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

Modelos de aprendizagem de máquina para identificar o risco do trabalho escravo contemporâneo em cidades brasileiras

Marlu da Silva Santos

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

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2020

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

da Silva Santos, Marlu
dD111m Modelos de aprendizagem de máquina para identificar o risco do trabalho escravo contemporâneo em cidades brasileiras / Marlu da Silva Santos; orientador Gladston Luiz da Silva. -- Brasília, 2020.
65 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2020.

1. Machine Learning. 2. Logistic Regression. 3. Gradient Boosting. 4. Data Mining. 5. Escravidão Contemporânea. I. Luiz da Silva, Gladston, orient. II. Título.

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

Modelos de aprendizagem de máquina para identificar o risco do trabalho escravo contemporâneo em cidades brasileiras

Marlu da Silva Santos

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

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

Prof. Dr. Adriano Lorena Inácio de Oliveira
Universidade Federal de Pernambuco

Prof. Dr. Marcelo Ladeira
Universidade de Brasília

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 2 de dezembro de 2020

Dedicatória

Aos meus avós (*in memorian*): Ildorce, Agenor e Maria Abadia. Aos meus pais Lucélia e Manoel, minha irmã Lorena, ao meu sobrinho Kauã e em especial a minha companheira Daiani por ser a minha fonte de inspiração na vida. Pelo amor e carinho que tenho por todos vocês dedico este trabalho.

A todas as pessoas que acreditam em um mundo com condições dignas e humanas de trabalho.

Agradecimentos

Ao professor Dr. Gladston, pela orientação e competência inestimável, por sempre realizar os direcionamentos necessários da pesquisa, atuando de forma cordial e profissional.

Meus sinceros agradecimentos a todos os professores e amigos do mestrado, em especial da linha de pesquisa de Ciência de Dados. Agradeço ao Maurício pela amizade e pela partilha de conhecimento.

A vida nos proporciona amigos, independente da quantidade o importante é ter algum, e tenho a satisfação em poder agradecer por escrito ao meu grande amigo Lincoln, que mesmo em outro continente me apoiou nas revisões.

Agradeço também ao Ministério Público do Trabalho por ter apoiado esta pesquisa.

Resumo

O crime do trabalho escravo contemporâneo permeia por centenas de países e extinguir essa violação humana é um dever global. No Brasil, o trabalho escravo contemporâneo é caracterizado pelo código penal. Uma série de problemas são encontrados pelos agentes responsáveis pela inibição deste crime. As principais dificuldades estão relacionadas em: atender as denúncias que necessitam ser priorizadas; identificar ou antecipar aos crimes; medir o nível de erradicação do trabalho escravo; e recursos insuficientes para atender as ocorrências. A existência de um mecanismo para prever o nível de risco associado a cada cidade pode ser uma ferramenta importante para um passo na erradicação do trabalho escravo contemporâneo. Este estudo propõe o uso de modelos preditivos para identificar o risco da escravidão contemporânea em cidades brasileiras utilizando dados socioeconômicos, demográficos e registros de operações de resgate. Como existem muitas denúncias deste tipo de crime, identificar o grau do risco em cada cidade é uma ferramenta essencial para auxiliar no planejamento das fiscalizações. O estudo utiliza a técnica *embedded* com regularização Lasso (L1) para seleção de variáveis. Um método comparativo de técnicas para o tratamento de dados desbalanceados foi aplicado, os resultados mostraram que para o contexto do problema a técnica indicada é *Random Oversampling* (ROS). No total, 16 modelos são avaliados, formados por 8 diferentes conjuntos de dados e dois classificadores: *Logistic Regression* (LR) e *Gradient Boosting Machine* (GBM). Os resultados indicam o modelo GBM com melhor performance, com acurácia de 77%, AUC 80% e G-mean 71%. Como validação do modelo um teste estatístico com reamostragem é aplicado utilizando *Bootstrapping* para 1000 iterações, cujos resultados apontam que o modelo se manteve robusto, visto que para um intervalo de confiança de 0.95, a acurácia ficou entre 87.5% e 87.8%. O melhor modelo foi validado com dados de fiscalização mais recentes, cujos resultados do levantamento revelaram estar coerentes com o teste estatístico do modelo, visto que de 96 novas ocorrências registradas para os anos de 2019 a junho de 2020, o modelo acertou 87,5% e errou 12,5%.

Palavras-chave: Machine Learning, Logistic Regression, Gradient Boosting, Data Mining, Imbalanced Dataset, Escravidão Contemporânea.

Abstract

Crime of contemporary slave labor pervades hundreds of countries and extinguishing this human violation is a global duty. In Brazil, contemporary slave labor is characterized by the penal code. A number of problems are encountered by the agents responsible for inhibiting this crime. The main difficulties are related to: attend to complaints that need to be prioritized; identify or anticipate crimes; measure the level of eradication of slave labor; and insufficient resources to deal with the occurrences. The existence of a mechanism to predict the level of risk associated with each city can be an important step towards the eradication of contemporary slave labor. This study proposes the use of predictive models to identify the risk of contemporary slavery in Brazilian cities using socioeconomic, demographic and rescue operation records. As there are many reports of this type of crime, identifying the degree of risk in each city is an essential tool to assist in planning inspections. The study uses the *embedded* technique with Lasso regularization (L1) to select variables. A comparative method of techniques for the treatment of unbalanced data was applied, the results showed that for the context of the problem the appropriate technique is *Random Oversampling* (ROS). In total, 16 models are evaluated, formed by 8 different data sets and two classifiers: *Logistic Regression* (LR) and *Gradient Boosting Machine* (GBM). The results indicate the GBM model with the best performance, with accuracy of 77%, AUC 80% and G-mean 71%. As a validation of the model, a statistical test with resampling is applied using Bootstrapping for 1000 iterations, which results show that the model remained robust, seen that for a confidence interval of 0.95, the accuracy was between 87.5% and 87.8%. The best model was validated with more recent inspection data, the results of the validation revealed to be consistent with the statistical test of the model, since of 96 new occurrences registered for the years 2019 to June 2020, the model got 87,5% right and 12,5% wrong.

Keywords: Machine learning, Logistic Regression, Gradient Boosting, Data Mining, Imbalanced Dataset, Modern Slavery.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ADASYN *Adaptive Synthetic Sampling Approach.*

AUC *Area Under the Curve.*

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

GBM *Gradient Boosting Machine.*

IDH *Índice de Desenvolvimento Humano.*

ILO *International Labour Organization.*

KDD *Knowledge Discovery in Databases.*

LASSO *Least Absolute Shrinkage and Selection.*

LR *Logistic Regression.*

MPF Ministério Público Federal.

MPT Ministério Público do Trabalho.

MPU Ministério Público da União.

PNUD Programa das Nações Unidas para o Desenvolvimento.

RFE *Recursive Feature Elimination.*

ROC *Receiver Operating Characteristic.*

ROS *Random Oversampling.*

RUS *Random Undersampling.*

SMOTE *Synthetic Minority Over-Sampling Technique.*

Capítulo 1

Introdução

A definição e fundamentação da escravidão moderna, ao que se refere aos aspectos jurídicos e sociais é interesse de pesquisa na atualidade [2]. Estudos demonstram o entendimento de que fazem parte do escopo do trabalho escravo: o tráfico humano [3], trabalho forçado, servidão por dívida e exploração infantil [4] [5]. O trabalho forçado é conceituado segundo a Convenção nº 29 da International Labour Organization (ILO) [6], como sendo, "(...)todo trabalho ou serviço exigido de um indivíduo sob ameaça de qualquer penalidade e para o qual ele não se ofereceu de espontânea vontade".

O termo é tipificado pela legislação brasileira no Artigo 149 do Código Penal[1] [2], que amplia o sentido de trabalho forçado e dispõe sobre a condição análoga à de escravo, referindo-se a alguém que seja submetido a trabalhos forçados ou jornadas exaustivas, condições degradantes de trabalho, restrição de locomoção do empregado em razão de dívidas contraídas com o empregador.

Alguns elementos contribuem para o risco do crime, que são, primeiramente a existência de um alvo vulnerável, podendo ser pessoas em situações de desassistência social. O segundo fator é na existência da oferta e na necessidade de alguém que se dispõe a fazer. Um terceiro ponto, trata-se da ausência de fiscalização, ou medidas preventivas, sejam públicas ou privadas, o que contribuem para o aumento de atos criminosos. O quarto elemento refere-se ao ambiente favorável à prática do crime e desfavorável a quem fiscaliza e previne [7].

A existência do risco para as formas de crime do trabalho escravo contemporâneo, vai de encontro com a abordagem dos elementos fundamentais que contribuem para qualquer forma de crime. Neste sentido, a existência dos fatores que contribuem para a ocorrência de práticas criminosas estão contidas nas práticas do trabalho escravo da era moderna.

[1] www.planalto.gov.br/ccivil_03/Decreto-Lei/Del2848.htm#art149
[2] www.planalto.gov.br/ccivil_03/decreto-lei/del2848.htm

A estimativa do número de pessoas em condições de trabalho escravo, tem sido pesquisada principalmente por instituições não governamentais[3,4,5,6]. Os relatórios técnicos produzidos por tais instituições, objetivam construir um melhor entendimento sobre o problema, assim como fornecer mecanismos que auxiliem em tomadas de decisões e políticas públicas [8] [4] [5].

No sentido de erradicar o trabalho escravo, diversas medidas são motivadas por instituições governamentais e instituições que aderem à causa da extinção do trabalho escravo. A recomendação 203 do protocolo de 2014 da ILO [9], que trata sobre a Conveção nº29 [6], prioriza que as nações membro deverão implementar políticas nacionais, planos de ações efetivas com supressão ao trabalho forçado ou compulsório em todas as suas formas por meio da prevenção, proteção e acesso a recursos, tais como indenização das vítimas e punição aos causadores.

O perfil do trabalhador brasileiro submetido ao trabalho escravo, segundo a publicação sobre a atuação do Ministério Público do Trabalho (MPT)[7], trata-se de pessoas em condições de extrema vulnerabilidade socioeconômica, oriundas em grande parte do meio rural, com baixa escolaridade, que migram em busca de alternativas de vida e de trabalho [10].

Ao se analisar o perfil dos principais atores envolvidos no trabalho escravo rural no Brasil, demonstra-se que a maioria dos resgatados tinham em seu histórico trabalho infantil, pobreza, analfabetismo ou baixa escolaridade. O estudo [11] faz o mapeamento de perfil típico do trabalhador submetido ao trabalho escravo contemporâneo. Trata-se de migrantes, principalmente da região nordeste do país, de sexo masculino, baixa escolaridade, que são aliciados para trabalhar em áreas principalmente vinculadas ao desmatamento [11].

O elevado índice de trabalho escravo contemporâneo é preocupante. De acordo com os dados estatísticos divulgados pela *International Labour Organization* (ILO), estima-se que em 2016 pelo menos 1,3 milhões de pessoas encontravam-se em algum tipo de situação de trabalho escravo moderno ou trabalho degradante nas Américas [4].

No Brasil existem algumas medidas realizadas pelo Governo Federal para a erradicação do trabalho escravo, iniciados nos anos de 2003 e 2008: grupos nacionais móveis de fiscalização; e, "lista suja de empresas"[8], referente ao cadastro de empregadores que tenham submetido trabalhadores a condições análogas à de escravo. Embora existam os

[3] `https://www.alliance87.org`

[4] `https://www.minderoo.com.au`

[5] `https://ilo.org`

[6] `https://www.antislavery.org`

[7] Instituição fiscalizadora da legislação trabalhista brasileira - `mpt.mp.br`

[8] Portaria interministerial nº 4, de 11 de maio de 2016, `http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&data=13/05/2016&pagina=178`

programas de combate à escravidão moderna, os índices ainda revelam elevadas taxas de ocorrências de resgates de pessoas nestas condições desumanas.

A forma de atuação do MPT, ocorre mediante articulação interinstitucional e pela via Judicial. Na articulação interinstitucional, são realizadas parcerias e desenvolvem-se projetos comuns. Pela via judicial, questionam-se junto ao poder Judiciário a ausência ou insuficiência de políticas públicas relacionadas ao tema, a fim de se obter uma determinação judicial dirigido-se ao poder Legislativo ou Executivo para que contemple medidas aptas a erradicar o trabalho escravo [10].

Uma das atribuições do Ministério Público é o trabalho conjunto com outros órgãos e entidades comprometidas com o tema, dentre os quais cito: Ministérios Públicos Federais e Estaduais, Secretaria do Trabalho, Ministério da Justiça, polícias Federal, Rodoviária, Civil e Militar, Universidades, poder Judiciário, Fórum de Direitos Humanos, entre outras entidades envolvidas com a questão do trabalho escravo [10].

O efeito de parcerias pode introduzir pessoas resgatadas do trabalho escravo no mercado de trabalho, oriundas de fiscalizações do MPT que culminam em reparação por dano moral coletivo causado tanto em termos de ajuste de conduta, quanto em ações civis públicas. A reintrodução do trabalhador no mercado de trabalho faz com que a pessoa evite ser aliciada ao trabalho escravo novamente [10] [12].

O MPT, tem como atribuição principal fiscalizar o cumprimento da legislação trabalhista, promover ação civil pública na esfera da Justiça do Trabalho, propor as ações necessárias à defesa dos direitos e interesses dos menores, incapazes e índios, decorrentes de relações de trabalho. Na atualidade, possui oito coordenadorias nacionais temáticas: Trabalho Escravo, Meio Ambiente do Trabalho, Fraudes Trabalhistas, Promoção da Igualdade, Criança e Adolescente, Trabalho Portuário e Aquaviário, Liberdade Sindical e Administração Pública, que promovem discussões sobre suas respectivas áreas, definem estratégias e articulam planos nacionais de ações.

Assim como os demais ramos do Ministério Público, o MPT exerce importante papel na resolução administrativa (extrajudicial) de conflitos. A partir do recebimento de denúncias, representações, ou por iniciativa própria, pode instaurar inquéritos civis e outros procedimentos administrativos, notificar as partes envolvidas para que compareçam a audiências, forneçam documentos e outras informações necessárias [10].

Este estudo propõe uma abordagem de uso de modelos de previsão para identificar o risco da escravidão contemporânea nas cidades brasileiras, com o intuito de auxiliar nas fiscalizações de combate ao trabalho escravo. Provendo como resultado o grau do risco associado para cada cidade, podendo ser utilizado como ferramenta para priorizar ações em determinadas regiões ou na escolha das denúncias a serem investigadas. Este trabalho faz uma abordagem ao uso da aprendizagem de máquina aplicada a um problema real. O

modelo de referência para mineração de dados *CRoss Industry Standard Process for Data Mining* (CRISP-DM)[1] é utilizado para conduzir tarefas que envolvam a mineração de dados. A partir de registros que evidenciam a existência de trabalho escravo, as cidades foram classificadas em duas categorias: cidade com trabalho escravo (1) ou cidade não rotuladas (0). Devido ao menor número de ocorrências de classe 1, por conseguinte, classe minoritária, foi utilizado o tratamento de dados desbalanceados e, devido à alta dimensionalidade foi necessário o uso de técnicas de seleção de variáveis. Ao final, dezesseis modelos preditivos com algoritmos classificadores de *Logistic Regression* (LR) e *Gradient Boosting Machine* (GBM) foram modelados e avaliados.

1.1 Justificativa

Segundo dados da Secretaria do Trabalho, entre 1995 e 2019 o Brasil teve 53.635 casos [9] de resgate de pessoas em situação de trabalho análogo à escravidão, sendo que somente no ano de 2018 foram resgatadas 1.744 pessoas. O Brasil é um país membro da ILO e possui acordos ratificados, dentre os quais a Convenção nº 29 [6], que refere-se ao compromisso em abolir a utilização do trabalho forçado, dentre outras exigências, que exige dos países membros a Recomendação nº 203 do Protocolo 2014 [9]. Tais acordos reforçam o reconhecimento da existência do problema pela nação e o interesse em sua resolução.

