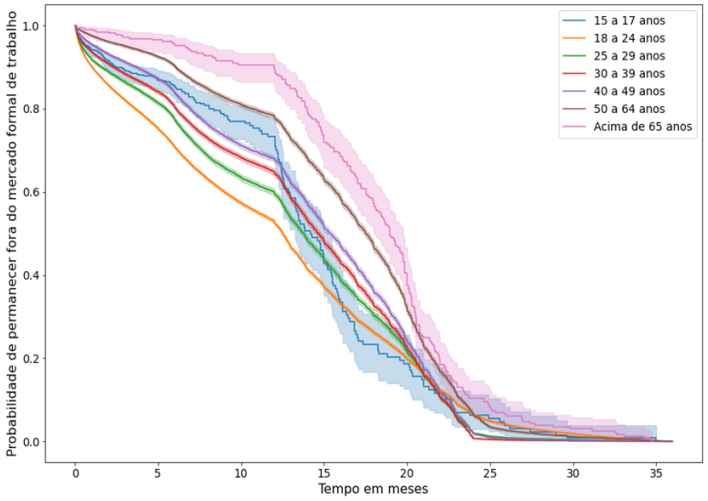

Figura 3.16: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por faixa etária.

No que tange ao atributo de raça e cor, a Figura 3.17 mostra que os trabalhadores que se identificaram como brancos e pardos na região em análise tiveram um menor tempo para a recolocação profissional em comparação com os demais. Trabalhadores indígenas e aqueles que se identificaram como de raça amarela apresentaram maior tempo para a recolocação, embora tenham sido também os grupos com resultados mais inconclusivos, conforme observado pela amplitude do intervalo de confiança.

Em relação ao tempo no emprego anterior, ou seja, o tempo em que o trabalhador permaneceu em atividade no último vínculo antes do desligamento, observou-se que os trabalhadores que conseguiram se recolocar no mercado de trabalho mais rapidamente foram aqueles que exerceram as atividades no último vínculo de emprego por menos tempo, conforme observado na Figura 3.18.

Além da avaliação das características dos trabalhadores registrados no Sine do Paraná, utilizou-se a análise de sobrevivência para verificar se há diferença no tempo para a recolocação profissional dos trabalhadores que utilizaram os serviços de intermediação do Sine, ou seja, os que foram encaminhados para pelo menos uma entrevista intermediada pelo Sine, daqueles trabalhadores que não utilizaram os serviços de intermediação,

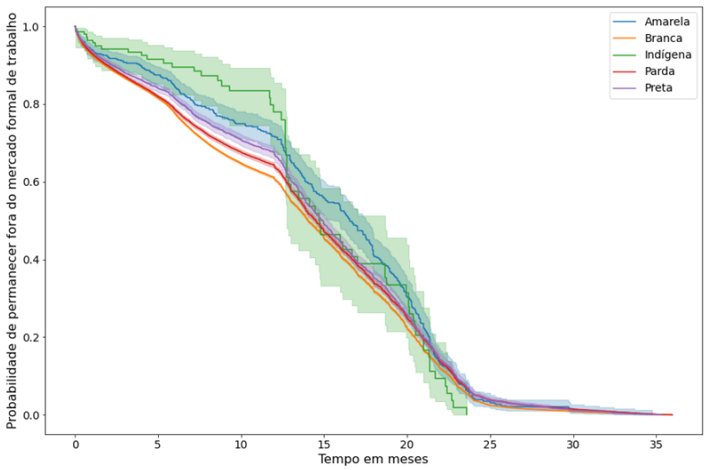

Figura 3.17: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por raça e cor.

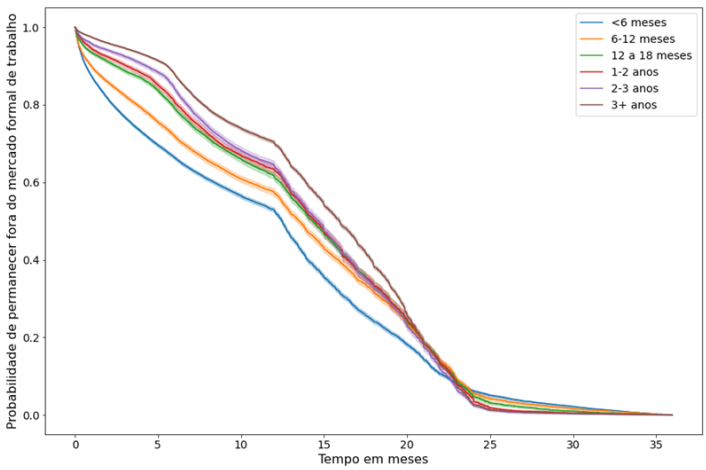

Figura 3.18: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná por faixa etária.

conforme observado na Figura 3.19. A partir da figura é possível observar que, considerando os primeiros 12 meses após a recolocação, um trabalhador registrado no Sine do Paraná que utilizou os serviços de intermediação de mão de obra, precisou de 41 dias (30 x 1,383) a menos em média do que um trabalhador nas mesmas condições mas que não foi intermediado pelo Sine. Contudo, convém destacar que essa diferença de tempo pode se dever a um viés de seleção, ou seja, os trabalhadores que estão mais interessados

a retornarem ao mercado de trabalho tendem a ser os que estão mais propensos utilizar serviços de recolocação profissional, como os de intermediação de mão de obra oferecidos pelo Sine.

