Investigação da Influência de Atributos que Caracterizam Municipalidades em Resultados Eleitorais

Nilton Cesar Sacco

Agosto / 2021

Dissertação de Mestrado em Ciência da Computação

Investigação da Influência de Atributos que Caracterizam Municipalidades em Resultados Eleitorais

Esse documento corresponde a dissertação apresentado à Banca Examinadora no curso de Mestrado em Ciência da Computação do UNIFACCAMP – Centro Universitário Campo Limpo Paulista.

Campo Limpo Paulista, 27 de agosto de 2021.

Nilton Cesar Sacco

Maria do Carmo Nicoletti (Orientadora)
Eduardo Javier Huerta Yero (Coorientador)

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ficha catalográfica elaborada pela

Biblioteca Central da Unifaccamp

S125i

Sacco, Nilton Cesar

Investigação da influência de atributos que caracterizam municipalidades em resultados eleitorais / Nilton Cesar Sacco. Campo Limpo Paulista, SP: Unifaccamp, 2021.

Orientadora: Profª. Drª. Maria do Carmo Nicoletti
Coorientador: Profº Dr. Eduardo Javier Huerta Yero

Dissertação (Programa de Mestrado Profissional em Ciência da Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Extração de dados. 2. Análise de dados. 3. Dados eleitorais. 4. IDH. 5. IDHM. I. Nicoletti, Maria do Carmo. II. Yero, Eduardo Javier Huerta. III. Centro Universitário Campo Limpo Paulista. IV. Título.

CDD-005.1

Resumo. *O trabalho de pesquisa descrito nesta dissertação teve como enfoque principal a investigação de possíveis relações entre determinados indicadores que caracterizam aspectos de eleitores e municípios brasileiros, com o resultado obtido em uma eleição em nível federal. Um dos indicadores escolhido foi o Índice de Desenvolvimento Humano Municipal (IDHM), que busca caracterizar numericamente o nível de desenvolvimento e a qualidade de vida oferecida por municípios brasileiros. O IDHM é derivado do Índice de Desenvolvimento Humano (IDH), um indicador global considerado uma das métricas mais amplamente aceitas para representar o status de desenvolvimento de um país. O trabalho realizado analisou os resultados das eleições federais à presidência, ocorridas em 2018, por meio de exames dos índices IDHM, bem como do número de habitantes de cada município. Apresenta e discute os resultados de experimentos realizados em conjunto de dados do TSE e do AtlasBrasil, subsidiado por análises estatísticas e em uma segunda etapa pelo uso de algoritmos de aprendizado supervisionado e não supervisionado usando os mesmos dados que foram submetidos à análise estatística inicial. Com base nos resultados obtidos, foi possível constatar que os municípios com maiores índices de IDHM e maior número de habitantes preferiram majoritariamente o PSL, enquanto o PT foi preferido por aqueles municípios que apresentam menores índices de IDHM e menor número de habitantes.*

Palavras-chave: *extração de dados, análise de dados, dados eleitorais, IDH, IDHM.*

Abstract*: The research work described in this dissertation had as its main focus the investigation of possible relationships between certain indicators that characterize aspects of Brazilian voters and municipalities, with the result obtained in an election at the federal level. One of the indicators chosen was the Municipal Human Development Index (IDHM), which seeks to numerically characterize the level of development and quality of life offered by Brazilian municipalities. The HDI is derived from the Human Development Index (HDI), a global indicator considered to be one of the most widely accepted metrics to represent the development status of a country. The work carried out analyzed the results of the federal presidential elections, held in 2018, through examinations of the IDHM indexes, as well as the number of inhabitants in each municipality. It presents and discusses the results of experiments carried out on datasets from the TSE and AtlasBrasil, supported by statistical analysis and in a second stage by the use of supervised and unsupervised learning algorithms using the same data that were submitted to the initial statistical analysis. Based on the results obtained, it was possible to verify that municipalities with higher MHDI indexes and greater number of inhabitants preferred the PSL, while PT was preferred by those municipalities with lower MHDI indexes and fewer inhabitants.*

Keywords: *data extraction, data analysis, electoral data, HDI, MHDI*

Agradecimentos

Agradeço primeiramente a Deus pelo dom da vida, pela sabedoria e por ter me dado forças para chegar até aqui.

Agradeço à minha esposa que me incentivou e me manteve motivado durante toda essa jornada e a meus pais e familiares que me apoiaram sempre.

Agradeço à Profa. Dra. Maria do Carmo Nicoletti pela dedicação, paciência, motivação, ensinamentos e por todo conhecimento compartilhado durante o processo de desenvolvimento desta dissertação.

Agradeço ao Prof. Dr. Eduardo Javier Huerta Yero pela confiança, incentivo, amizade e pelo conhecimento compartilhado.