Ao considerar o índice elevado de pessoas submetidas a situações de trabalho escravo contemporâneo [8] [4] [5], entidades passam a trabalhar de forma conjunta para sua erradicação. Entretanto, apesar das ações existirem, há muito a ser realizado, tanto de forma preventiva, quanto de forma repressiva. A fiscalização é um dos eixos que atuam de forma direta no combate a este tipo de crime, promovida por meio de auditorias, onde as operações são motivadas principalmente por meio de denúncias, onde uma pequena parte é atendida devido a grande demanda, a ausência de mecanismos para auxiliar na escolha das regiões a serem investigadas é uma realidade.

Por meio da análise dos dados da Secretaria do Trabalho, é de conhecimento que determinadas regiões do Brasil possuem maior incidência de trabalho em condições análogas a de escravidão. No entanto, poucos estudos abordam os fatores que encadeiam as ocorrências do trabalho forçado. Um dos motivos que distancia o entendimento deste problema, além da complexidade ao tema, está na relação de uma rede sistêmica de variáveis que o cerca.

O trabalho decente e crescimento econômico faz parte do escopo da agenda 2030 [13], que refere-se aos objetivos de desenvolvimento sustentável. Para cumprir esta meta os

[9] Radar: `https://sit.trabalho.gov.br/radar/`

países deverão adotar medidas e políticas que avancem no sentido de contribuir para a proteção dos direitos trabalhistas, promover ambientes seguros para o trabalho independente do gênero da pessoa, e ainda proteger trabalhadores das formas de emprego precário. É adequado dizer que os esforços para a eliminação das formas de trabalho escravo contemporâneo são aderentes à agenda 2030, mais particularmente ao objetivo indicado no item 8.8.

Neste sentido, ações e estudos que busquem o entendimento das variáveis que contribuem para o risco de ocorrências do trabalho escravo ajudarão a auxiliar na construção de políticas voltadas para erradicar esta forma degradante de trabalho, buscando diminuir o índice de ocorrências e, consequentemente, melhora nas condições de trabalho.

O estudo busca contribuir com o desenvolvimento técnico aplicado ao tema, propondo uma abordagem e modelos que podem ser adaptados por outras regiões ou países. A pesquisa trata de um problema do mundo real e que ocorre em diversas regiões do globo.

A proposta abordada neste trabalho é de interesse para as instituições que possuem em seu escopo a busca pela solução do problema do trabalho escravo contemporâneo. Utilizando os dados disponíveis na instituição e transformando-os em conhecimento aplicado. Os resultados desta pesquisa poderão ser utilizados por grupos de trabalhos atuantes. Auxiliando nas tomadas de decisões, antecipação a eventos e principalmente em ações específicas da Coordenadoria Nacional de Erradicação do Trabalho Escravo. O enfoque da pesquisa, na previsão do risco do trabalho escravo nas cidades Brasileiras, está diretamente relacionado ao plano de atuação da instituição do MPT e das instituições parceiras.

A hipótese da pesquisa é de que é possível identificar as variáveis que relacionam-se com a ocorrência de trabalho escravo, a partir de dados de IDH dos municípios e de bases de dados com registros das operações de resgate, assim como identificar o risco do trabalho escravo contemporâneo a partir da aplicação de modelos de aprendizagem de máquina.

1.2 Objetivos

Este trabalho tem como objetivo geral a construção de modelos de previsão para identificar o risco do trabalho escravo contemporâneo em cidades brasileiras. Para essa finalidade, são aplicadas técnicas de mineração de dados em conjunto com o levantamento de pesquisas relacionadas ao tema, bem como, o uso de base de dados originárias de operações de resgate do trabalho escravo e indicadores sociais.

Como objetivos específicos, cita-se: i) realizar a seleção sistêmica de variáveis que correlacionam com o trabalho escravo; ii) realizar o estudo comparativo (*benchmarking*) das técnicas de balanceamento de dados; iii) construir e avaliar modelos de aprendizagem de máquina; iv) construir um teste estatístico com reamostragem para avaliar a robustez

do modelo; v) determinar o grau do risco e ranking do trabalho escravo para cada cidade; vi) avaliar os resultados obtidos em conjunto com as equipes de atuação da erradicação do trabalho escravo nas instituições do MPT e Secretaria do Trabalho;

1.3 Metodologia

Nesta seção é abordada a metodologia seguida por este estudo. Os conjunto de dados utilizados são: o Atlas dos Municípios do Brasil[10] e Operações de Resgate do Trabalho Escravo[11] de 2003 à 2018. As ferramentas utilizadas são: linguagem de programação *Python* com pacotes *scikit-learn*, *imbalanced-learn*, *matplotlib*, *plotnine* e *multiprocessing*.

- Em primeiro momento, é feito o levantamento de informações sobre os grupos de trabalhos atuantes na erradicação do trabalho escravo, caracterizando o entendimento do negócio e do problema a ser abordado, conforme apresentado na Seção 3.1;
- Após o entendimento do negócio, é realizado o entendimento dos dados, nesta etapa as diferentes bases de dados são coletadas e analisadas, um pré-processamento inicial é necessário para realizar a limpeza dos dados.
- Com a limpeza inicial realizada a próxima etapa está em realizar a preparação dos dados, o que resultará em uma versão unificada do conjunto de dados classificados com a variável alvo dependente e as variáveis independentes.
- A etapa de modelagem destina-se na construção dos modelos de previsão ao risco do trabalho escravo. De acordo com o estudo levantado, é feita a escolha ao uso da regressão logística [14] [15] e GBM para a classificação. O conjunto de dados é dividido em treinamento (70%) e teste (30%) conforme abordagem *holdout*. O treinamento é realizado sobre os dados re-amostrados para compor o balanceamento das classes. Uma abordagem comparativa entre métodos de balanceamento é realizada para a escolha da melhor técnica.
- As validações em treino são realizadas com *Cross-validation* com 10 *folds*. Os modelos são avaliados com as métricas da acurácia, AUC, *F1-measure*, *precision*, *recall* e *G-mean*. A performance será constatada pelo o uso da curva ROC. Um teste estatístico é realizado utilizando a técnica de *bootstrapping* para 1000 iterações.
- Na etapa final, é aplicada a implantação do modelo selecionado, onde é apresentado o *ranking* dos resultados para as partes interessadas. Uma etapa adicional é aplicada

[10]Disponível em `http://www.atlasbrasil.org.br/2013/pt/download/`, acessado em 27/09/2018.
[11]Fornecido por `http://observatorioescravo.mpt.mp.br`, acessado em 25/09/2018.

ao modelo, realizando a discretização para uma escala da escravidão contemporânea [16]. Um levantamento comparativo entre o modelo proposto e os dados dos anos de 2019 a junho de 2020, referente as fiscalizações da Secretaria do Trabalho é realizado com a finalidade de atestar na prática o modelo desenvolvido.

1.4 Estrutura

Este trabalho segue estruturado em quatro capítulos. Os capítulos subsequentes ao Capítulo 1 são organizados em: Capítulo 2, que se destina a revisão bibliográfica utilizada neste trabalho. O Capítulo 3 refere-se à solução proposta, onde será apresentado de forma aprofundada as etapas apresentadas na metodologia, o entendimento do negócio, o entendimento dos dados, preparação dos dados, modelagem e avaliação. No Capítulo 4 os resultados são apresentados. Por fim, as conclusões do estudo são expostas, seguida do referencial bibliográfico.

Capítulo 2

Revisão bibliográfica

Este capítulo aborda todos os artigos, pesquisas e trabalhos acadêmicos de maior relevância utilizados para o desenvolvimento deste estudo. Em primeiro momento é feita a contextualização de mineração de dados e aprendizagem de máquina. Em seguida são apresentados as etapas do modelo de referência de mineração de dados CRISP-DM [1], uma breve diferenciação é feita entre os processos do *Knowledge Discovery in Databases* (KDD) [17].

Posteriormente, são apresentadas as técnicas para seleção de variáveis, representando uma etapa importante na análise e preparação dos dados. Após as pesquisas sobre seleção de variáveis, a revisão da literatura é realizada sobre os algoritmos supervisionados que serão utilizados nos modelos. Como a abordagem desta pesquisa utiliza classes desbalanceadas, se faz necessário o estudo sobre técnicas para o balanceamento de dados, um tópico sobre re-amostragem é utilizado para essa discussão.

Ao final deste capítulo são apresentadas as técnicas de validação cruzada, avaliação de modelos e os trabalhos relacionados, respectivamente. A avaliação de modelos, possui a finalidade de aferir a performance e a acurácia para que as otimizações sejam realizadas. Os trabalhos relacionados, utilizam modelos estatísticos para resolução de problemas de estimativas de previsão com o enfoque ao tema da escravidão contemporânea.

2.1 Mineração de dados

Um grande volume de dados proveniente de aplicações computacionais tem emergido de diferentes áreas da industria e sociedade. A mineração de dados surge com a necessidade de extrair informações relevantes desta grande concentração de dados. As informações coletadas podem ser utilizadas para a descoberta de conhecimento, que trata-se do processo de obter informações não triviais, consideradas previamente desconhecidas e que podem ser de grande valor [18] [17] [19].

O uso de mineração de dados e a descoberta de conhecimento podem ser aplicadas em diferentes abordagens, como em tomada de decisão, visualização de dados, aprendizagem de máquina [20], entre outras aplicações [19]. Para este estudo, a mineração de dados tem papel bem definido, por concentrar as principais tarefas computacionais que são realizadas nas bases de dados, onde o objetivo está na construção de conhecimento para a resolução do problema.

2.2 Aprendizagem de Máquina

Aprendizado de máquina é um método computacional que utiliza técnicas estatísticas e matemáticas para realizar tarefas de previsões, fazendo com que melhore os resultados esperados com a experiências passadas. A aprendizagem de máquina faz intercessão entre duas ciências principais, Computação e Estatística [21] [22].

A teoria por trás da aprendizagem de máquina está centrada em buscar respostas para questões ligadas ao desenvolvimento de software de computadores que aprendem com suas próprias experiências, além da teoria estatística e computacional que envolvem os sistemas de aprendizagem. Em um escopo maior está a inteligência artificial (IA)[23] [24], que utiliza a aprendizagem de máquina para resolução de problemas, como de visão computacional, reconhecimento de fala, processamento de linguagem natural e várias outras aplicações [21] [25]. Para um melhor entendimento a aprendizagem de máquina está contida dentro de um conceito maior chamado inteligencia artificial, fazendo uma analogia a conjuntos, seria representado por $B \subset A$, onde B trata-se da teoria de aprendizagem de máquina. Uma ideia mais completa, seria representada por $C \subset B \subset A$, onde C representaria uma abordagem em *Deep Learning* [26].

2.3 CRISP-DM e KDD

Os processos de referência são utilizados como uma camada para auxiliar nas etapas que utilizam técnicas de Mineração de dados. As principais referências de processos utilizados são: CRISP-DM [1] [27] e KDD[17]. Nesta seção é dada uma breve descrição aos processos do KDD e posteriormente com maior enfoque os processos do CRISP-DM, por tratar-se do modelo de referência escolhido neste trabalho. As principais diferenças entre os dois modelos decorrem das características explícitas dos processos de Entendimento do Negócio e Implantação, contidas no CRISP-DM [27].

Os processos do KDD são definidos em 5 estágios: Seleção - onde é criado a base de dados alvo, com o enfoque na seleção das variáveis ou amostras de dados; Pré-processamento - consiste na limpeza e no pré-processamento, com o objetivo de se obter dados consis-

tentes; Transformação - são realizadas transformações nos dados, utilizando métodos ou redução de dimensões; Mineração de Dados - essa etapa destina-se na criação do modelo ou busca por padrões, vai depender do objetivo principal de interesse; e Interpretação e Avaliação - nesta etapa os resultados são interpretados e avaliados [17].

O CRISP-DM aborda seis estágios de processos, que são: Entendimento do negócio, Entendimento dos dados, Preparação dos dados, Modelagem, Avaliação e ao final, Implantação [1]. A Figura 2.1 apresenta de forma ilustrada cada uma das etapas.

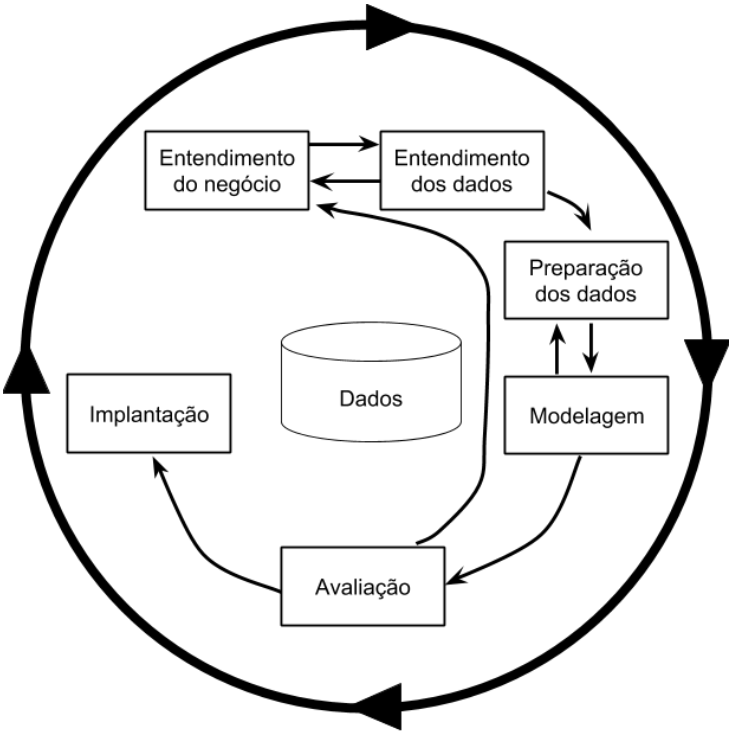


Figura 2.1: Modelo de referência de processos - CRISP-DM (adaptado de [1])

Entendimento do negócio: trata-se da fase inicial, onde os requisitos do ponto de vista do domínio do negócio são entendidos e planejados, convergindo conhecimento sobre o problema de mineração de dados a ser resolvido.

Entendimento dos dados: é feito o entendimento inicial dos dados, em um conjunto de dados procura-se absorver conhecimento sobre os dados, identificando subgrupos de interesse, promover os primeiros *insights* e levantar hipóteses sobre informações ocultas.

Preparação dos dados nesta etapa o objetivo é chegar em um conjunto de dados final, para isso tarefas de transformação e limpeza dos dados podem ser realizadas de forma contínuas.

Modelagem: neste estágio são aplicadas diferentes técnicas para construção de um modelo que satisfaça a resolução do problema. É comum que seja necessário voltar ao estágio de preparação dos dados, pois em muitos dos casos existirá a necessidade do ajuste de variáveis ou parâmetros.

Avaliação: é uma etapa importante para verificar se o modelo criado atende aos objetivos do negócio. É decidido passar para etapa de implantação, após a avaliação do modelo.

Implantação: é nesta etapa em que o modelo é utilizado para agregar algum valor ao domínio do negócio, podendo ser utilizado para construir ou auxiliar em tomadas de decisões, o formato de implantação pode variar entre relatório ou um sistema.

2.4 Seleção de variáveis

O aumento significativo dos dados e consequentemente dos problemas a serem resolvidos, torna a seleção de variáveis uma importante tarefa a ser realizada quando existe a necessidade de redução de dimensionalidade dos dados e melhoria na acurácia dos modelos de aprendizagem. Neste sentido, o uso de diferentes métodos para seleção de variáveis estão sendo pesquisados para aplicações em problemas de aprendizagem de máquina [28] [29] [30] [31] [32].

A seleção de variáveis ou a eliminação de características não representativas do volume dos dados tem demonstrado ser necessárias para a preparação dos dados de forma eficiente. A seleção de variáveis pode ser classificada nas perspectivas de métodos supervisionados, não supervisionados e semi-supervisionados [33].