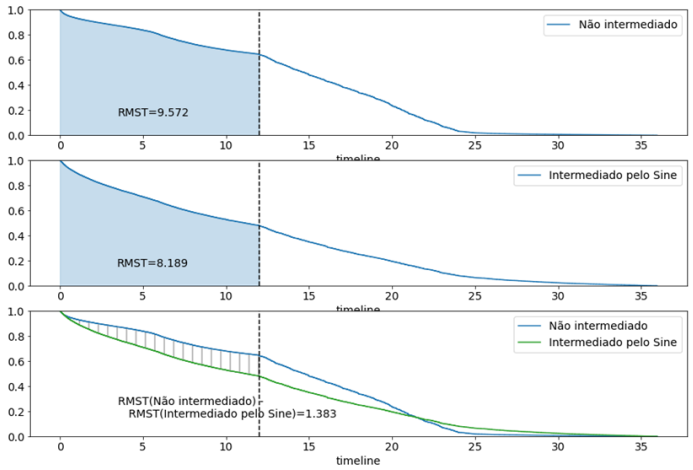

Figura 3.19: Curva de sobrevivência do tempo para a recolocação no mercado de trabalho dos trabalhadores registrados no Sine no Paraná que utilizaram ou não os serviços de intermediação de mão de obra.

3.2 Modelagem e Validação dos Modelos

Esta seção detalha o treinamento e o teste dos cinco modelos avaliados no presente trabalho. Essa fase se dividiu em duas iterações onde foi utilizada a linguagem de programação *Python*[1] na construção de todos os modelos.

Na primeira iteração, inicialmente foi treinado o modelo preditivo *Logistic Regression* (LR), cuja performance foi utilizada como base de comparação com os demais modelos. Em seguida, foram treinados os modelos *Random Forest* (RF), *Gradient Boosting* (GBT) e *XGBoost*, cujos desempenhos foram comparados com a linha base. Foi utilizada a biblioteca *scikit-learn*[2] na construção dos três primeiros modelos (LR, RF e GBT) e a biblioteca *xgboost*[3] na construção do último (*XGBoost*).

O treinamento da primeira iteração utilizou uma amostra aleatória estratificada de 20.000 observações, onde 80% foram separadas para treinamento e 20% para testes. Os estratos foram baseados no risco do trabalhador se tornar um desempregado de longa

[1]Disponível em https://www.python.org
[2]Disponível em https://scikit-learn.org
[3]Disponível em https://xgboost.readthedocs.io/en/stable/index.html

duração. Posteriormente, utilizou-se o *Grid Search* para avaliar a performance, medida pelo AUC, de diferentes combinações de parâmetros, conforme apresentado na Tabela 3.2.

Tabela 3.2: Parâmetros avaliados por *Grid Search* dos algoritmos utilizados na primeira iteração

Algoritmo	**Parâmetro**	**Valores testados**	**Modelo Final**
LR	C	0.01, 0.1, 1, 10, 100	0.1
	max_iter	100, 500	100
	penalty	l1, l2	l2
RF	n_estimators	10, 100, 500	500
	max_features	sqrt, log2, None	None
	max_depth	5, 10, 20	10
	class_weight	None, 0:1,1:5, 0:1,1:10, 0:1,1:25	None
GBT	n_estimators	10, 50, 100, 250	250
	max_depth	5, 7, 9, 11, 13, 15	15
	min_samples_split	range(200,1001,200)	200
	min_samples_leaf	30, 40, 50,60,70	30
	max_features	7, 9, 11, 13, 15, 17, 19	19
	subsample	0.6, 0.7, 0.75, 0.8, 0.85, 0.9	0.9
XGBoost	booster	gbtree, gblinear	gbtree
	eta	0.01, 0.2,0.3	0.01
	colsample_bytree	0.5, 0.75, 1	0.5
	objective	reg:squarederror, binary:logistic	binary:logistic
	eval_metric	error, auc	error

A Figura 3.20 apresenta a performance dos modelos a partir da combinação dos melhores parâmetros. É possível observar que, embora a Regressão Logística tenha apresentado a maior amplitude no ganho de desempenho a partir do ajuste de parâmetros, o XGBoost obteve resultado superior em relação a todos os modelos testados na primeira iteração.

Na segunda iteração, o treinamento foi realizado com apenas dois modelos, o melhor modelo da etapa anterior, o XGBoost, além de um modelo baseado em Rede Neural, em que os dados passaram por uma etapa adicional de transformação dos códigos CBO. Nessa etapa, os códigos CBO foram substituídos por sua representação vetorial, conforme descrito anteriormente.