Às professoras Dra. Ana Maria Monteiro, Dra. Verônica Oliveira de Carvalho pela disposição e boa vontade em participar da banca examinadora.

Aos professores e funcionários do programa do Centro Universitário Campo Limpo Paulista.

Sumário

Glossário

AM – Aprendizado de Máquina.

CSV – Valores Separados por Vírgula, um tipo de arquivo que pode ser editado no Excel

IA – Inteligência Artificial.

IDH – Índice de Desenvolvimento Humano

IDHM – Índice de Desenvolvimento Humano Municipal

FJP - Fundação João Pinheiro

IBGE – Instituto Brasileiro de Geografia e Estatística

IPEA - Instituto de Pesquisa Econômica Aplicada

LGBT - Lésbicas, Gays, Bissexuais e Transgênero

PIB – Produto Interno Bruto

PNUD - Programa das Nações Unidas para o Desenvolvimento

PPC – Poder de Paridade de Compra

PSL – Partido Social Liberal

PT – Partido dos Trabalhadores

RNB - Renda Nacional Bruta

RPC – Renda Per Capta

TSE – Tribunal Superior Eleitoral

tweets - mensagem publicada no Twitter

Lista de Tabelas

Lista de Figuras

Lista de Algoritmos

Capítulo 1

Introdução

Este documento descreve a pesquisa em nível de mestrado, realizada junto ao PMCC da UNIFACCAMP, intitulado “Investigação da Influência de Atributos que Caracterizam Municipalidades em Resultados Eleitorais”, sob a orientação da Profa. Maria do Carmo Nicoletti e coorientação do Prof. Eduardo Javier Huerta Yero.

1.1 Contextualização

O conceito de dado praticamente permeia todo o conhecimento humano. De uma maneira simplista, dado pode ser caracterizado como observações ou, então, resultados de medições coletadas via instrumentação ou ainda, resultados de simples contagens/identificações.

Quando dados são organizados, interpretados ou estruturados de maneira a terem um significado ou então, alguma utilidade, são considerados informação. Já o conceito de conhecimento é bem mais abrangente e, apesar de ser intrinsicamente dependente dos conceitos de dado e informação, vai bem além, envolvendo consciência e cognição, como definido em [Wikipedia 2020], ao considerar conhecimento como uma familiaridade, consciência ou entendimento de alguém ou algo, como fatos, informações, descrições ou habilidades, adquiridos por meio da experiência, ou da educação, pela percepção, descoberta ou aprendizado.

Invariavelmente os três conceitos, dado, informação e conhecimento são abordados de maneira inter-relacionada, devido à interdependência entre eles. Uma abordagem simplificada que é comumente aceita é apresentada em [Tuomi 1999], em que dados são tratados como fatos simples que se tornam informações, à medida que são combinados em estruturas significativas e que, subsequentemente, se tornam conhecimento à medida que informações significativas são colocadas em um contexto e quando podem ser usadas para fazer previsões.

Na literatura existem inúmeros trabalhos abordando os três conceitos, nos mais variados contextos do conhecimento humano (ver [Tuomi 1999] [Tan *et al*. 2006] [Zins

2007] [Diffen 2020]). Em vários trabalhos com enfoque em uso/análise de dados, dados podem também ser referenciados como dados brutos *i.e.*, dados obtidos que não passaram por um processo de tratamento para eliminar ou, possivelmente, recuperar dados com problemas por exemplo, incompletos ou espúrios.

É fato que um volume imenso de dados brutos vem sendo produzidos diariamente, por fontes dos mais variados tipos, tais como redes sociais, órgãos governamentais, empresas estatais, empresas privadas, empresas sem fins lucrativos, bancos, universidades, meios de comunicação, sites informativos e de divulgação, e muitos outros. Muitos desses dados acabam se tornando obsoletos ao longo do tempo devido, principalmente, à sua natureza volátil.

Muitas organizações, que ainda não estão sensibilizadas com relação à importância que os dados que produzem têm no planejamento de suas atividades, acabam ignorando muitos dos dados produzidos. Eventualmente tais dados se tornam desatualizados ao longo do tempo e, então, são descartados. Já organizações mais ágeis e sensíveis à importância que dados desempenham em quase todas as atividades que implementam/planejam, buscam desenvolver procedimentos que utilizam muitos dos dados produzidos para subsidiar a definição/refinamento de muitos de seus processos e, particularmente, os processos administrativos relacionados a tomadas de decisão.

Dados produzidos ou coletados servem, por exemplo, para que organizações, por meio da análise de tais dados, conheçam a opinião de seus clientes sobre seus produtos/serviços ou, então, para poderem corrigir problemas relacionados aos seus produtos/serviços.