Os problemas de classificação ou regressão estão geralmente relacionadas a técnicas de seleção de variáveis supervisionadas. Com o rótulo preditor, a importância da característica da variável é avaliada e selecionada de acordo com sua classe ou com o propósito da regressão usada. Já os problemas de *clustering* podem ser aplicados a técnicas de seleção não supervisionada por não necessitar de rótulos preditores [33]. Este trabalho utiliza como formato a seleção de variáveis por meio de um rótulo supervisor.

Diferentes estratégias podem ser aplicadas para diferentes tipos de dados, as principais são *wrapper*, *filter*, *embedded*. Na estratégia *wrapper* um algoritmo de aprendizagem busca por um subconjunto de variáveis e avalia a qualidade das características selecionadas com base no aprendizado, mas essa estratégia oferece baixa performance para ser aplicada com alta dimensão de dados [33]. Outra estratégia é o *filter*, que consiste em duas etapas: primeiro no rankeamento da importância da característica com base em critério de avaliação, podendo ser univariado (rankeamento individual de variáveis) ou multivariado

(rankeamento em lote); na segunda etapa é realizado o filtro das características de baixa relevância [33] [29].

Já a estratégia *embedded* agrega modelos de aprendizagem a seleção de variáveis, unindo as estratégias *wrapper* e *filter*, essa estratégia demonstra ser mais eficiente. As técnicas mais utilizadas são os modelos de regularização [34], onde o objetivo é minimizar os erros de ajuste [33] [29]. Em um estudo recente, foi proposto o uso da eliminação de variáveis recursivas com teste de sensibilidade, utilizando *Support Vector Machine* (SVM), cujos resultados demonstraram ser satisfatórios em um problema de grande dimensionalidade de variáveis [31].

2.4.1 Regularização

A regularização pode ser empregada quando existe um problema matemático que pode ser solucionado computacionalmente e que não está bem definido. Dois métodos de *Shinrikage* principais são utilizados para reduzir o número de atributos com pouca significância. De uma forma geral as técnicas de regularização utilizam a penalização da função de custo. Para este estudo a regularização será utilizada para seleção de variáveis e redução de *overfitting*. A formulação geral pode ser dada pela Equação 2.1, apresentada em [35].

$$\min_{f \epsilon H} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \tag{2.1}$$

Onde, $J(f)$ é a penalização dada para a função de custo $L(y_i, f(x_i))$, H é o espaço das funções no qual $J(f)$ está definido e λ é o parâmetro de suavização.

LASSO

O *Least Absolute Shrinkage and Selection Operator* (LASSO) é um método cuja a proposta está em minimizar a soma residual dos quadrados, podendo ser usada para estimação de modelos. Produz boa capacidade para interpretação, com aplicação de forma geral em modelos, incluindo para a seleção de variáveis e regularização [34]. O uso do LASSO pode ser estendido para modelos lineares generalizados e equações de estimativas generalizadas. Conforme definição dada [35], a Equação 2.2 define de forma objetiva a regressão, onde dado um conjunto de dados (x_i, y_i), com valores em $i = 1, 2, ..., N$, sendo $x_i = (x_{i1}, ..., x_{ip})^T$ são as variáveis independentes e y_i a variável dependente, onde β_0 é uma constante. Assumindo existir padronização nas variáveis de entrada x_i. A equação do LASSO é dado por:

$$\widehat{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{para} \quad \sum_j |\beta_j| \leq t \tag{2.2}$$

Regressão *Ridge*

Diferente do LASSO, que penaliza de acordo com a soma do valor absoluto dos coeficientes, reduzindo o erro quadrático, a Regressão *Ridge* penaliza o tamanho dos coeficientes de regressão, por meio de um estimador *Ridge*, diminuindo a importância do atributo [35] [36]. A estimativa de *Ridge* $\hat{\beta}$ é definida como um valor de β que é minimizado, como formalizado pela Equação 2.3.

$$\widehat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.3}$$

Onde, $\lambda \geq 0$ vai aumentar a capacidade de encolhimento (*shrinkage*), os coeficientes vão convergir a 0. Na Equação 2.3, a penalização do coeficiente é feita por $\lambda \sum_{j=1}^{p} \beta_j^2$.

2.5 Algoritmos de classificação

Nesta pesquisa a Regressão Logística (LR) foi utilizada como classificador, pois foi amplamente utilizada nos trabalhos relacionados. Um classificador adicional, *Gradient Boosting Machine* (GBM), é usado na comparação. De fato, consideramos vários outros classificadores tradicionais. No entanto, esses dois classificadores foram escolhidos para este estudo, pois eles tiveram um desempenho melhor que os outros e simplificaram a análise de *benchmarking* entre as técnicas de balanceamento.

A regressão logística pertence ao grupo de *Generalized Linear Models* (GLM), e está fortemente relacionada a problemas que existam a necessidade de descrever relações entre variáveis independentes (preditoras) e variáveis dependentes (resposta) com saída binária ou dicotômico, ao contrário de regressão linear que comumente a variável alvo é contínua [37]. O modelo logístico é representado na Equação 2.4.

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{2.4}$$

Onde, os termos α e β são representados por parâmetros desconhecidos que são necessários ser estimados com base nos valores obtidos de X_i e P(X) é a probabilidade a ser obtida para o conjunto de variáveis independentes.

O algoritmo *Gradient Boosting Machine* (GBM) foi introduzido por Friedman [38], possuindo como base principal a técnica de *Boosting*. O algoritmo GBM tenta otimizar a função de custo, utiliza os algoritmos *Boosting* para transformar aprendizagens fracas em novas predições e com gradiente descendente auxiliar nas novas árvores que são adicionadas ao modelo.

2.6 Reamostragem

A reamostragem faz uso de dados derivados observados ou gerados, utilizados para produzir novas amostras hipotéticas que simulam a população subjacente. Os dados reamostrados são amplamente utilizados, principalmente em situações em que as abordagens paramétricas são de difícil aplicação. De uma forma geral, os dados reamostrados são centenas e ou milhares de novos ensaios. Os algorítimos realizam massivas repetições, o que não é um problema para a realidade computacional cotidiana. Mas que em décadas passadas, esse método era de difícil aplicação [39].

A técnica mais utilizada é o *Bootstrapping*, mas também existem outros métodos como *Jackknife*, que usa inferência estatística para estimar a tendência e o erro padrão [40]. Também existem outros tipos de reamostragem que são: simulação de Monte Carlo e Teste de Randomização (permutação) [41].

Com base na revisão bibliográfica [39] [40] [41] [42], a escolha do método de reamostragem escolhido para aplicação neste estudo, foi o método *Bootstrapping*, apresentado com mais detalhe na Seção 2.6.1.

Existem algumas recomendações ao utilizar técnicas de reamostragem. Algumas são descritas por Philip Crowley [39]. Tem-se como exemplo, utilizar uma maior quantidade de ensaios, ou seja, repetições de proporção massiva. O mesmo autor indica o uso do *Bootstrapping* como superior para abordagens em que exigem a necessidade do intervalo de confiança e teste de hipótese.

2.6.1 *Bootsrapping*

O *Bootstrapping* usa amostragem aleatória com reposição e possibilita estimar as propriedades de um estimador. A técnica envolve em realizar repetidas reamostragens, gerando uma estimativa empírica de uma amostra de distribuição estatística. O uso de *Bootstrapping* também é aplicado em abordagens de inferência estatística [42].

A Figura 2.2, aborda de forma abstrata os passos da técnica *Bootstrapping* para o teste de robustez com intervalo de confiança. Em um primeiro momento é obtido o conjunto da amostra original. Em seguida é realizada a reamostragem aleatória com reposição para n repetições. São calculados as estatísticas θ (parâmetro a ser estimado) para cada amostra gerada, obtendo os estimadores $\widehat{\theta}$. Com base no cálculo estatístico é construído a distribuição da amostragem e então usada para fazer a inferência pretendida, utilizando o intervalo de confiança de $\widehat{\theta}$.

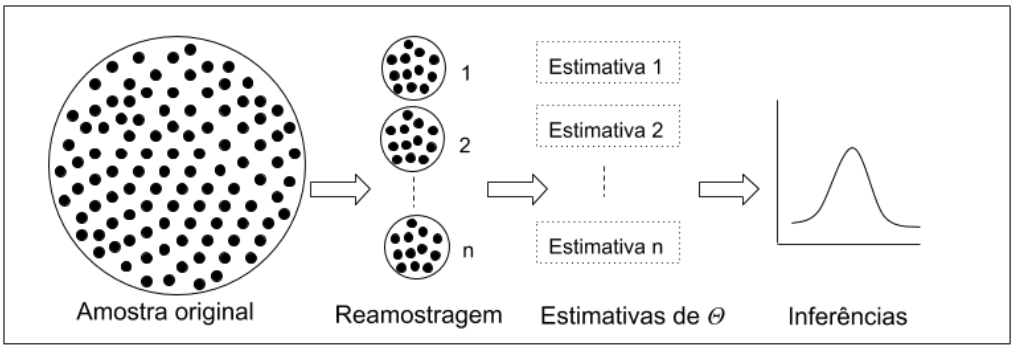


Figura 2.2: Passos do uso do *Bootstrapping*.

2.7 Técnicas de Balanceamento

Em abordagens de aprendizagem de máquina que utilizam a classificação binária, o desbalanceamento de classes gera problemas de desempenho para o modelo. As classes majoritárias tendem a influenciar os resultados se comparada às classes minoritárias [43], [44], [45]. O tratamento para classes desbalanceadas devem ser realizadas em dados de treinamento, no próprio nível dos dados [46], o qual, é a abordagem deste estudo.

Diferentes técnicas podem ser utilizadas para o balanceamento dos dados, como as abordagens em *Random Undersampling* (RUS) e *Random Oversampling* (ROS). Na técnica de RUS a classe majoritária é reduzida para o tamanho da classe minoritária, já a ROS faz cópias dos dados da classe minoritária de forma randômica. Outras técnicas de amostragem inteligente são derivadas destas principais, a saber, *Synthetic Minority Over-Sampling Technique* (SMOTE) [47], *Adaptive Synthetic Sampling Approach* (ADASYN) [48], *Cluster* [49], SMOTE-Tomek que combina a técnica de SMOTE com técnicas de limpeza de dados (Tomek) [50] para reduzir a sobreposição gerada pelos métodos de *oversampling* e SMOTE-ENN o qual faz a limpeza aprofundada. Bons resultados foram obtidos no estudo de Zhu et al. [46] utilizando técnicas de re-amostragem ROS e SMOTE-Tomek com classificador *Random Forest* [51].

2.7.1 *Random Undersampling*

A técnica de *Undersampling* faz a remoção aleatória de instâncias da classe majoritária, com ou sem reposição, isto é, com reposição implica na possibilidade da instância retirada ser coletada novamente para formar a nova amostra balanceada. Já sem reposição, a instância da classe majoritária é removida, e não repetida para formar a nova amostra balanceada. O maior problema nesta abordagem é o descarte de informações potencialmente úteis para sua classificação.

2.7.2 *Random Oversampling*

Ao contrário da abordagem que usa a classe majoritária, o *oversampling* utiliza a classe minoritária para balancear o conjunto de dados, duplicando aleatoriamente com ou sem reposição as observações de dados da classe minoritária. Algumas técnicas mais utilizadas derivadas de *oversampling* são: SMOTE e ADASYN.

A técnica de SMOTE cria dados sintéticos da classe minoritária, utilizando os k vizinhos mais próximos do espaço das observações. Para uma amostra atual x_i, uma nova amostra é criada x' selecionando um vetor $v(x_i - x_k)$ entre um dos k vizinhos próximos e o ponto de dados atual x_i. O cálculo consiste no produto do vetor selecionado, por um número aleatório p de valor entre 0 e 1. O novo ponto de dados x' será a adição do resultado com o ponto de dados atual.

$$x' = x_i + p \times v(x_i - x_k) \tag{2.5}$$

O mesmo cálculo da Equação 2.5 é usado para a técnica de ADASYN, na prática a diferença entre SMOTE, é que ADASYN gera amostras proporcionais de classe oposta de uma determinada vizinhança selecionada. Na técnica ADASYN é considerado a distribuição da densidade, que determinará a quantidade de amostras sintéticas a serem geradas para um ponto selecionado.

2.8 Validação cruzada

É uma técnica utilizada para avaliar o poder de generalização de um modelo, que normalmente está associado a um problema de previsão. O conjunto de dados é dividido de forma mutua e exclusiva em novos subconjuntos de dados. Um subconjunto de dados pode ser utilizado para treinar um modelo, estimando os coeficientes do modelo, outros subconjuntos podem ser utilizados de forma independente para validação e testes do modelo [52] [35].

Existem três formas de separação de dados mais utilizadas que são: *holdout*, *k-fold* e o *leave-one-out*. O *holdout* pode fazer a divisão dos dados de forma fracionada, em uma proporção de 2/3 dos dados para treinamento do modelo e uma proporção de 1/3 dos dados para o teste do modelo, mas também é possível, mas não muito comum, uma divisão proporcionalmente equivalentes para treino e testes. A técnica de *k-fold* faz a divisão dos dados em k subconjuntos de tamanho iguais, onde um conjunto de k é reservado para teste e os demais $k-1$ são destinados para treinamento e avaliação do modelo. O processo de treinamento e teste são repetidas k vezes, ao final é feita a avaliação do modelo por meio da média do desempenho de cada modelo. Outra técnica de divisão de dados é a

leave-one-out, que trata-se de uma forma específica do *k-fold*, onde $k = N$, ou seja, k é definido com o tamanho N, onde N representa o mesmo tamanho do conjunto de dados para o modelo [52] [35].

2.9 Avaliação de modelos

A avaliação de modelos é uma importante etapa, principalmente para construção de modelos de aprendizagem, pois por meio do uso de métricas são aferidas as taxa de acertos dos modelos. Uma abordagem avaliativa de modelos contribui para a comparação de diferentes algoritmos e técnicas.

2.9.1 Matriz de confusão

A matriz de confusão, é montada com base em um problema de previsão para os casos onde existem uma classe a ser prevista (dicotômica), a tabela é constituída com verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos [35]. Conforme tabela 2.1.