Além disso, diferente da amostra de 20.000 observações da primeira iteração, essa etapa utilizou dados resultantes de uma amostragem aleatória estratificada de 2.000 observações. Os estratos foram baseados no risco do trabalhador se tornar um desempregado de longa duração. O tamanho da amostra foi menor que na iteração 1 em função do alto tempo para treinamento da rede neural. Já em relação à proporção utilizada para treinamento e testes, foram mantidos os mesmos percentuais, ou seja, 80% e e 20% para treinamento e testes, respectivamente.

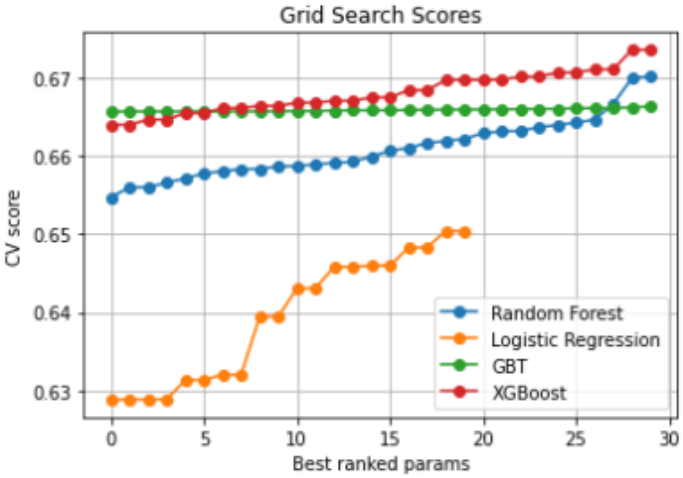

Figura 3.20: Curvas com os resultados dos melhores parâmetros dos modelos testados na primeira iteração.

Semelhante ao que foi feito na primeira iteração, foram testadas várias combinações de parâmetros a fim de identificar aquela cuja utilização resulta na melhor capacidade preditiva do modelo. A Tabela 3.3 apresenta os parâmetros utilizados no treinamento da rede neural, assim como os melhores valores de parâmetros encontrados.

Tabela 3.3: Parâmetros avaliados por *Grid Search* do algoritmo utilizado na segunda iteração

Algoritmo	**Parâmetro**	**Valores testados**	**Modelo Final**
Rede Neural	batch_size	10, 20	20
	epochs	10, 50	10
	optimizer	SGD, RMSprop	RMSprop
	init_mode	uniform, lecun_uniform	lecun_uniform
	activation	softmax, softplus	softplus
	dropout_rate	0.0, 0.1, 0.2	0.2
	neurons	1, 5, 10	5

Por fim, foram comparados os desempenhos dos algoritmos baseados em XGBoost e em Rede Neural. Os resultados com diferentes combinações de parâmetros podem ser observados na Figura 3.21, onde é possível verificar a melhor performance do XGBoost em relação à rede neural no caso em análise.

Uma vez escolhido o modelo com melhor capacidade preditiva, foi necessário garantir que as características que mais impactaram na classificação do trabalhador em alto ou baixo risco fossem interpretadas corretamente pelos gestores regionais do Sine. Isso porque, a partir da identificação das características relevantes para a classificação de um

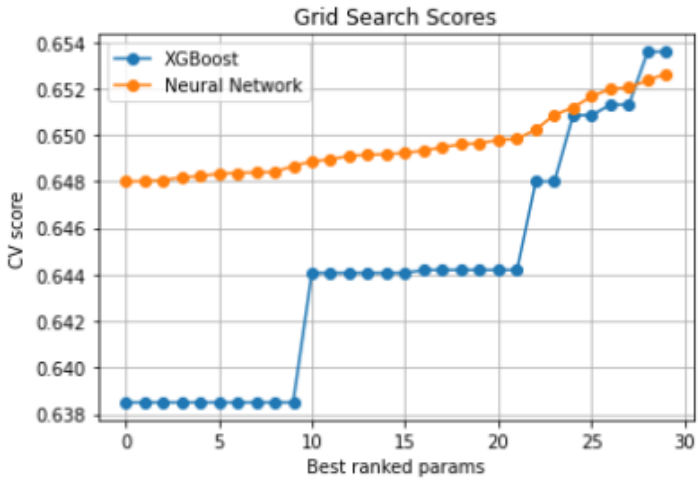

Figura 3.21: Curvas com os resultados dos melhores parâmetros dos modelos testados na segunda iteração.

trabalhador como de alto risco, foi possível planejar ações mais adequadas à mitigação ou redução dos problemas encontrados por esse trabalhador.

Por este motivo, utilizou-se implementação do Shap em python[4] para explicar o modelo estatístico selecionado na etapa anterior. A seguir, serão apresentados os resultados de análises dos fatores mais importantes no caso geral e em observações específicas.

A Figura 3.22 apresenta a relevância dos principais atributos na classificação de risco de trabalhadores. Cores mais próximas do vermelho indicam uma maior probabilidade de um indivíduo ser classificado como de alto risco. Já valores mais próximos do azul indicam maior potencial da classe de baixo risco ser escolhida.

Considerando que os atributos utilizados neste trabalho são categóricos, o gráfico deve ser interpretado de forma que, valores acima de zero indicam que a observação possui aquela característica, enquanto valores abaixo de zero significam a ausência dessa característica. Como exemplo, ao analisar a faixa etária de 18 a 24 anos, pode-se verificar que indivíduos nessa faixa de idade possuem menor probabilidade de serem considerados de alto risco, o que pode ser concluído a partir da observação da concentração de pontos azuis acima de zero, ou seja, quando a observação possui essa característica. Por outro lado, trabalhadores na faixa etária de 50 a 64 anos possuem uma maior probabilidade de ser considerados de alto risco, como observado pelos pontos vermelhos para valores acima de zero.

No que tange a observações individuais, a Figura 3.23 apresenta os fatores mais relevantes para que uma dada observação tenha sido classificada como de alto risco. Os atributos são apresentados em ordem de importância, reforçando a classificação, atribu-

[4]Disponível em https://shap.readthedocs.io/en/latest/

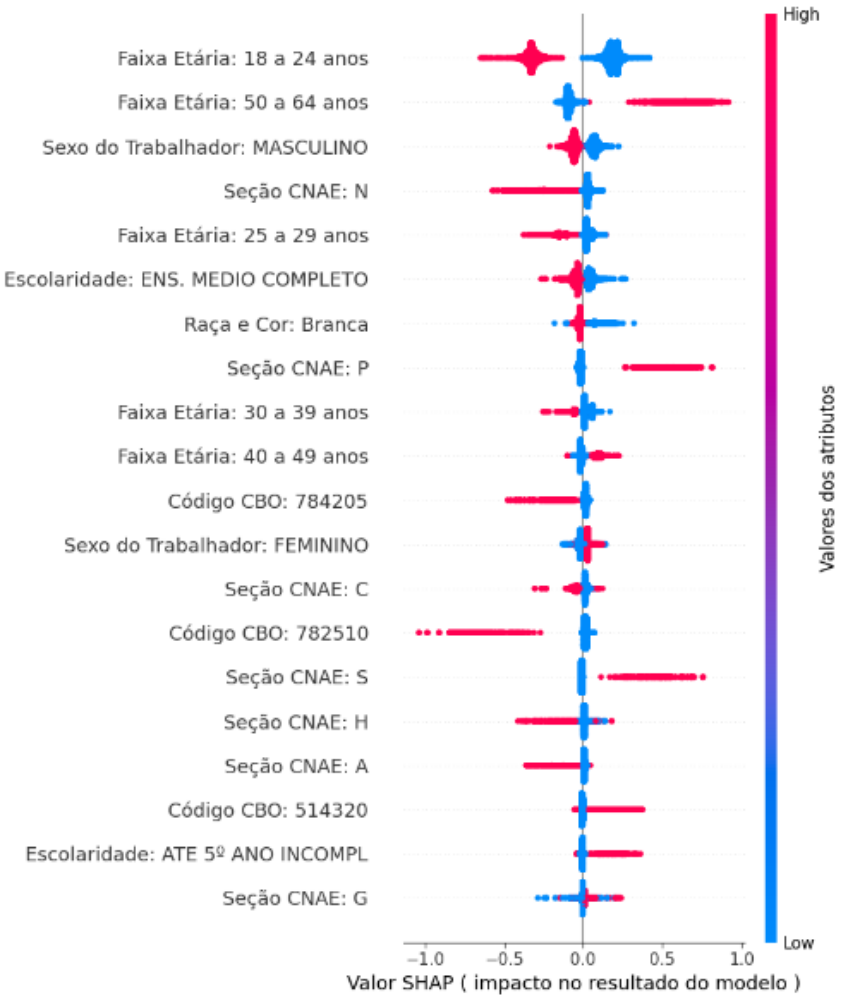

Figura 3.22: Valores SHAP dos atributos mais relevantes na classificação de risco de trabalhadores.

tos seguidos de uma barra vermelha, ou enfraquecendo-a, atributos seguidos de uma barra azul. No exemplo em análise a variável faixa etária foi o fator de maior importância para a classificação de alto risco, seguido do sexo feminino.

Já a Figura 3.24 apresenta um exemplo de observação classificada de baixo risco. Nessa situação, o indivíduo apresentou a faixa etária, entre 18 e 24 anos, como principal fator para a escolha dessa classe, seguida do código CBO da atividade que o trabalhador exercia no momento do último desligamento, ou seja, o 717020, que identifica os ajudantes de obras civis.

Embora exista poucos dados disponíveis relacionados à acurácia dos modelos de perfilamento estatístico utilizados em outros países, os resultados obtidos no presente trabalho são compatíveis com experiências internacionais já realizadas, conforme as informações disponíveis [48].