	Classe prevista	
Classe original	*Positivo*	*Negativo*
Positivo	Verdadeiro positivo (TP)	Falso Negativo (FN)
Negativo	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Tabela 2.1: Matriz de confusão

Para avaliação será utilizadas as métricas derivadas da matriz de confusão, onde será medido a acurácia, sensibilidade, especificidade e precisão. A precisão é obtida pela proporção de previsões corretas, sobre os indicados pelo modelo como positivos, ou valor preditivo positivo (VPP), formulada pela Equação 2.6:

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

A taxa de sensibilidade, também chamada de *recall* é obtida pela proporção de verdadeiros positivos sobre os verdadeiros positivos e falsos negativos, em outras palavras, valores originais positivos indicados pelo modelo, sobre o total real de positivos, ou taxa positiva verdadeira (TPV), representada pela Equação 2.7.

$$recall = \frac{TP}{TP + FN} \tag{2.7}$$

A especificidade é a proporção de verdadeiros negativos sobre o total de verdadeiros negativos e falso positivos, ou seja, é a taxa total de negativos dentre os valores originais e previstos, de forma simples, taxa negativa verdadeira (TNV), dada pela Equação 2.8

$$specificity = \frac{TN}{TN + FP} \tag{2.8}$$

A acurácia é definida pela proporção dos resultados verdadeiros (positivos e negativos), sobre o total do número de casos previstos e originais, a acurácia é formulada matematicamente pela equação:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

Em uma abordagem de dados desbalanceados a métrica de avaliação pode interferir nos resultados. Zong et al [44] and Liu et al [43] usam média geométrica (G-mean), cujo objetivo é maximizar a precisão em cada uma das classes, mantendo o equilíbrio. Representado pela Equação 2.10, onde é obtido a raiz do produto do *recall* pela especificidade.

$$G\text{-}mean = \sqrt{recall\,.\,specificity} \tag{2.10}$$

2.9.2 *F-measure*

Em uma classificação binária o *F-score* ou *F-measure* considera os valores de *recall* e da precisão para que então seja calculado a pontuação. Trata-se de uma métrica para o teste de precisão, sendo a pontuação uma média harmônica, com resultado entre 0 e 1, onde o seu melhor valor é obtido em 1. O *F-measure* é formulado pela Equação 2.11, onde $0 < \beta < 1$ ou $\beta > 1$ [35].

$$F\text{-}measure = (1 + \beta^2)\frac{precision\,.\,recall}{(\beta^2.precision) + recall} \tag{2.11}$$

Para $\beta = 1$, caso a precisão e a sensibilidade tenha a mesma ponderação, na literatura é conhecido como F1-score ou F1-measure, formulado pela Equação 2.12.

$$F1\text{-}measure = \frac{2\,.\,precision\,.\,recall}{precision + recall} \tag{2.12}$$

2.9.3 Curva ROC

O *Receiver Operating Characteristic* (ROC), popularmente conhecido na literatura como curva ROC, é uma forma de representação de um sistema classificador binário que resulta em um gráfico de performance, com objetivo de apresentar o desempenho do classificador entre as taxas de erro. A curva ROC é formulada pela taxa de verdadeiros positivos

contra a taxa de falsos negativos. Inicialmente a curva ROC foi concebida para teoria de detecção de sinais, mais tarde começou a ser utilizada por estudos psicológicos e médicos [53] [54] [35]. A *Area Under the Curve* (AUC) apresenta a probabilidade de classificação de instâncias positivas de um determinado classificador. A métrica é utilizada para sumarizar em um valor a curva ROC [55].

2.10 Trabalhos relacionados

As pesquisas foram direcionadas aos estudos que tratam sobre a escravidão moderna. Os resultados demonstram existir artigos com diferentes abordagens, com aspecto jurídicos, sociais e também as pesquisas que modelam estatisticamente este tipo de crime (o trabalho escravo). Os estudos encontrados empregam um escopo global, em uma escala para países e continentes. Os trabalhos com o uso de técnicas estatísticas foram selecionados para o estudo de trabalhos relacionados.

Duas fontes de pesquisas publicaram relatórios na linha de estimativas e previsões para o trabalho escravo. O primeiro [16] faz estimativas e identificação da vulnerabilidade para o trabalho escravo e exploração sexual, de adultos e crianças. Para os casos de trabalho escravo em adultos foi utilizado um censo, realizado entre os anos de 2014 a 2016, aplicado em 48 países, com base nas respostas são mapeadas e quantificadas a escravidão por país e continente. Para os casos de exploração sexual de adultos e crianças, são utilizados a base das estimativas para o trabalho escravo em adultos e também uma base de dados de tráfico humano com 21 (vinte e uma) variáveis, onde 3 modelos são construídos utilizando regressão logística, por meio do *odds ratio* (probabilidade) e não uma classificação binária, podendo então, identificar a taxa de vulnerabilidade. Uma ressalva é feita referente ao calculo do índice de exploração sexual infantil que utiliza os dados do trabalho escravo adulto para estimar o trabalho escravo infantil.

Em maiores detalhes o entendimento do modelo para exploração sexual em adultos é calculado com o *logit*, dada pela Equação 2.13, o resultado será o *odds ratio*, com o uso da exponencial natural é obtido o valor final, dada pela Equação 2.14.

$$ln\frac{p}{1-p} = exp(\alpha + \beta + \lambda) \tag{2.13}$$

$$exp(x) = e^x \qquad \text{onde, } e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \tag{2.14}$$

O segundo trabalho [56], são construídos modelos de previsão de risco para o trabalho escravo forçado e casamento forçado, resultando em um relatório global [5]. Utilizando como método a base de dados do censo realizado entre 2014 a 2016. O trabalho usou

dimensões de variáveis para a vulnerabilidade por seguimento: governança; acesso social (áreas com índice de desnutrição); desigualdade; desassistidos juridicamente (como imigrantes); e efeitos de conflitos. Aplicando pesos em determinados grupos de países e por meio de técnica de regressão logística é calculada a probabilidade de trabalho escravo forçado e casamento forçado. Dois grupos de modelos são desenvolvidos, modelos de níveis individuais e multi-nível [56] [5].

Os modelos de níveis individuais são construídos baseados no *logit*, dada na Equação 2.15, onde o *logit* da probabilidade p é feita para cada registro i, contendo o termo constante β_0, um vetor de variáveis de controle demográfico, de variáveis independentes x com coeficientes desconhecidos β_1, um vetor de variáveis dependentes y com coeficientes desconhecidos β_2, adicionado um termo de erro individual ε_i.

$$ln\left(\frac{p}{1-p}\right)_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \varepsilon_i \tag{2.15}$$

Para os modelos multinível, o objetivo é modelar a nível de país, incluindo os países que não possuem dados sobre trabalho escravo, a formalização é dada pela Equação 2.16, onde são feitas alterações na equação 2.15, adicionando j para classificar individualmente os países, u_j representa um coeficiente aleatório que pode variar por país, v_j representando a pontuação da vulnerabilidade do país, associado a um coeficiente desconhecido β_3.

$$ln\left(\frac{p}{1-p}\right)_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 y_{ij} + \beta_3 v_{ij} + u_j + \varepsilon_{ij} \tag{2.16}$$

Em outros trabalhos [57] [58] são aplicadas técnicas de estimativa para prever a quantidade de escravos em países da Europa, usando como referência os trabalhos de [4] [16]. A Tabela 2.2 apresenta as técnicas utilizadas pelos trabalhos relacionados.

Referência	**Técnica**	**Avaliação**
[16]	*Regressão Logística*	*standard error*
[56]	*Regressão Logística* *Modelo multinível*	*log-likehood* *wald test* *AIC* *BIC* R^2 *AUC*
[57], [58]	*Statistical estimates*	*standard error*

Tabela 2.2: Trabalhos relacionados e técnicas utilizadas

Novas abordagens podem ser propostas, fundamentando-se em uma apuração criteriosa para seleção de variáveis, uma vez que nenhum dos trabalhos apresentaram a aplicação

de técnicas para avaliar a seleção das variáveis. O mesmo censo foi utilizado como fonte primária de dados em ambos os trabalhos relacionados.

Esta pesquisa gerou uma publicação internacional na área de aprendizagem de máquina [59], o que foi motivado por não ter encontrado na literatura nenhum trabalho para o tema em estudo, que utilize técnicas de mineração de dados e comparação entre classificadores. Deste modo, a solução proposta por esta pesquisa promove o estado da arte para o problema levantado na Seção 1. Propondo o uso de modelos de previsão para identificar o risco do trabalho escravo nas cidades, utilizando conjunto de dados reais de operações de resgate do trabalho escravo. Por meio de aprendizagem supervisionada, é proposto de forma comparativa diferentes técnicas de classificação, utilizando seleção de variáveis e uso de diferentes métricas de avaliação de modelos, proporcionando assim, maior exatidão.

Capítulo 3

Solução proposta

Este capítulo aborda o detalhamento das etapas necessárias para o desenho da solução proposta, conforme a metodologia apresentada na Seção 1.3.

O desenho da solução proposta pode ser observado na Figura 3.1, maiores detalhes são fornecidos no decorrer deste capítulo, mas de forma sumarizada, inicialmente foram utilizados dois conjuntos de dados, os quais passaram por um processamento, resultando em um único conjunto de dados classificado de alta dimensionalidade com 237 variáveis, deste modo, foi necessário reduzir a dimensionalidade com técnica de seleção de variáveis, dentre as técnicas testadas foram: *Recursive Feature Elimination*(RFE); univariada; e LASSO. A regularização com LASSO foi a escolhida por apresentar melhores resultados.

Após a seleção das variáveis de interesse, os dados foram divididos de forma aleatória e estratificada em dois conjuntos, treinamento e teste. Os conjuntos de dados são desbalanceados, ou seja, existem em menores quantidades os registros com a classe 1. Com o conjunto de treinamento foi feito o *benchmarking* para a escolha da melhor técnica de balanceamento. Após a escolha da melhor técnica é feito o treinamento do modelo, utilizando o conjunto de dados de treino. Em seguida, o modelo de teste é executado com o melhor modelo treinado e avaliado. Um teste estatístico de robustez do modelo foi construído, utilizando *Bootstrapping* com 1000 repetições. Ao final a probabilidade da previsão é utilizada, obtendo um valor mínimo de 0 (zero) e um valor máximo de 1 (um), resultando no grau do risco.

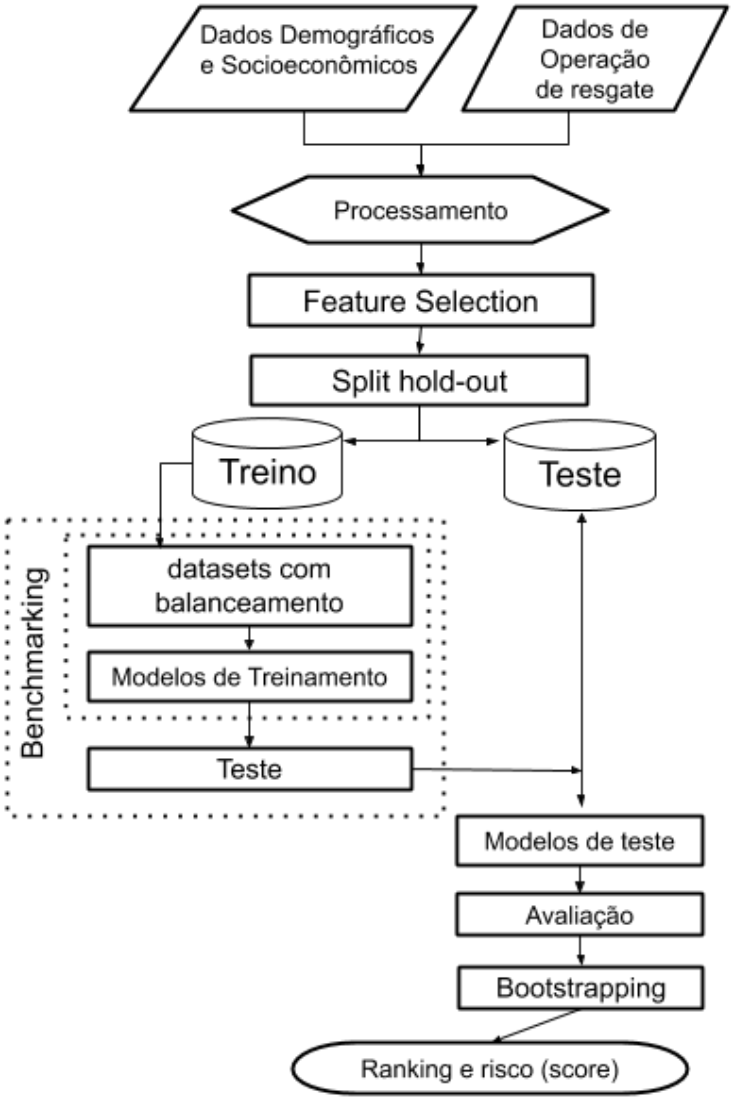


Figura 3.1: Fluxo das etapas da solução proposta.

3.1 Entendimento do negócio

O entendimento do domínio do negócio é iniciado no Capítulo 1, onde é exposto de forma introdutória o contexto no qual o problema se insere. Esta seção complementa as ideias iniciais, apresentando com maior grau de detalhes o funcionamento institucional e os mecanismos reativos atuais para o combate ao trabalho escravo.

É importante ter em mente que a problemática exposta nesta pesquisa, vai além da singularidade institucional, ou seja, a complexidade do tema está conectada a várias obrigações de diversas instituições, sendo um dever de diferentes instituições a disponibilidade e garantias da reparação trabalhista a toda sociedade.

O Ministério Público da União (MPU), compreende a esfera federal, trata-se de uma nomenclatura institucional organizada por meio dos ramos constituídos, são eles: Ministério Público Federal, Ministério Público do Trabalho, Ministério Público do Distrito Federal e Ministério Público Militar, possuem áreas de atuação que complementam-se. Em um aspecto geral têm áreas de atuação distinta, mas existe o escopo comum, o qual todos os ramos são protetores da lei, trabalhando para que os direitos coletivos sejam sempre respeitados e preservados.

O MPT, até o momento, possui 8 (oito) áreas de atuação por meio das coordenadorias nacionais, são elas: Administração Pública, criança e adolescente, fraudes trabalhistas, liberdade sindical, meio ambiente do trabalho, promoção da igualdade, trabalho escravo, trabalho portuário e aquaviário. Em detalhes essas áreas atuam da seguinte maneira:

- A coordenadoria da Administração Pública possui como objetivo a promoção de ações integradas de combate às irregularidades trabalhistas na Administração Pública, atuando ativamente nos temas relativos ao trabalho na Administração Pública Direta e Indireta, meio ambiente de trabalho dos servidores estatutários, celetistas e terceirizados, concurso público, terceirização, nulidade da contratação além da responsabilidade solidária ou subsidiária da Administração Pública nas questões trabalhistas.

- Atuação do combate a exploração do trabalho de criança e adolescente. Objetiva-se na promoção de políticas públicas para prevenção e a erradicação do trabalho infantil informal; efetivação da aprendizagem; proteção de atletas mirins; trabalho infantil artístico; exploração sexual comercial; autorizações judiciais para o trabalho antes da idade mínima; trabalho infantil doméstico e atuação em fiscalizações do trabalho de crianças em lixões.

- A coordenadoria de fraudes trabalhistas possui como principal foco às fraudes por meio de cooperativas intermediadoras de mão de obra, terceirizações ilegais, falsas sociedades de empregados e outras fraudes relacionadas a questões trabalhistas, utilizadas para fraudar a lei trabalhista.

- A liberdade sindical e a pacificação dos conflitos coletivos de trabalho também é uma área de atuação do MPT, garantindo os direitos sindicais, assegurando o direito de greve, atuando como mediador em conflitos coletivos de trabalho.

- Referente ao meio ambiente de trabalho, a instituição foca na segurança e saúde do trabalhador, atuando na redução do risco de acidentes e melhorias das condições de trabalho.

- Outra frente de atuação, que tem relação direta com o bem estar do trabalhador, é a da coordenação da igualdade de trabalho, onde preza pelo combate a discriminação no trabalho, na inclusão das pessoas com deficiência ao ambiente de trabalho.
- A atuação da coordenadoria de combate ao trabalho escravo, faz investigações de situações em condições análogas às de trabalho escravo, levando em consideração as ocorrências de servidão por dívidas, trabalho forçado, jornadas exaustivas e condições degradantes de trabalho, alojamento precarizado, ausência de água potável, alimentação inadequada, desrespeito às normas de segurança e saúde do trabalho, falta de registro, maus tratos e violência.
- Por fim, a coordenadoria do trabalho portuário e aquaviária investiga irregularidades em trabalhos dos portos, bem como a formalização dos trabalhadores da pesca, navegações marítimas e fluviais.

A fiscalização é um dos eixos que atuam de forma direta no combate a este tipo de crime, promovida por meio de auditorias, onde as operações são motivadas principalmente por meio de denúncias, mas apenas uma parte é atendida devido a grande demanda reportada. A ausência de mecanismos para auxiliar na escolha das regiões a serem investigadas é uma realidade.

Atualmente as principais entidades governamentais atuantes em alguma das etapas da fiscalização do trabalho escravo e tráfico humano são: Secretaria do Trabalho, MPT, MPF e Tribunais do Trabalho. A Figura 3.2 ilustra de forma simplificada a relação das principais entidades envolvidas no processo para atender a demanda da fiscalização.

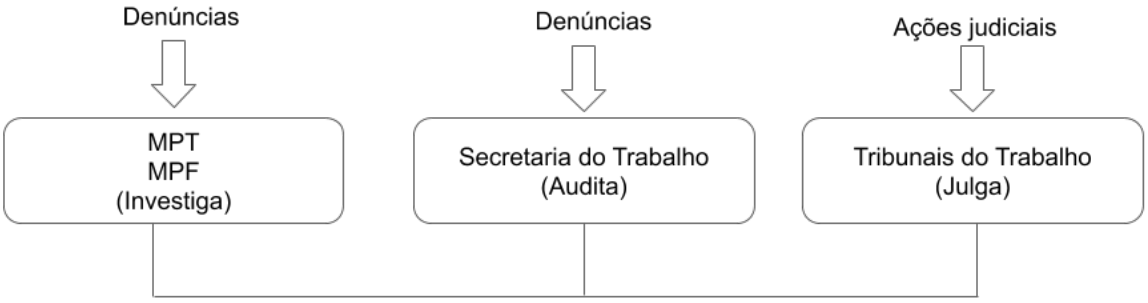


Figura 3.2: Relacionamento entre denúncias e entidades envolvidas.

As denúncias recebidas pelas entidades são verificadas, quando constatado a validade da denúncia, um processo interno é iniciado. Em casos de denúncias procedentes os fatos são investigados. É comum, o trabalho conjunto entre entidades, assim melhores resultados são alcançados. Ao final da investigação ou da auditoria, se materializado o

crime, as devidas medidas são tomadas com base na lei vigente, podendo ações judiciais serem instauradas em Tribunais de Justiça do Trabalho.

Por se tratar de um crime que envolve a dignada da pessoa humana, o trabalho escravo é combatido por outras instituições além do MPT. A Secretaria do Trabalho é uma instituição federal que atual na fiscalização deste crime. Assim como no MPT, a Secretaria do Trabalho recebe denúncias por canais de atendimento ao público e também de pedidos de fiscalização diretamente do MPT.

Embora o procedimento de atuação e sistemas de denúncia sejam independentes, entre si, tanto o MPT como a Secretaria do Trabalho, fazem uma pré triagem das denúncias encaminhas. Caso os critérios mínimos sejam procedentes, a denúncia é analisada podendo virar um procedimento investigatório ou fiscalizatório, conforme for a instituição.

Outros agentes públicos estão envolvidos neste contexto, como a polícia federal que atua em parceria com a Secretaria do Trabalho e o Ministério Público. Realizando diligências estratégicas e de apoio aos auditores, procuradores e corpo técnico. Em muitas situações envolve risco a segurança dos envolvidos.

Organizações não governamentais, cooperativas e sindicatos, também atuam de forma presente, pois possuem contato frequente com trabalhadores, principalmente os rurais. Essas entidades fomentam as instituições com informações importantes para fiscalização e investigação. E em muitas circunstancias são agentes promotores das denúncias.

Em um formato geral e simplificado, a figura 3.3, representa as fases do processo da denúncia. Inicialmente as denúncias chegam por meio de entidades representativas ou pela própria pessoa acometida da suposta violação trabalhista, nesta etapa é feito o registro da denúncia. Em seguida, é verificado a viabilidade da denúncia, tais como datas, nomes e o objeto da reclamação. Caso seja uma denúncia viável, ela parte para ser analisada no qual é feito o filtro e categorização, após a etapa da análise caso a denúncia persista coerente é dado o prosseguimento, formalizando ação investigativa e ou fiscalizadora.

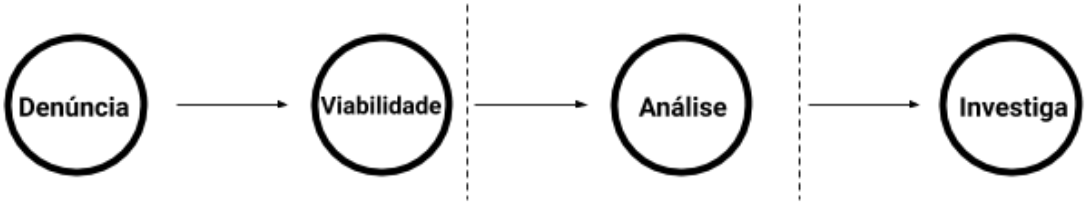


Figura 3.3: Fluxo do sistema de denúncia.

Atualmente, não existem mecanismos para priorizar de forma cientifica regiões ou áreas com maior vulnerabilidade. Também não existem métricas fundamentadas para medir o nível da erradicação do trabalho escravo contemporâneo e assim obter a evolução dos efeitos das ações executadas.

3.2 Entendimento dos dados

Dois conjuntos de dados são utilizados para formar a base classificada com as ocorrências de trabalho escravo. O primeiro conjunto de dados analisado foram os dados de operações de resgate, fornecidos pelo MPT[1], com quatro dimensões, a saber: *nm_municipio, nm_uf, ano, trabalhadores_resgatados*, descritos respectivamente por nome da cidade, código do Estado, ano do resgate e o número de trabalhadores resgatados, os dados são de 2003 à 2018, totalizando 3.318 registros, com 44.238 trabalhadores resgatados. O conjunto de dados das operações de resgate foi utilizado para classificar as cidades, como é apresentado na Seção 3.3. Os dados com os registros de trabalhadores resgatados também podem ser obtidos por meio do portal radar[2].

O segundo conjunto de dados analisado, refere-se ao Atlas do Desenvolvimento Humano dos Municípios do Brasil[3], fornecido pelo PNUD em parceria com o Instituto de Pesquisa Econômica Aplicada (Ipea) e a Fundação João Pinheiro. A base de dados do Atlas traz indicadores de demografia, educação, renda, trabalho, habitação e vulnerabilidade e foram construídos com base nos dados extraídos dos Censos Demográficos de 1991, 2000 e 2010, que em formato bruto representa 16.695 registros, com 237 dimensões (indicadores). Os indicadores oferecem um panorama do cenário socioeconômico do país, promovendo a reflexão sobre o desenvolvimento nacional.

Um pré-processamento é realizado, utilizando como critério o último Censo realizado de 2010, desta forma foram selecionados os conjuntos de dados do Atlas contendo os registros de 2010, resultando nas dimensões: 5565x237, neste momento não foram eliminadas nenhuma variável. Após a separação dos registros de interesse, não foram encontrados dados faltantes.

3.3 Preparação dos Dados

A preparação dos dados utiliza como entrada os dois conjuntos de dados pré-processados, conforme descrito na Seção 3.2. Nesta etapa, é feita a normalização textual das variáveis categóricas, pois são usadas para gerar um atributo derivado identificador em ambos conjuntos de dados, responsável pela identificação de cada registro, por meio dos atributos: Estado e nome da cidade.

Com o atributo identificador é possível criar o novo conjunto de dados contendo todos os campos dos dados do Atlas (indicadores), acrescentando duas colunas: número de

[1]Observatorio: `http://observatorioescravo.mpt.mp.br`, acessado em 25/09/2018.

[2]Radar: `https://sit.trabalho.gov.br/radar/`

[3]Atlas:`http://atlasbrasil.org.br/2013/data/rawData/atlas2013_dadosbrutos_pt.xlsx`, acessado em 27/09/2018.

trabalhadores resgatados e variável alvo do trabalho escravo. O critério de classificação utilizado foi determinado por: enquanto $(T > 0)$, então a variável *escravo* deve receber valor igual a 1, o contrário, $escravo = 0$, onde T representa o número de trabalhadores resgatados na cidade.

Os dados são classificados em cidades que ocorreram o trabalho escravo, com o registro de 767 (13.78%) casos confirmados, de valor 1, e cidades que não ocorreram o trabalho escravo, com 4798 (86.22%) casos não confirmados e que não houveram relatos, de valor 0. A Figura 3.4 apresenta o novo conjunto de dados rotulado e construído com base nos conjuntos de dados de entrada.

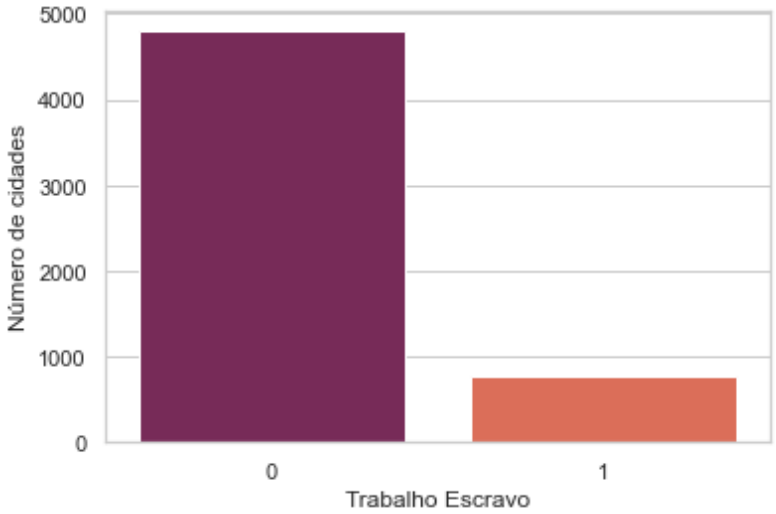


Figura 3.4: Novo conjunto de dados rotulado.

Como pode ser observado o novo conjunto de dados rotulado é desbalanceado. Uma abordagem ao tratamento desta condição e o estudo comparativo das técnicas para tratamento de dados desbalanceados é apresentado na Seção 3.4.2.

A seleção de variáveis independentes é necessária, pois o conjunto de dados possui 237 variáveis, sendo 232 contínuas e 5 categóricas. A Seção 3.3.1 apresenta o algoritmo responsável por selecionar as variáveis de interesse para o modelo.

3.3.1 Algoritmo para seleção de variáveis

Para a redução de dimensionalidade é aplicada a técnica *embedded* com regularização L1 (lasso), por apresentar melhores resultados diante de outras técnicas testadas. A aplicação do algoritmo 1 resultou na seleção de 16 variáveis, que estão apresentadas na Tabela 3.1. A ideia por trás do algoritmo é selecionar os coeficientes diferentes de zero, por meio de um modelo de aprendizagem.

O conjunto de dados final apresenta dimensões de 5565 x 17, onde 5565 representa o número de cidades, e na dimensão das variáveis selecionadas foi adicionada a variável alvo (dependente), totalizando 17 dimensões.

Algorithm 1: seleciona variáveis com uso de aprendizagem de máquina

Result: Retorna variáveis de importância para o modelo

Input: X uma lista de variáveis independentes, $X = \langle x_1, x_2, \ldots, x_n \rangle$ e y variável preditora.

Output: *vet_naozero* vetor de variáveis selecionadas

Function `FeatureSelection(`X, y`):`

```
vet_variaveis ⟵ pega_variaveis(X)
modelo ⟵ LogisticRegression(C = 0.1, penalty =' l1')
modelo_treinado ⟵ modelo.fit(X, y)
vet_zero ⟵ [...]
vet_naozero ⟵ [...]
foreach i ∈ range(vet_variaveis) do
    coeficiente ⟵ modelo_treinado.pega_coeficiente[0, i]
    if coeficiente == 0 then
        vet_zero.adiciona(pega_variavel(X[i]))
    else
        vet_naozero.adiciona((coeficiente, pega_variavel(X[i]) ))
    end
end
Return orderna(vet_naozero);
```

End Function

Alguns algoritmos apresentam melhores resultados com baixa dispersão de dados, como o caso da regressão logística. Como forma de padronizar e tornar os dados mais homogêneos, foi aplicada a normalização linear, com o propósito de harmonizar a escala de atributos para a faixa de 0 a 1, para tanto, a Equação 3.1 é aplicada.

$$\frac{X_i^k - X_{min}^k}{X_{max}^k - X_{min}^k} \tag{3.1}$$

Onde, X representa o valor do conjunto de dados, i a observação corrente analisada e k o atributo (coluna) do conjunto de dados, assim,X_i^k é o valor em que se deseja colocar na escala, já X_{max}^k representa o valor máximo do atributo k e X_{min}^k o valor mínimo.

A descrição dos dados encontra-se na Tabela 3.2, onde estão apresentadas a contagem de frequência para cada variável selecionada, os valores máximos, mínimos, média, desvio padrão e respectivas distribuições por quartis.

Tabela 3.1: Variáveis selecionadas

Nome da variável	Descrição
prentrab	Percentual da renda proveniente de rendimentos do trabalho
parede	Percentual da população que vivem em domicílios com paredes que não são de alvenaria ou adequadas
t_atraso_1_basico	percentual da população de 6 a 17 anos de idade frequentando o ensino básico, que tem 1 ano de atraso
t_agua	percentual da população que vive com água encanada
pren10ricos	Percentual da renda total apropriada pelos 10% da população com maior renda domiciliar per capita
t_super25m	Percentual da população de 25 anos ou mais com superior completo
trabsc	Percentual de ocupados de 18 anos ou mais que são empregados sem carteira
t_freq4a5	Taxa da população de 4 a 5 anos de idade que estava frequentando a escola
ren2	Percentual dos ocupados com rendimento de até 2 salários mínimos
pren40	Percentual da renda total apropriada pelos 40% da população com menor renda domiciliar per capita
rind	Renda domiciliar per capita média dos extremamente pobres
pren20	Percentual da renda total apropriada pelos 20% da população com menor renda domiciliar per capita
t_flpre	Taxa de frequência à pré-escola
p_transf	Percentual dos ocupados na indústria de transformação
ren1	Percentual dos ocupados com rendimento de até 1 salário mínimo
t_env	Taxa de envelhecimento

Tabela 3.2: Análise descritiva.

	prentrab	parede	t_atraso_1_basico	t_agua	pren10ricos	t_super25m	trabsc	t_freq4a5	ren2	pren40	rind	pren20	t_flpre	p_transf	ren1	t_env
count	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0
mean	68.48	5.36	18.29	85.59	38.20	5.49	25.22	78.45	82.37	12.11	32.03	3.72	54.77	9.61	39.30	8.39
std	10.79	9.41	4.01	14.7	5.91	3.25	9.85	15.46	10.35	3.25	9.60	1.54	15.92	8.92	21.58	2.42
min	27.43	0.00	4.33	0.15	22.26	0.28	3.03	13.03	36.49	0.00	0.00	0.00	3.91	0.00	4.53	1.46
25%	61.06	0.41	15.72	79.65	34.17	3.24	17.64	70.37	74.90	9.920	27.44	2.45	44.26	3.33	19.60	6.78
50%	70.58	1.64	18.80	90.28	37.62	4.81	24.75	82.14	83.86	12.17	32.51	3.72	55.56	6.53	35.81	8.38
75%	76.60	5.82	21.03	96.26	41.63	6.95	32.04	90.13	91.57	14.45	37.09	4.90	66.07	13.31	58.55	9.96
max	95.24	82.74	32.31	100.00	75.34	33.68	62.23	100.00	99.140	22.50	70.00	9.26	100.00	65.11	89.33	20.42

3.4 Modelagem

A etapa de modelagem visa à construção do modelo, após a preparação dos dados, uma vez obtido as variáveis de interesse na Seção 3.3, o processo para escolha do melhor modelo é iniciado para então obter os modelos de previsão de risco do trabalho escravo.

De acordo com o levantamento bibliográfico, conforme Seção 2.5 e as pesquisas prévias realizadas sobre trabalhos com propósitos alinhados a problemática levantada nesta pesquisa, apresentado na Seção 2.10, aponta a técnica de LR como mais utilizada. No entanto, por se tratar de um tema inédito ao uso de aprendizagem de máquina, não foi identificado o uso de outros classificadores.

Um classificador adicional, GBM, foi usado na comparação. Foram considerados muitos outros classificadores tradicionais, no entanto, foi feita a escolha por utilizar dois classificadores no estudo, LR e GBM, por apresentarem melhores resultados, assim como para simplificar a análise de *benchmarking* entre as técnicas de balanceamento.