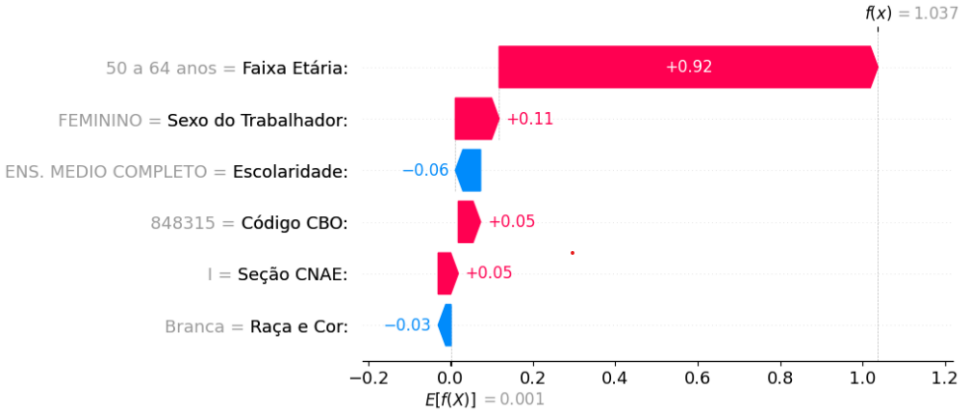

Figura 3.23: Atributos mais relevantes para a classificação do indivíduo como de alto risco.

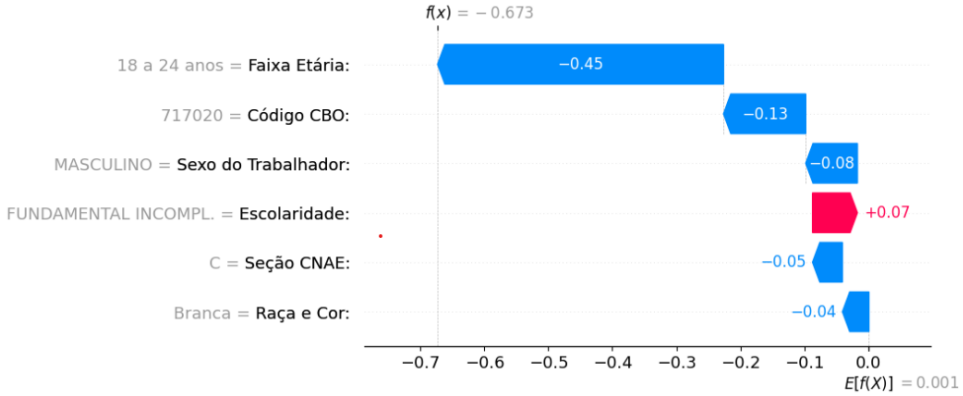

Figura 3.24: Atributos mais relevantes para a classificação do indivíduo como de baixo risco.

No que tange à interpretabilidade dos modelos, considerando a necessidade de utilização dos resultados do modelo de forma a subsidiar o trabalho nos postos do Sine e permitir uma avaliação mais robusta durante o atendimento aos trabalhadores, entende-se que as visualizações proporcionadas pela utilização das técnicas do Shap são adequadas. Isso permitirá que intervenções sobre os trabalhadores com maior risco de permanecerem fora do mercado de trabalho sejam mais focalizadas, garantindo, dessa forma, que os gastos com qualificação e orientação profissional, além da intermediação de mão de obra, sejam mais efetivos.

Ademais, considerando que apenas no ano de 2022 foram gastos cerca de R$ 35 bilhões em parcelas do seguro-desemprego, a redução do número de parcelas pagas em função da recolocação antecipada dos trabalhadores segurados possui um grande potencial de economia de recursos públicos.

Capítulo 4

Conclusões e trabalhos futuros

Este capítulo apresenta as conclusões, os resultados obtidos e as sugestões de trabalhos futuros que poderão ser realizados de forma a aprimorar os mecanismos utilizados nesta pesquisa.

4.1 Conclusões

Este trabalho foi realizado com o objetivo de auxiliar os gestores do Sistema Nacional de Emprego (Sine) na identificação de trabalhadores que apresentem maiores dificuldades para a reinserção no mercado formal de trabalho. Utilizou-se como critério o risco desses trabalhadores se tornarem desempregados de longa duração, ou seja, de precisarem de um período superior a de 12 meses para a recolocação profissional. A fim atender a essa necessidade, foram estabelecidos objetivos específicos que nortearam o objetivo final do trabalho.

Com este propósito, foi criado um modelo de perfilamento de trabalhadores baseado em aprendizado de máquina, para a separação de trabalhadores em grupos utilizando como critério de agregação o risco desses trabalhadores permanecerem fora do mercado formal de trabalho por um período superior a 12 meses. Com isso, foi possível atingir o objetivo específico 1 deste trabalho, qual seja, detectar trabalhadores cujas estimativas de tempo em desemprego sejam maiores que 12 meses.

No que tange ao objetivo específico 2, referente à identificação de caraterísticas de trabalhadores que mais impactam na duração do tempo em desemprego, foi treinado um explicador baseado no Shap. O algoritimo utilizado permitiu a avaliação dos atributos mais importantes para o tempo em desemprego, seja em relação a toda a população, seja a respeito de observações individuais.

Além disso, o explicador permitiu uma melhor interpretação dos motivos pelos quais o algoritmo de classificação atribuiu uma determinada classe a uma observação, o que facilitou a utilização do mecanismo nos postos do Sine.

Por fim, considerando a diferença na dinâmica do mercado de trabalho formal nas diferentes regiões do Brasil, optou-se por selecionar as informações, ou seja, os dados referentes aos trabalhadores de um Estado do Brasil, o estado do Paraná, antes do treinamento dos algoritmos. Ademais, foi realizado, sobre a população do Sine desse estado, um estudo de análise de sobrevivência do tempo em desemprego a fim de se entender melhor os diferentes atributos que impactam sobre o tempo em desemprego. Com isso, foi alcançado o objetivo específico 3 do trabalho, que buscou indicar os fatores mais relevantes para o tempo em desemprego por região.

4.2 Resultados Obtidos

Com a realização deste estudo foi possível obter alguns resultados principais e que impactam em setores públicos que trabalham com a empregabilidade de trabalhadores, indicados a seguir:

1. A criação de uma metodologia para a limpeza, preparação e integração das bases da Rais, Caged e Sine.
2. A utilização de técnicas de aprendizado de máquina na identificação de trabalhadores com maior risco de se tornarem desempregados de longa duração sobre a base de dados do Sine.
3. A criação de um mecanismo de fácil interpretação para a identificação das características mais relevantes para o tempo em desemprego, tanto de uma população quanto de observações específicas, em determinada região.

Embora os resultados do modelo estatístico de perfilamento proposto neste trabalho possam ser considerados em uma primeira análise aceitáveis, eles são compatíveis com experiências de outros países.

Além disso, a identificação antecipada dos trabalhadores que possuem maior probabilidade de permanecerem fora do mercado formal de trabalho permite a realização de uma série de medidas para a redução do tempo em desemprego. Desta forma, há a possibilidade da redução de gastos com o seguro-desemprego, que foram em torno de R$ 35 bilhões no ano de 2022[53].

4.3 Trabalhos Futuros

Considerando a melhoria da capacidade preditiva do algoritmo, adicionar ao treinamento do modelo de aprendizado de máquina informações relacionadas a mudanças de ocupação, setor da economia e residência podem permitir, além de otimizar a performance do algoritmo, indicar para os trabalhadores, além dos gestores da política pública de emprego, ganhos em relação a essas transições.

Ainda sobre o aprimoramento da capacidade preditiva do algoritmo de perfilamento, estudos adicionais podem utilizar, além das bases de dados da Rais, do Sine e do Caged, informações do programa Bolsa Família[54] e do Seguro Desemprego[55]. Essas bases, além de terem o potencial de melhorar a identificação dos grupos com risco de permanecerem mais tempo fora do mercado formal de trabalho, possibilitam a utilização adicional de atributos relacionados a grupos prioritários, como quilombolas. Dessa forma, o uso dessas informações tem grande potencial de ampliar a abrangência social do mecanismo de perfilamento.

Ademais, estudos adicionais podem vincular os atributos mais relacionados ao desemprego de longa duração a serviços públicos que possuem maior probabilidade de reduzir a relevância dessas características. Por exemplo, ofertando requalificação profissional a trabalhadores em que a ocupação anterior foi indicada como fator de risco.

Informações sobre tendências futuras do mercado de trabalho, seja como resultante de estudos de análise preditiva sobre as bases do Caged, seja como resultado da ampliação planejada do mercado de trabalho, como por exemplo a instalação de um parque industrial em determinada região e em um momento futuro, podem resultar em grande potencial de melhoria da capacidade preditiva do algoritmo de perfilamento. Isso porque admissões ou desligamentos em momentos de aumento de contratações ou de desligamentos possuem contextos opostos.

Referências

[1] Ramos, Jorge: *Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional*. Tese de Doutoramento, dezembro 2016. ix, 7

[2] Semaan, Gustavo, Germano Ferraz, Rodrigo Erthal Wilson, Débora Alvernaz e Jose Brito: *Uma análise do acervo da revista produção online*. Revista Produção Online, 20:1279–1300, dezembro 2020. ix, 11

[3] Tan, Pang Ning, Michael Steinbach, Anuj Karpatne e Vipin Kumar: *Introduction to Data Mining*. Pearson, 2018. ix, 11, 12, 15, 16, 17

[4] Zou, Kelly H., A. James O'Malley e Laura Mauri: *Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models*. Circulation, 115(5):654–657, 2007. `https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.105.594929`. ix, 18

[5] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. `https://faculty.marshall.usc.edu/gareth-james/ISL/`. ix, 18, 19, 20, 22

[6] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. ix, 21

[7] Aggarwal, Charu C.: *Neural Networks and Deep Learning*. Springer, Cham, 2018, ISBN 978-3-319-94462-3. ix, 23

[8] Ribeiro, Marco Tulio, Sameer Singh e Carlos Guestrin: *Why should i trust you? explaining the predictions of any classifier*. Em *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, páginas 1135–1144, 2016. ix, 24, 25

[9] Hasluck, Chris: *The use of statistical profiling for targeting employment services: The international experience*. New European Approaches to Long-Term Unemployment: What Role for Public Employment Services and What Market for Private Stakeholders?, 2008. 1, 2