3.4.1 Parametrização

Os parâmetros dos algoritmos podem ser ajustados para se obter melhor desempenho dos modelos. Alguns algoritmos são conhecidos por proporcionar diversas combinações, como é o caso do GBM. Neste estudo foi utilizada a parametrização padrão da biblioteca *scikit-learn*.

Para o classificador LR, foram utilizados os seguintes parâmetros: **penalidade**, que compreende em *L1*, *L2*, *elasticnet* e *none* (sem penalidade); **Dual**, que é utilizado em conjunto com a penalidade L2, de tipo boleano, com o padrão *false*; **Tol** é um critério de parada, de tipo *float*; **C**, de tipo *float*, quanto menor o valor mais alta será a regularização. **Fit_intercept**, de tipo boleano, especifica se uma constante deverá ser adicionada a uma função de decisão. O parâmetro **intercept_scaling** é util somente quando utiliza um problema de resolução *liblinear* utilizado em conjunto com o parâmetro *self.fit_intercept*, se o valor deste parâmetro aumentar ocorrerá a diminuição da regularização em uma variável sintética. O **class_weight** refere-se ao peso da classe e, caso não haja uma atribuição prévia, todas as classe devem ter peso 1 (um). Para salvar o estado da instância utiliza-se o parâmetro **random_state**. O parâmetro **solver** tem fundamental importância, que pode ser escolhido entre *newton-cg*, *lbfgs*, *liblinear*, *sag*, saga, são algoritmos de otimização de problema. Cada um com uma característica. A exemplo, o *lbfgs* utiliza uma estimativa para a matriz de Hessian, possui como característica o armazenamento apenas de alguns vetores que representam a aproximação implícita, este algoritmo é análogo ao método de *newton-cg*, mas com uma matriz aproximada. Outro parâmetro importante é o **max_iter** que oferece o número máximo de iterações para os algoritmos de otimização de problemas. Outros parâmetros são **multi_class**, **verbose**, **warm_start**, **n_jobs** e **l1_ratio**. A Tabela 3.3 apresenta os valores dos parâmetros utilizados para o classificador LR.

Tabela 3.3: Parâmetros do algoritmo de classificação - Regressão Logística

parâmetro	valor
penalty	l2
dual	False
tol	0.0001
C	1.0
fit_intercept	True
intercept_scaling	1
class_weight	None
random_state=	one
solver	'lbfgs'
max_iter	100
multi_class	'auto'
verbose	0
warm_start	False
n_jobs	None
l1_ratio	None

Os principais parâmetros do algoritmo GBM são: **loss** que é a função de perda a ser otimizada, que pode ser de dois tipos *deviance* (desvio) ou exponencial; **learning_rate** a taxa de aprendizagem reduz a contribuição de cada árvore; **n_estimators** trata-se do número de estágios a serem executados, um grande número resulta em melhor desempenho; **criterion** é usado para aferir a qualidade de um *split*, as métricas suportados são *friedman_mse* para o erro quadrático médio com pontuação, *MSE* para o erro quadrático médio e *MAE* para erro absoluto médio. A Tabela 3.4 apresenta os valores dos parâmetros utilizados para o classificador GBM.

Tabela 3.4: Principais parâmetros do algoritmo de classificação - GBM

parâmetro	valor
loss	'deviance'
learning_rate	0.1
n_estimators	100
criterion	'friedman_mse'

3.4.2 *Benchmarking*

Por se tratar de um problema de classificação em dados desbalanceados, o estudo levantado na Seção 2.6 sugere o balanceamento dos dados em treinamento. Devido à ausência de trabalhos de contexto similar a este estudo, optou-se por realizar o *benchmarking* para a escolha da melhor técnica de balanceamento de dados para a realidade do problema.

Por meio das médias aritméticas das técnicas LR e GBM, para $k = (1, ..., n)$, a média é dada pela Equação 3.2, onde k é a técnica de classificação e n é a quantidade máxima de técnicas para classificação, no estudo representado por $n = 2$, e $metric_k$ é o resultado da métrica para cada técnica de classificação. O método de balancemaneto é escolhido usando o critério da melhor média.

$$\sum_{k=1}^{n} \frac{metric_k}{n} \tag{3.2}$$

As médias são obtidas para cada métrica (acurácia, auc, precisão, *recall*, *f1* e *G-mean*), conforme apresentado nas Tabelas: 4.3 e 4.4.

3.4.3 Treinamento e Teste

O conjunto de dados foi dividido em treinamento (70%) e teste (30%), de acordo com a abordagem *hold-out*. O teste foi realizado em um conjunto de dados desbalanceado (30%). Por se tratar de um problema de classificação em dados desbalanceados, o estudo levantado na Seção 2.6 sugere o balanceamento de dados em treinamento. Devido à ausência de trabalhos com o contexto semelhante ao presente estudo, optou-se por realizar o *benchmarking* para escolher a melhor técnica de balanceamento de dados para a realidade do problema.

O treinamento foi realizado para um conjunto desbalanceado (base) usando as técnicas de balanceamento RUS, ClusterCentroids, ROS, SMOTE, SMOTEENN, SMOTETomek e ADASYN. Com a escolha do conjunto de dados a ser utilizada foi feito o treinamento utilizando validação cruzada com *10-folds*. O modelo foi testado utilizando os dados de teste (30%) que é desconhecida pelo modelo treinado. O fluxo dos processos da solução proposta é apresentado na Figura 3.1.

No total, 16 modelos (8x2) foram avaliados, baseado em 8 diferentes conjuntos de dados e 2 classificadores. Os resultados são apresentados no Capítulo 4.

3.5 Avaliação dos Modelos

Esta etapa é destinada em avaliar o desempenho dos modelos. Na prática a avaliação é integrada com a seção de modelagem, pois é a partir das avaliações obtidas que melhorias são projetadas nos modelos. No estudo, os processos que estão relacionados a esta seção são: conjunto de dados para treinamento, tratamento de dados desbalanceados, seleção de modelos com validação cruzada e finalmente as respectivas Avaliações.

Os métodos de balanceamento são avaliados, como descrito na Seção 3.4.2. Os valores médios são obtidos para cada métrica (acurácia, auc, precisão, *recall*, *f1* e *G-mean*), conforme formulado na Equação 3.2.

Com a escolha da técnica de balanceamento a ser usada, o treinamento é feito usando validação cruzada com *10-folds*. O modelo é testado usando o conjunto de teste (30%) ao qual é desbalanceado e desconhecido para o modelo de treinamento.

A avaliação final do modelo selecionado leva em conta como fator decisivo as métrica de acurácia e AUC. Os resultados são apresentados na Tabela 4.5. Com o objetivo de testar o modelo final, um teste estatístico de robustez é aplicado utilizando o *Bootstrapping* com 1000 repetições.

Como parte da avaliação do modelo final, neste estudo utilizamos o *Bootstrapping* para obter amostras aleatórias com reposição, em seguida é calculado a estatística da população gerada. Ao final é calculado o intervalo de confiança. Este passos descrevem um teste de robustez para o modelo proposto.

O Algoritmo 2 de forma simplificada, reproduz o teste estatístico com intervalo de confiança de 95% utilizando o *Bootstrapping* em uma abordagem de aprendizagem de máquina. São realizadas 1000 (mil) repetições e para cada repetição é feita a amostragem com divisão dos dados em treino e teste, de forma aleatória e com reposição. O modelo é treinado e aplicado a predição, o resultado da acurácia é armazenado em uma lista. Ao final do ciclo de repetições o intervalo de confiança é calculado. O algoritmo retorna a

acurácia máxima e mínima para o intervalo de confiança estabelecido.

Algorithm 2: Faz reamostragem e calcula o intervalo de confiança em um modelo de aprendizagem

Result: Retorna o intervalo de confiança de uma distribuição de amostragem

Input: X conjunto de dados de variáveis independentes, $X = \langle x_1, x_2, \ldots, x_n \rangle$ e y conjunto com variável dependente.

Output: max maior valor em um intervalo de confiança, min menor valor em um intervalo de confiança

Function `Bootstrapping_Learning(`X, y`):`

$resultados \longleftarrow lista()$
$n\_iteracoes \longleftarrow 1000$
$size\_treino \longleftarrow 0.3$
$size\_teste \longleftarrow 0.7$
$i \longleftarrow 1$
while $i \leq n\_iteracoes$ **do**
  $amostra\_treino \longleftarrow$
   $amostragem\_aleatoria\_reposicao(X, y, size\_treino)$
  $amostra\_teste \longleftarrow amostragem\_aleatoria\_reposicao(X, y, size\_teste)$
  $modelo \longleftarrow algoritmo\_classificacao()$
  $modelo\_treinado \longleftarrow modelo.fit(amostra\_treino)$
  $modelo\_previsao \longleftarrow modelo\_treinado.predict(amostra\_teste)$
  $acuracia \longleftarrow metrica(modelo\_previsao)$
  $resultados \longleftarrow resultados.add(acuracia)$
  $i++$
end
$max, min \longleftarrow Calcula\_IntervaloConfianca(resultados, 0.95)$
Return max, min;
End Function

Capítulo 4

Resultados

Neste capítulo são apresentados os resultados obtidos por cada etapa desenvolvida na solução proposta, organizados primeiramente pelos resultados do estudo comparativo das técnicas de balanceamento. Em seguida são apresentados os gráficos com os resultados de curva ROC dos modelos, cuja técnica de balanceamento foi escolhida previamente. Após, uma tabela com avaliação dos modelos por métrica é apresentada, onde pode ser identificado o melhor modelo no estudo. Ao final, os resultados do teste estatístico de robustez é observado, seguido pelo respectivo gráfico. Como parte dos resultados, a Figura 4.3 facilita a visualização que expõe os níveis e a distribuição do risco de escravidão encontrados.

Os resultados do estudo comparativo das técnicas de balancemaneto são apresentados nas Tabelas 4.1 e 4.2, representando respectivamente os resultados de treino e teste.

A Tabela 4.1 expõe os resultados em treinamento do *benchmarking*. É possível observar melhores resultados com a técnica de SMOTEEN para todas as métricas aplicadas. Indicando que essa técnica tem excelente comportamento para um conjunto de dados conhecido pelo modelo. O algoritmo de classificação GBM é superior aos resultados de LR.

Tabela 4.1: *Benchmarking* das técnicas de balanceamento em um dataset de treino

	accuracy		**auc**		**precision**		**recall**		**F1**		**G-mean**	
Método	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*
base	0.8652	0.8652	0.8105	0.8030	0.5953	0.5980	0.2037	0.2074	0.3022	0.3057	0.4461	0.4499
RUS	0.7279	0.7092	0.7972	0.7821	0.7240	0.7033	0.7387	0.7316	0.7302	0.7154	0.7280	0.7089
ClusterCentroids	0.7521	0.7468	0.8235	0.8331	0.7554	0.7533	0.7494	0.7387	0.7509	0.7429	0.7522	0.7468
ROS	0.7473	0.8303	0.8144	0.9019	0.7469	0.8069	0.7488	0.8690	0.7475	0.8365	0.7473	0.8294
SMOTE	0.7618	0.8126	0.8304	0.8900	0.7547	0.7929	0.7763	0.8471	0.7649	0.8187	0.7617	0.8119
SMOTEEN	0.8436	0.8938	0.9129	0.9567	0.8636	0.8943	0.8912	0.9417	0.8770	0.9173	0.8251	0.8753
SMOTETomek	0.7660	0.8226	0.8330	0.8943	0.7589	0.7996	0.7804	0.8615	0.7693	0.8293	0.7659	0.8217
ADASYN	0.7369	0.8032	0.8018	0.8739	0.7308	0.7763	0.7514	0.8536	0.7409	0.8130	0.7368	0.8015

Quando o conjunto de dados é desconhecido para o modelo, os resultados mudam, diferenciando do que foi apontado anteriormente na Tabela 4.1. A Tabela 4.2 expõe os

Tabela 4.2: *Benchmarking* das técnicas de balanceamento em um dataset de teste

	accuracy		**auc**		**precision**		**recall**		**F1**		**G-mean**	
Método	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*
base	0.8790	0.8766	0.8030	0.8029	0.5468	0.5135	0.1682	0.1826	0.2573	0.2695	0.4061	0.4221
RUS	0.7425	0.7155	0.8044	0.7928	0.2881	0.2645	0.7259	0.7211	0.4125	0.3870	0.7353	0.7179
ClusterCentroids	0.6958	0.6293	0.7884	0.7532	0.2508	0.2157	0.7259	0.7500	0.3728	0.3351	0.7085	0.6775
ROS	0.7371	0.7742	0.8043	0.8054	0.2824	0.3066	0.7211	0.6442	0.4059	0.4155	0.7302	0.7146
SMOTE	0.7353	0.7550	0.8004	0.7994	0.2784	0.2866	0.7067	0.6490	0.3994	0.3976	0.7228	0.7070
SMOTEEN	0.6137	0.6634	8022	0.8035	0.2230	0.2449	0.8461	0.8173	0.3530	0.3769	0.7009	0.7241
SMOTETomek	0.7383	0.7622	0.7998	0.7918	0.2793	0.2895	0.6971	0.6250	0.3989	0.3957	0.7202	0.6990
ADASYN	0.7149	0.7389	0.7981	0.7930	0.2697	0.2816	0.7548	0.7067	0.3974	0.4027	0.7317	0.7248

resultados em teste (validação), e é observado que para a métrica de acurácia o uso sem técnica de balanceamento (base) destaca-se, já para a métrica de AUC a técnica de ROS é melhor para os dois algoritmos de classificação. Com a métrica de precisão, novamente o destaque é o não uso de técnica de balanceamento (base), mas com a métrica de recall, é observado que a técnica de SMOTEEN possui bons resultados se comparado as outras técnicas de balanceamento. Para a métrica de F1, ROS possui melhor resultado absoluto, e por fim, com a métrica de G-mean as técnicas de RUS, ROS e ADASYN destacam-se.

A Tabela 4.3 apresenta as médias do *benchmarking* em treinamento, obtidas dos algoritmos de classificação (LR e GBM) para cada tipo de balanceamento em treinamento. Os resultados sugerem que as técnicas de balanceamento SMOTE-ENN, SMOTETomek, SMOTE e ROS tiveram melhores resultados para as métricas de AUC, F1 e G-mean.

Tabela 4.3: Médias das métricas por tipo de balanceamento (amostragem) - Treino

Avaliações das médias para o tipo de amostra - Treinamento						
Tipo	$\overline{acc}$	$\overline{\boldsymbol{auc}}$	$\overline{precision}$	$\overline{recall}$	$\overline{f1}$	$\overline{G\text{-}mean}$
base	0.8643	0.8062	0.5775	0.1994	0.2951	0.4410
RUS	0.7136	0.7820	0.7104	0.7369	0.7211	0.7148
ClusterCentroids	0.7566	0.8291	0.7589	0.7558	0.7566	0.7584
ROS	0.7906	0.8584	0.7786	0.8108	0.7940	0.7902
SMOTE	0.7867	0.8585	0.7732	0.8112	0.7914	0.7865
SMOTEEN	0.8691	0.9353	0.8800	0.9158	0.8975	0.8512
SMOTETomek	0.7928	0.8630	0.7786	0.8180	0.7975	0.7924
ADASYN	0.7706	0.8381	0.7535	0.8045	0.7778	0.7698

A seleção da melhor técnica para o balanceamento dos dados para o contexto deste trabalho foi validada com o uso do conjunto de dados de teste, que representa 30% dos dados. A Tabela 4.4 apresenta os resultados, mostrando que uma base desbalanceada (base) apresenta baixos índices de recall, f1 e consequentemente baixo valor para G-mean,

sugerindo que o uso unicamente de uma métrica não é uma forma segura de avaliação para o conjunto de dados desbalanceados.

Ao contrário, quando testado os conjuntos de dados com as técnicas de balanceamento, os valores de recall, f1 e G-mean são melhorados. A melhor técnica avaliada por acurácia, AUC, F1 e G-mean foi ROS para o contexto proposto.