[10] Bacchetta, Marc, Ekkehard Ernst e Juana P. Bustamante: *Globalization and informal jobs in developing countries*. WTO, janeiro 2009, ISBN 9789287045409. 1

[11] OECD, Inter American Development Bank e World Association of Public Employment Services: *The World of Public Employment Services*. IDB, Washington, D.C., 2016. https://www.oecd-ilibrary.org/content/publication/9789264251854-en. 1, 28

[12] Gazier, Bernard: *Employability—the complexity of a policy notion*. Em *Employability: From theory to practice*, páginas 3–24. Routledge, New York, 2017. 2, 3, 36

[13] Rudolph, Helmut: *Profiling as an instrument for early identification of people at risk of long-term unemployment*. Em *Employability: From theory to practice*, páginas 25–50. Routledge, New York, 2017. 3, 4

[14] Silva Costa, Simone da: *Pandemia e desemprego no brasil*. Revista de Administração Pública, 54, agosto 2020, ISSN 1982-3134. 3, 5

[15] Ruiz, Angélica Aparecida Parreira Lemos, Maurício Augusto de Souza Ruiz, Angela Maria Grossi e Juliano Maurício de Carvalho: *Pandemia covid-19 e a aceleração da transformação digital nos serviços públicos: uma proposta de intervenção cidadã unesp prep@ara*. Mídia, cultura inovativa e economia criativa em tempos pandêmicos, 2020. 3

[16] Presidência da República: *Decreto 8.373/2014*, 2014. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/decreto/d8373.htm. 3

[17] Governo Federal Brasileiro: *Conheça o esocial*, 2019. https://www.gov.br/esocial/pt-br/centrais-de-conteudo, acesso em 19/03/2022. 3

[18] Presidência da República: *Decreto no 10854/2021*, 2021. http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Decreto/D10854.htm#art187. 3

[19] Presidência da República: *Lei no 4.923/1965*, 1965. http://www.planalto.gov.br/ccivil_03/leis/l4923.htm. 3

[20] Lipp, Robert: *Job seeker profiling: The australian experience*. Em *EU-Profiling Seminar*, 2005. 4, 27

[21] O'Connell, Philip, Seamus Mcguinness e Elish Kelly: *A statistical profiling model of long-term unemployment risk in ireland*. Economic and Social Research Institute (ESRI), Papers, janeiro 2010. 5, 28

[22] Mattei, Lauro e Vicente Loeblein Heinen: *Impactos da crise da covid-19 no mercado de trabalho brasileiro*. Brazilian Journal of Political Economy, 40, dezembro 2020, ISSN 1809-4538. 5

[23] Organisation for Economic Co-operation and Development: *Long-term unemployment rate*, 1999. https://data.oecd.org/unemp/long-term-unemployment-rate.htm, acesso em 19/03/2022. 5

[24] Heerdt, Mauri Luiz e Vilson Leonel: *Metodologia científica e da pesquisa: livro didático*, volume 7. Palhoça - UnisulVirtual, Santa Catarina, 5ª edição, 2007, ISBN 9788522458233. 6

[25] Wazlawick, Raul Sidnei: *Metodologia de Pesquisa para Ciência da Computação*. Elsevier, second edição, 2014, ISBN 978-85-352-7782-1. https://www.sciencedirect.com/book/9788535277821. 6

[26] Gil, Antonio Carlos: *Como elaborar projetos de pesquisa*. Atlas, São Paulo, 2010, ISBN 9788522458233. 6

[27] Universidade FUMEC: *A ciência e seus métodos*, 2016. http://ppg.fumec.br/ecc/wp-content/uploads/2016/12/MetodCientifica_02.pdf, acesso em 22/03/2022. 6

[28] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer e Rudiger Wirth: *Crisp-dm 1.0 step-by-step data mining guide*. Relatório Técnico, The CRISP-DM consortium, August 2000. https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf. 7

[29] Elias, Peter: *Occupational Classification (ISCO-88): Concepts, Methods, Reliability, Validity and Cross-National Comparability*. Oecd labour market and social policy occasional papers 20, OECD Publishing, janeiro 1997. https://ideas.repec.org/p/oec/elsaaa/20-en.html. 9

[30] Geron, Aurelien: *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Sebastopol, CA, 2017, ISBN 978-1491962299. 13, 16

[31] Guo, Cheng e Felix Berkhahn: *Entity embeddings of categorical variables*, 2016. https://arxiv.org/abs/1604.06737. 13

[32] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent e Christian Jauvin: *A neural probabilistic language model*. JOURNAL OF MACHINE LEARNING RESEARCH, 3:1137–1155, 2003. 14

[33] Kleinbaum, David G. e Mitchel Klein: *Survival Analysis : A self learning book*. Springer Science and Business Media, LLC, New York, 2012. 14

[34] Ciuca, Vasilica e Monica Matei: *Survival analysis for the unemployment duration*. Em *Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation*, volume 1, páginas 354–359, 2010. 14