Tabela 4.4: Médias das métricas por tipo de balanceamento (amostragem) - Teste

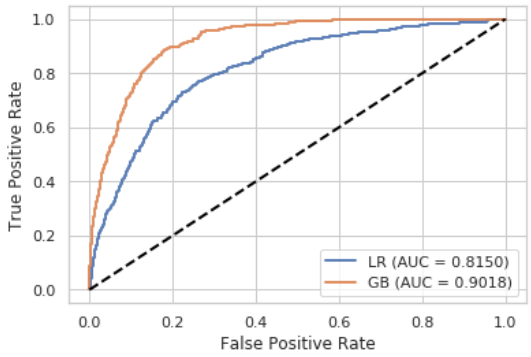
Avaliações das médias para o tipo de amostra - Teste						
Tipo	$\overline{acc}$	$\overline{\boldsymbol{auc}}$	$\overline{precision}$	$\overline{recall}$	$\overline{F1}$	$\overline{G\text{-}mean}$
base	0.8781	0.8032	0.5337	0.1754	0.2639	0.4141
RUS	0.7296	0.7989	0.2784	0.7331	0.4034	0.7310
ClusterCentroids	0.6416	0.7625	0.2259	0.7524	0.3464	0.6850
ROS	0.7556	**0.8048**	0.2945	0.6826	**0.4107**	**0.7224**
SMOTE	0.7452	0.7997	0.2825	0.6778	0.3985	0.7149
SMOTEEN	0.6389	0.8028	0.2341	0.8317	0.3652	0.7127
SMOTETomek	0.7502	0.7958	0.2844	0.6610	0.3973	0.7096
ADASYN	0.7272	0.7957	0.2759	0.7307	0.4003	0.7284

4.1 Avaliação dos modelos selecionados

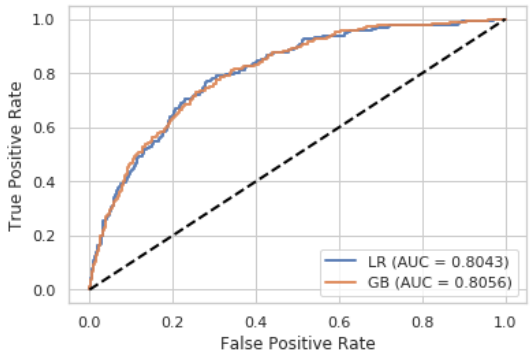
De acordo com as avaliações apresentadas na Seção 3.5, a técnica de balanceamento ROS obteve melhores índices, sendo escolhida para compor os modelos que representam a solução proposta. Os modelos finais, utilizam os algoritmos de classificação LR e GBM.

A Fig.4.1 apresenta a *Receiver Operating Characteristic* (ROC) *curves* de classes positivas, no estudo correspondem aos valores binários 1, dos modelos de treinamento e teste, ambas usam balanceamento ROS. Como pode ser observado, nos modelos de treinamento a *Area Under the Curve* (AUC) é inferior para o modelo de LR se comparado ao modelo GBM que possui maior AUC. Com um valor superior de AUC indica melhor capacidade de previsão do modelo, com melhores taxas de verdadeiros positivos.

Os resultados das métricas obtidas para os dois modelos na validação em teste são apresentadas na tabela 4.5. É possível observar que GBM apresenta melhores resultados. O modelo LR possui valores com as métricas de recall e G-mean superior ao GBM. O modelo GBM possui melhor performance com resultados superiores na acurácia, AUC e F1.



(a) Modelos de treinamento



(b) Modelos de teste

Figura 4.1: As curvas ROC para os modelos nos conjuntos de dados ROS

Tabela 4.5: Avaliação dos modelos do conjunto de dados em teste

Modelo	*accuracy*	*AUC*	*precision*	*recall*	*f1*	*G-mean*
LR	0.7371	0.8043	0.2824	0.7211	0.4059	0.7302
GBM	0.7742	0.8056	0.3066	0.6442	0.4155	0.7146

4.2 Teste estatístico

O algoritmo do teste estatístico foi apresentado na Seção 3.5, o qual foi exposto de forma simplificada os passos seguidos para realizar uma inferência por intervalo de confiança, utilizando o *Bootstrapping* com 1000 iterações. O teste estatístico foi aplicado ao melhor modelo selecionado, GBM com ROS.

A Figura 4.2 apresenta o histograma da acurácia em um intervalo de confiança de 0.95, com 1000 iterações. Os resultados mostram que a acurácia se manteve entre 87.5 e 87.8 para um intervalo de confiança de 95%.

O resultado da Figura 4.2 também apresenta uma distribuição Gaussiana, ou seja, uma distribuição normal, portanto um forte indicativo de que os resultados do teste estatístico de robustez estão sem viés ou *overfitting*. Confirmando a coerência do modelo.

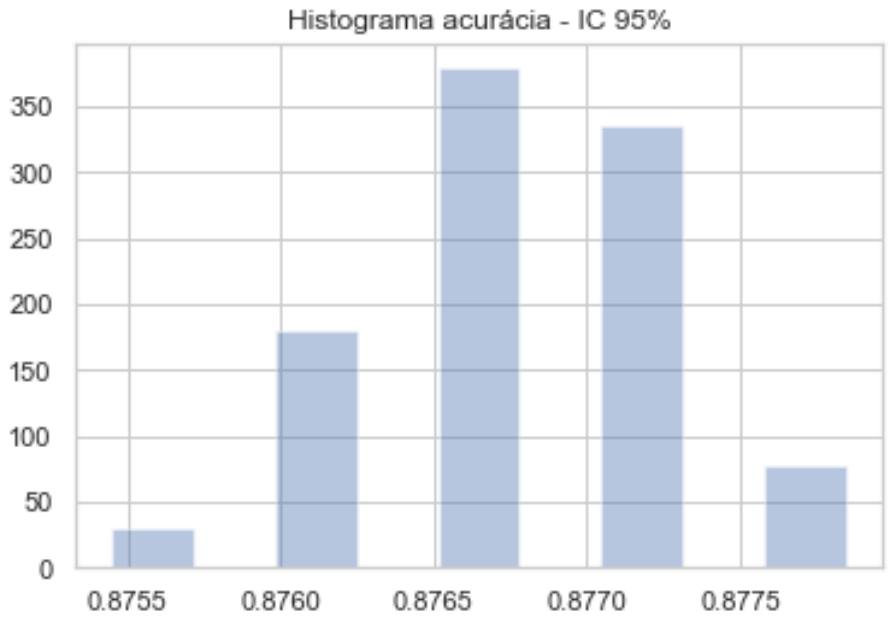


Figura 4.2: Histograma da acurácia para um intervalo de confiança de 0.95 com 1000 iterações.

4.3 Implantação

Selecionou-se o modelo com melhor resultado, GBM com a técnica ROS para balanceamento em treinamento, descrito na seção 4.1, nesta fase do estudo, a critério de uma pré-implantação, resultou no grau de risco para cada cidade.

O grau de risco para cada cidade foi determinado, conforme ilustrado na Figura 4.3. Para isso, utilizou-se do poder preditivo do modelo para obter a probabilidade de cada classe, no caso, a classe de interesse com valor 1.

Com a probabilidade de ocorrência da classe, o próximo passo foi discretizar os valores em uma escala de escravidão, em 4 níveis de risco (0 - baixo, 1- médio, 2- alto e 3 - altíssimo). Uma observação importante, é a de que mesmo para o nível mais baixo, indica um risco existente associado e não a ausência do risco.

Na Figura 4.3, o eixo x corresponde aos Estados. No alinhamento vertical, cada ponto no gráfico representa uma cidade para um respectivo Estado (os nomes foram anonimizado de 1 a 27). Por outro lado, o eixo y contém as pontuações de risco para cada cidade. A discretização por níveis de escravidão revela o agrupamento de cidades em diferentes níveis.

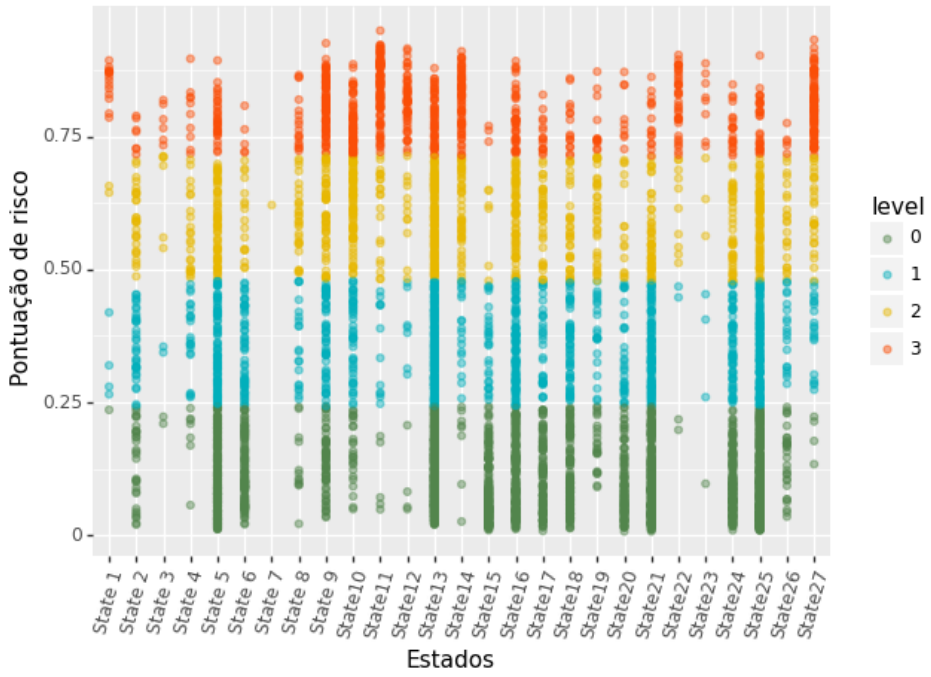


Figura 4.3: Aplicação do modelo. Pontuação do risco de escravidão para cada cidade, classificadas em quatro níveis.

A implantação de modelos de aprendizagem de máquina em escala de produção envolve o planejamento da estrutura que irá sustentar a solução proposta. Além de alinhar a estratégia do intervalo de tempo em que o modelo será atualizado, ou seja, o tempo necessário para um novo treinamento do modelo proposto.

Como resultado desta pesquisa, foi possível implantar um painel dinâmico, que pode ser consultado pelo grupo de trabalho responsável pelas investigações do trabalho escravo no Brasil. A Figura 4.4 apresenta o aspecto visual do painel.

Figura 4.4: Painel dinâmico desenvolvido com o modelo da pesquisa.

Com a solução proposta na pesquisa, foi possível identificar as variáveis de interesse que representam o modelo resultante. A partir das variáveis identificadas é possível iniciar o planejamento para uma escala produtiva. Este estudo utiliza os dados do PNUD, que por sua vez, utiliza os indicadores do último Censo para sua formulação, conforme descrito na seção 3.2. Uma vez conhecendo as variáveis utilizadas no modelo, a coleta dos dados poderá ser antecipada e ser realizada diretamente com as prefeituras.

4.3.1 Levantamento comparativo do modelo proposto

Este estudo treinou o modelo com um conjunto de dados histórico até 2018, o que possibilitou a constatação do modelo em uma situação da vida real, pois novos dados de fiscalização podem ser utilizados para comparar com o modelo.

Os resultados gerados com a previsão do modelo, foram confrontados com os novos dados de fiscalização da Secretaria do Trabalho de 2019 até junho 2020, disponíveis no portal radar[1].

Essa subseção apresenta o levantamento comparativo da consulta dos registros de trabalho escravo, os resultados são apresentados na Tabela 4.6 com a quantidade de cidades e os respectivos níveis de risco associados.

A Tabela 4.6 contem em formato sumarizado as classificações obtidas com o modelo proposto para as cidades com registro de trabalho escravo. A taxa de erro é determinada pela indicação de risco baixo dado pelo modelo. Os níveis de riscos diferentes deste são

[1]Radar: `https://sit.trabalho.gov.br/radar/`, acessado em 15/10/2020.

considerados dentro da margem de acerto, ou seja, para a classificação de risco de nível médio, alto e altíssimo é considerado uma previsão de acerto do modelo.

Tabela 4.6: Resumo da classificação de risco obtida utilizando o modelo proposto para as cidades de trabalho escravo confirmado de 2019 a junho de 2020

Quantidade de cidades por nível de risco de 2019 a junho de 2020	
Nível do risco	Quantidade cidades
Altíssimo	44
Alto	29
Médio	11
Baixo	12

Para a consulta realizada, o total de 96 cidades sofreram registro de ocorrências de trabalho escravo comprovado. O modelo proposto nesta pesquisa conseguiu prever a classificação do risco corretamente em 84 cidades, apenas 12 cidades o modelo sugeriu baixo risco associado, resultando em 12,5 percentuais em taxa de erro e 87,5% em acertos.

Capítulo 5

Conclusão

Esta pesquisa abordou o uso do aprendizado de máquina para prever o risco do trabalho escravo contemporâneo em cidades brasileiras. O estudo identificou o grau do risco de escravidão para cada cidade do país, utilizando o poder de previsão do modelo. O que foi possível por meio da probabilidade da predição modelada.

A solução proposta contou com etapas diversas em um contexto de aprendizagem de máquina. Quanto à seleção de variáveis foi utilizada a técnica *embedded* com aprendizagem e regularização L1, o que ocasionou a redução de dimensionalidade para 16 variáveis.

Por tratar-se de um problema de conjunto de dados desbalanceado e na ausência de trabalhos em contextos semelhantes, foi realizado o estudo comparativo (*benchmarking*) entre as principais técnicas de balanceamento de dados existentes na literatura. O uso da reamostragem contribuiu para viabilização do teste estatístico o qual foi utilizado *Bootstrapping* com 1000 iterações. Ao final, foi feito o levantamento comparativo colocando em prática o modelo proposto na pesquisa com as ocorrências já registradas de trabalho escravo.

Os objetivos estabelecidos foram alcançados:

i. As principais variáveis preditoras foram identificadas;

ii. O *benchmarking* foi construído para selecionar o método mais eficiente de balanceamento de dados;

iii. Os modelos de aprendizagem foram validados para o problema proposto;

iv. Um teste estatístico com reamostragem foi aplicado para avaliação do modelo;

v. Foi identificada a pontuação e o nível de risco de cada cidade;

vi. O modelo foi implantado em um painel dinâmico, possibilitando consultas.

Com a identificação das variáveis o estudo contribuiu para o entendimento do problema abordado, auxiliando para que estudos futuros possam ser direcionados. A relação das variáveis selecionadas com o modelo proposto, confirmou a importância de existir dados sobre índices de desenvolvimento humano.

O *benchmarking* foi uma ferramenta essencial que serviu como apoio na escolha técnica do melhor método de balanceamento de dados para o escopo proposto. O método apontado com melhores resultados no estudo comparativo foi o ROS.

A capacidade de generalização do modelo foi observada com o uso da técnica de validação cruzada na etapa de modelagem. Seguindo as recomendações bibliográficas, todos os modelos apresentados no estudo foram validados em uma base desconhecida para o modelo, o que contribuiu para mitigar qualquer possibilidade de *overfitting* ou viés.

As pontuações (*score*) de risco podem ser utilizadas para o planejamento de ações estratégicas preventivas, que atualmente não possuem abordagens preditivas. O grau do risco associado também pode ser usado como entrada no sistema de denúncias da instituição para classificar e ponderar as denúncias que serão atendidas com prioridade.

Foi possível constatar com os modelos selecionados que as previsões têm um desempenho aceitável para o estudo proposto, que é a identificação do risco da escravidão contemporânea para cada cidade brasileira. O modelo de melhor desempenho foi o GBM, treinado no conjunto de dados ROS, com 77% de acurácia, 80% de AUC e G-mean com 71%. As taxas de falsos positivos são mais evidentes no modelo de teste, pois este possui uma taxa baixa de precisão e o conjunto de dados de teste é desbalanceado, resultando em um F1 com 41%.

O teste estatístico com *Bootstrapping* apresentou bons resultados, com um intervalo de confiança de 0.95, a acurácia ficou entre 87.5% e 87.8%. O levantamento comparativo do modelo proposto com a base de dados de fiscalização para os anos de 2019 a junho de 2020 da Secretaria do Trabalho, revelou 87,5% de acerto e 12,5% de taxa percentual de erro.

Os resultados do levantamento comparativo demonstraram ser coerentes com o teste estatístico, pois a taxa de acerto ficou dentro da margem prevista do teste de robustez para 1000 iterações. O algoritmo de predição do grau do risco classificou corretamente 84 ocorrências das 96 confirmadas na consulta do levantamento, sendo que para 44 cidades atribuiu o nível de grau de risco em altíssimo, 29 cidades com nível de grau de risco em alto e 11 em nível médio. O resultado do levantamento comparativo confirma a robustez do modelo, assim como foi indicado pelo teste estatístico.

A solução proposta forneceu o grau do risco do trabalho escravo para cada cidade brasileira, estes resultados podem ser aplicados, mas não exclusivamente, nas seguintes aplicações:

- Como ponderador no sistema de denúncias do MPT;
- Indicador em painéis estratégicos e relatórios;
- Métrica de medição para a erradicação do trabalho escravo;
- Indicador no sistema de denúncias da Secretaria do Trabalho;
- Monitor do grau de risco e nível do trabalho escravo para consulta do Legislativo, Judiciário e Executivo;
- Aplicações que requeiram um indicador que contenha o grau de risco associado ao trabalho escravo para cada cidade do Brasil.

As principais contribuições deste trabalho são resumidas em três grupos:

- Tecnológica, cujo o modelo proposto neste estudo gerou um conjunto de dados rotulado contendo o grau de risco e níveis associados, o que potencializa o uso em aplicações destinadas a temática do trabalho escravo contemporâneo;
- Inovadora, pois o método proposto abordou de forma inédita um problema real. O trabalho prova como verdadeira a hipótese da pesquisa, de que é possível identificar variáveis preditoras que expliquem as ocorrências do trabalho escravo, e ainda provou que é possível prever o risco associado para cada cidade.
- Científica, novas pesquisas podem utilizar este estudo [59] como fonte motivadora ou referencial. O método abordado nesta pesquisa pode ser estendido para outros temas, ou ainda ser utilizada em outros países, pois o problema é de alcance global.

Toda pesquisa deve ser vista como um projeto, isso envolve cumprir prazos e seguir um cronograma determinado, visando alcançar os objetivos estabelecidos. Assim como em projetos, é natural que toda pesquisa desenvolvida gere alguma limitação. Os fatores limitantes identificados nesta pesquisa são:

- Os modelos não diferenciam as condições urbanas e rurais, agrupando-as em uma única classe;
- Este estudo não ponderou regiões ou cidades com maiores índices de trabalhadores encontrados em situações de escravo;
- O estudo limitou-se a dados socioeconômicos e demográficos.

A continuidade deste pesquisa é fortemente indicada, pois devido a problemática do tema, deve ser considerado uma discussão interdisciplinar e aprofundada, que envolva a análise de aspecto social e econômico das variáveis preditoras selecionadas no estudo.

Como trabalhos futuros, recomenda-se:

- Novas abordagens usando diferentes algoritmos de classificação podem ser explorados;
- Um novo conjunto de dados sobre o censo populacional pode ser experimentado;
- O uso de *gridsearch* pode ser aplicado a modelos paramétricos, visando melhorar os resultados;
- Uma abordagem multiclasse pode ser aplicada tratando atividade escrava não identificada, atividade urbana e rural.

Referências

[1] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer e Rüdiger Wirth: *CRISP-DM 1.0 Step-by-step*. Relatório Técnico, 2000, ISBN 9780769532677. x, 4, 8, 9, 10

[2] Scarpa, Silvia: *The Nebulous Definition of Slavery : Legal versus Sociological Definitions of Slavery*. The Palgrave International Handbook of Human Trafficking, páginas 1–14, 2019. 1

[3] United Nations Office on Drugs and Crime: *United Nations convention against transnational organized crime and protocols thereto*. Trends in Organized Crime, 5(4):11–21, 2007, ISSN 1084-4791. 1

[4] International Labor Organization: *Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. Relatório Técnico, 2017, ISBN 9789221301318. 1, 2, 4, 20

[5] Walk Free Foundation: *The Global Slavery Index*. Relatório Técnico, 2018. 1, 2, 4, 19, 20

[6] Members of the International Labour Organization: *Forced Labour Convention - C029*. Geneva, 1930. The General Conference of the International Labour Organization. 1, 2, 4

[7] David, Fiona: *Modern Slavery - From Statistics to Prevention*. Chance, 30(3):54–60, 2017, ISSN 0933-2480. 1

[8] Datta, M. Narayan, Fiona David, Kevin Bales e Nick Grono: *The Global Slavery Index*. Relatório Técnico, Walk Free Foundation, 2013. 2, 4

[9] International Labour Organization: *R203 - Forced Labour (Supplementary Measures) Recommendation*, 2014. 2, 4

[10] MPT: *30 anos de constituição federal: Atuação do MPT*. MPT, Brasília, 2018, ISBN 978-85-66507-25-6. 2, 3

[11] Mello, Neli Aparecida De, Julio Hato e Eduardo Paulon Girardi: *Atlas do trabalho escravo no Brasil*. ISBN 9788586928093. 2

[12] ILO: *Perfil dos principais atores envolvidos no trabalho escravo rural no Brasil*. 2011, ISBN 9789228254938. 3

[13] United Nations: *Transforming our world: The 2030 agenda for sustainable development*. United Nations Department of Economic and Social Affairs, 2015. 4

[14] Liu, Chuanhai: *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*. Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family, páginas 227–238, 2005. 6

[15] Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau e Yu Sung Su: *A WEAKLY INFORMATIVE DEFAULT PRIOR DISTRIBUTION FOR LOGISTIC AND OTHER REGRESSION MODELS*. IMS Journals and Publications - Annals of Applied Statistics, 2(4):1360–1383, 2008, ISSN 1932-6157. 6

[16] International Labour Office e Walk Free Foundation: *Methodology of the global estimates of modern slavery: Forced labour and forced marriage*. Relatório Técnico, International Labour Office, Geneva, 2017, ISBN 9789221301318. `https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_575479.pdf`. 7, 19, 20

[17] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 17(3):37–54, 1996, ISSN 16113349. 8, 9, 10

[18] M. Weiss, Sholom e Nitin Indurkhya: *Predictive Data Mining: A pratical guide*. 1997, ISBN 978-1558604032. 8

[19] Chen, Ming syan, Jiawei Han e Philip S Yu: *Data Mining: An Overview from a Database Perspective*. Ieee Transactions on Knowledge and Data Engineering, 8(6):866–883, 1996. 8, 9

[20] Samuel, A L: *Some Studies in Machine Learning Using the Game of Checkers*. IBM JOURNAL, páginas 211–229, 1959. 9

[21] Jordan, M. I. e T. M. Mitchell: *Machine learning: Trends, perspectives, and prospects*. Science, 349(6245):255–260, 2015, ISSN 10959203. 9

[22] Mohri, Mehryar, Afshin Rostamizadeh e Ameet Talwalkar: *Foundations of Machine Learning*, volume 53. The MIT Press, Cambridge, 2012, ISBN 978-0-262-01825-8. 9

[23] McCarthy, John, Marvin L. Minsky e Claude E. Shannon: *A proposal for the Dartmouth summer research project on artificial intelligence - August 31, 1955*. Ai Magazine, 27(4):12–14, 2006, ISSN 2371-9621. 9

[24] Hendler, James: *Avoiding Another AI Winter*. IEEE Intelligent Systems, 23(2):2–4, 2008. 9

[25] Holzinger, Andreas, Peter Kieseberg, Edgar Weippl e A Min Tjoa: *Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI*. 11015:1–8, 2018. 9

[26] Goodfellow, Ian, Yoshua Bengio e Aaron Courville: *Deep Learning*. MIT press, Cambridge, 2016, ISBN 9780262035613. 9

[27] Azevendo, Ana e M.F Santos: *KDD , SEMMA AND CRISP-DM : A parallel overview*. IADS-DM, 2008. 9

[28] Chandrashekar, Girish e Ferat Sahin: *A survey on feature selection methods*. Computers and Electrical Engineering, 40(1):16–28, 2014, ISSN 00457906. 11

[29] Cai, Jie, Jiawei Luo, Shulin Wang e Sheng Yang: *Feature selection in machine learning: A new perspective*. Neurocomputing, 300:70–79, 2018, ISSN 18728286. 11, 12

[30] Xue, Bing, Mengjie Zhang, Will N. Browne e Xin Yao: *A Survey on Evolutionary Computation Approaches to Feature Selection*. IEEE Transactions on Evolutionary Computation, 20(4):606–626, 2016, ISSN 1089778X. 11

[31] Escanilla, Nicholas Sean, Lisa Hellerstein, Ross Kleiman, Zhaobin Kuang, James Shull e David Page: *Recursive Feature Elimination by Sensitivity Testing*. Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, páginas 40–47, 2019. 11, 12

[32] Ang, Jun Chin, Andri Mirzal, Habibollah Haron e Haza Nuzly Abdull Hamed: *Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(5):971–989, 2016, ISSN 15455963. 11

[33] Li, Jundong, Kewei Cheng, Wang Suhang, Fred Morstatter, Robert P. Trevino, Jiliang Tang e Huan Liu: *Feature Selection: A Data Perspective*. ACM Computing Surveys, 50(6), 2017. 11, 12

[34] Tibshirani, Robert: *Regression Shrinkage and Selection Via the Lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1995. 12

[35] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning The Elements of Statistical Learning*. 2017. 12, 13, 16, 17, 18, 19

[36] Kennard, Robert W e Arthur E Hoerl: *Ridge regression: Biased estimation for nonorthogonal problems*. Technometrics, 12(1):55–67, 1970. 13

[37] Hosmer, David W. e Stanley Lemeshow: *Applied Logistic Regression*. Wiley-Interscience Publication, 2ª edição, 2000, ISBN 0471356328. 13

[38] Friedman, Jerome H.: *Greedy Function Approximation: A Gradient Boosting Machine*. Annals of statistics, páginas 1189–1232, 2001. 13

[39] Crowley, Philip H: *Resampling Methods for Computation - Intensive Data Analysis in Ecology and Evolution*. Relatório Técnico, 1992. `www.annualreviews.org/aronline`. 14

[40] B. Efron: *Bootstrap Methods: Another Look at the Jackknife*. Annals of Statistics, Volume 7(Number 1):1–26, 1979. `https://projecteuclid.org/euclid.aos/1176344552`. 14

[41] Phillip I. Good: *Resampling Methods*. Huntington Beach, 2006, ISBN 978-0-8176-4386-7, 978-0-8176-4444-4. https://doi.org/10.1007/0-8176-4444-X. 14

[42] Mooney, Christopher Z e Robert D Duval: *Bootstrapping: A nonparametric approach to statistical inference*. Número 95. sage, 1993. 14

[43] Liu, Xu ying, Jianxin Wu e Zhi hua Zhou: *Exploratory Undersampling for Class-Imbalance Learning*. Ieee Transactions on Systems, Man and Cybernetics, 39(2):1–14, 2009. 15, 18

[44] Zong, Weiwei, Guang Bin Huang e Yiqiang Chen: *Weighted extreme learning machine for imbalanced learning*. Neurocomputing, 101:229–242, 2013. 15, 18

[45] Wang, Shuo, Leandro L Minku e Xin Yao: *A Systematic Study of Online Class Imbalance Learning With Concept Drift*. IEEE Transactions on Neural Networks and Learning Systems, páginas 1–20, 2018. 15

[46] Zhu, Bing, Bart Baesens, Aimée Backiel e Seppe K.L.M. Vanden Broucke: *Benchmarking sampling techniques for imbalance learning in churn prediction*. Journal of the Operational Research Society, 69(1):49–65, 2018, ISSN 14769360. 15

[47] Chawla, N.V., Bowyer K.W. Hall L.O. Kegelmeyer W.P.: *SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research*. Journal of Artificial Intelligence Research, 16:321–357, 2002, ISSN 10769757. 15

[48] He, Haibo, Yang Bai, Edwardo Garcia A. e Shutao Li: *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning*. IEEE International Joint Conference on Neural Networks, IJCNN '08, (3):1322–1328, 2008. 15

[49] Yen, Show Jane e Yue Shi Lee: *Cluster-based under-sampling approaches for imbalanced data distributions*. Expert Systems with Applications, 36(3 PART 1):5718–5727, 2009, ISSN 09574174. 15

[50] Tomek, Ivan: *Tomek Link: Two Modifications of CNN*. IEEE Trans. Systems, Man and Cybernetics, páginas 769–772, 1976. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4309452. 15

[51] Breiman, Leo: *Random forests*. Ensemble Machine Learning: Methods and Applications, 45:5–32, 2001. 15

[52] Kohavi, R: *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Em *Proceedings of the 14th international joint conference on Artificial intelligence - IJCAI*, volume 2, páginas 1137–1143, 1995. 16, 17

[53] Swets, John A: *Measuring the Accuracy of Diagnostic Systems*. Science, 240(4857):1285–1293, 1988. 19

[54] Chawla, Nitesh V: *Data Mining for Imbalanced Datasets: An Overview*. Data Mining and Knowledge Discovery Handbook 2nd ed., 2010. 19

[55] P. Bradley, Andrew: *The use of the area under the roc curve in the evaluation of machine learning algorithms*. Pattern Recognition Society, 30(7):1145–1159, 1997. 19

[56] Diego-Rosell, Pablo e Jacqueline Joudo Larsen: *Modelling the Risk of Modern Slavery*. SSRN Electronic Journal, 2018. 19, 20

[57] Datta, Monti N. e Kevin Bales: *Slavery in Europe: Part 1, Estimating the Dark Figure*. Human Rights Quarterly, 35(4):817–829, 2013. 20

[58] Datta, Monti N. e Kevin Bales: *Slavery in Europe: Part 2, Testing a Predictive Model*. Human Rights Quarterly, 36(2):277–295, 2014. 20

[59] Da Silva Santos, M., M. Ladeira, G.C.G. Van Erven e G. Luiz Da Silva: *Machine learning models to identify the risk of modern slavery in Brazilian Cities*. Em *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 2019, ISBN 9781728145495. 21, 46

Apêndice A

Artigo publicado - ICMLA 2019

Machine Learning Models to Identify the Risk of Modern Slavery in Brazilian Cities

Abstract: The scope of modern slavery encompasses human trafficking, forced labor, debt bondage and child labor. This article proposes the use of predictive models to identify the risk of modern slavery in Brazilian cities using real socioeconomic, demographic and rescue operations data. The study uses the embedded technique with Lasso regularization (L1) to select variables. A comparative analyze of techniques for treatment of imbalanced data was applied and the results indicated the Random Over-Sampling (ROS) as the best one. In total, 16 models are evaluated, consisting of 8 different data sets and two classifiers: Logistic Regression (LR) and Gradient Boosting Machine (GBM). The results indicate that the GBM model has better performance and efficiency, with accuracy of 77%, AUC 80% and G-mean of 71%.

Published in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)

Date of Conference: 16-19 Dec. 2019.

Date Added to IEEE Xplore: 17 February 2020.

Electronic ISBN: 978-1-7281-4550-1

Print on Demand(PoD) ISBN: 978-1-7281-4551-8

DOI: 10.1109/ICMLA.2019.00132

Publisher: IEEE

Conference Location: Boca Raton, FL, USA, USA

Link: https://ieeexplore.ieee.org/document/8999093