[35] Witten, Ian H., Eibe Frank e Mark A. Hall: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3ª edição, 2011, ISBN 978-0-12-374856-0. http://www.sciencedirect.com/science/book/9780123748560. 15

[36] Bradley, Andrew P.: *The use of the area under the roc curve in the evaluation of machine learning algorithms.* Pattern Recognition, 30(7):1145–1159, 1997, ISSN 0031-3203. https://www.sciencedirect.com/science/article/pii/S0031320396001422. 17

[37] Breiman, Leo: *Random forests.* Machine Learning, 45(1):5–32, 2001, ISSN 0885-6125. 21

[38] Friedman, Jerome H.: *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 29:1189–1232, 2000. 22

[39] Chen, Tianqi e Carlos Guestrin: *Xgboost: A scalable tree boosting system.* Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, página 785–794, New York, NY, USA, 2016. Association for Computing Machinery, ISBN 9781450342322. https://doi.org/10.1145/2939672.2939785. 22

[40] Lundberg, Scott M. e Su In Lee: *A unified approach to interpreting model predictions.* Em *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, página 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc., ISBN 9781510860964. 24

[41] Bhattacharya, A.: *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more.* Packt Publishing, 2022, ISBN 9781803234168. https://books.google.com.br/books?id=Kal3EAAAQBAJ. 25

[42] Choi, Jin A e Kiho Lim: *Identifying machine learning techniques for classification of target advertising.* ICT Express, 6(3):175–180, 2020, ISSN 2405-9595. https://www.sciencedirect.com/science/article/pii/S2405959520301090. 25

[43] Lo, Siaw, David Cornforth e Raymond Chiong: *Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on twitter.* dezembro 2014, ISBN 978-3-319-14062-9. 25

[44] Dabbah, Mohammad A., Angus B. Reed, Adam T. C. Booth, Arrash Yassaee, Aleksa Despotovic, Benjamin Klasmer, Emily Binning, Mert Aral, David Plans, Davide Morelli, Alain B. Labrique e Diwakar Mohan: *Machine learning approach to dynamic risk modeling of mortality in covid-19: a uk biobank study.* Scientific Reports, 11:16936, dezembro 2021, ISSN 2045-2322. 26

[45] Kuno, Toshiki, Takahisa Mikami, Yuki Sahashi, Yohei Numasawa, Masahiro Suzuki, Shigetaka Noma, Keiichi Fukuda e Shun Kohsaka: *Machine learning prediction model of acute kidney injury after percutaneous coronary intervention.* Scientific Reports, 12:749, dezembro 2022, ISSN 2045-2322. 26

[46] Kim, Seungwook, Daeyoung Choi, Eunjung Lee e Wonjong Rhee: *Churn prediction of mobile and online casual games using play log data.* PLOS ONE, 12:e0180735, julho 2017, ISSN 1932-6203. 26

[47] Serrano, Emilio, Mari Carmen Suárez-Figueroa, Jacinto González-Pachón e Asunción Gómez-Pérez: *Toward proactive social inclusion powered by machine learning*. Knowledge and Information Systems, 58:651–667, março 2019, ISSN 0219-1377. 26

[48] Desiere, Sam, Kristine Langenbucher e Ludo Struyven: *Statistical profiling in public employment services: An international comparison*. 2019. 26, 50

[49] Eberts, Randall W. e Christopher J. O'Leary: *Design of the wprs system and evaluation in michigan*. Em Balducchi, David E. (editor): *Worker Profiling and Reemployment Services (WPRS) National Colloquium*, páginas 187–199. U.S. Department of Labor, Employment and Training Administration, 1996. 27

[50] Froelich, Markus, Michael Lechner e Heidi Steiger: *Statistically assisted programme selection - international experiences and potential benefits for switzerland*. Swiss Journal of Economics and Statistics, 139:311–331, janeiro 2003. `https://www.alexandria.unisg.ch/15873/`. 27

[51] Rosholm, Michael, Michael Svarer e Bo Hammer: *A danish profiling system*. SSRN Electronic Journal, 2004, ISSN 1556-5068. 27, 28

[52] Troya, Íñigo Martínez de Rituerto de, Ruqian Chen, Laura O Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B Brás e Leid Zejnilovic: *Predicting, explaining, and understanding risk of long-term unemployment*. Em *32nd Conference on Neural Information Processing Systems*, 2018. 28, 29

[53] Governo Federal Brasileiro: *Painel de informações do seguro-desemprego*, 2019. `https://app.powerbi.com/view?r=eyJrIjoiOGZhZDFmNTgtOTU5YS00YzYzLWFmOTgtMWVjZGYzMTM1MmZlIiwidCI6IjNlYzkyOTY5LTVhN` `pageName=ReportSectionc8988ed07c71babe2419`, acesso em 02/07/2023. 53

[54] Presidência da República: *Medida provisória n° 1164, de 2023*, 2023. `https://www.congressonacional.leg.br/materias/medidas-provisorias/-/mpv/156003`. 54

[55] Presidência da República: *Lei no 7.998/1990*, 1990. `https://www.planalto.gov.br/ccivil_03/leis/L7998compilado.htm`. 54