Outro exemplo da coleta e uso de tais dados é aquele realizado em hospitais quando do monitoramento de pacientes, por meio de sensores colocados em pacientes que estão sob cuidados médicos contínuos, com o objetivo de coletar informações sobre suas principais funções vitais, para acompanhamento *online* de suas situações clínicas, como descrito em [Wesley *et al*. 2003] [An *et al*. 2012]. Também, imagens digitalizadas podem contribuir tanto para a equipe médica substanciar com mais precisão diagnósticos, quanto para acompanhar a resposta/evolução do estado do paciente aos tratamentos.

Não pode ser esquecido, entretanto, que existem ainda muitas áreas em que o uso de dados passíveis de serem coletados ainda não foram devidamente explorados, devido às

dificuldades práticas envolvidas com a própria área. É o caso, por exemplo, de dados relativos a áreas relacionadas a atividades ilegais, que permeiam muitas das atividades legais em comunidades sociais [Kejriwal & Szekely 2007].

Atualmente o Brasil é uma fonte geradora de grande volume de dados disponíveis em repositórios públicos, estruturados, que livremente podem ser acessados, utilizados, modificados e compartilhados. Algumas leis, como a Lei do Acesso à Informação [Lei do Acesso 2020] e a Lei da Transparência [Portal da Transparência 2020], que garantem o acesso à informação e a veracidade dos dados disponibilizados, incentivaram os órgãos públicos brasileiros a disponibilizar em seus portais os detalhes dos processos e movimentações do dinheiro público, como também, dados que indicam a qualidade de vida nos municípios brasileiros e uma maior clareza nos processos e resultados eleitorais. Podemos citar como exemplo o Programa das Nações Unidas para o Desenvolvimento (PNUD), que tem disponibilizado dados importantes sobre o desenvolvimento humano nos países e colaborado com índices que caracterizam a qualidade de vida, as condições da saúde e da educação da população em todos os municípios brasileiros, como também o Tribunal Superior Eleitoral (TSE) que tem contribuído com dados dos processos eleitorais nas instâncias municipal, estadual e federal.

Em 2018 a polarização do cenário político brasileiro atingiu seu auge durante as eleições gerais. Os candidatos à presidência chegaram ao segundo turno representando visões político/governamentais radicalmente opostas para o futuro do Brasil. De um lado o Partido dos Trabalhadores (PT), um partido de esquerda, que havia permanecido no poder por quatro mandatos consecutivos, tentava retomar a presidência depois que seu último presidente foi deposto por um polêmico processo de *impeachment* e muitas de suas figuras políticas mais importantes se envolveram em um enorme escândalo de corrupção. Do outro lado o candidato do Partido Social Liberal (PSL), um partido de extrema direita com base no liberalismo econômico, militarismo, anticomunismo e conservadorismo social, que entre suas posições mais radicais e polêmicas se destacavam a admiração por torturadores do período do regime militar brasileiro, seus comentários discriminatórios dirigidos à comunidade LGBT e sua visão sobre a necessidade de endurecimento das políticas na luta contra o crime. A eleição foi vencida pelo candidato do PSL com 55,13% dos votos válidos [TSE 2018].

Todo esse contexto político referente ao segundo turno das eleições de 2018, se tornou uma boa oportunidade para estudos sobre a escolha dos eleitores em uma disputa tão bipolarizada. Conhecer o vencedor das eleições em cada município, poderia incutir na descoberta de padrões que relacionem o resultado das votações ao desenvolvimento de cada município e ao tamanho de sua população.

Durante o levantamento bibliográfico para o desenvolvimento deste trabalho, não foram encontrados trabalhos publicados sobre a previsão dos resultados das eleições presidenciais no Brasil, principalmente com relação ao uso da situação de desenvolvimento de cada município como um potencial preditor de resultados eleitorais, podendo afirmar que a metodologia proposta difere das demais metodologias encontradas na literatura.

1.2 Objetivo da Pesquisa e Organização da Dissertação

O objetivo da pesquisa descrita neste documento é o de investigar possíveis correlações e tendências, entre determinados indicadores, que caracterizam aspectos de eleitores e municípios brasileiros, com os resultados obtidos na última eleição presidencial, que aconteceu no Brasil em 2018.

A investigação foi planejada para ser conduzida por meio de tratamento estatístico de dados e, em algumas situações, algoritmos de AM foram utilizados para a indução de modelos e identificação de atributos relevantes.

Não há nenhuma intenção de avançar o conhecimento de AM propondo novos algoritmos ou variações de algoritmos já existentes. O trabalho desenvolvido com vista ao objetivo pretendido está descrito nesta dissertação, que está organizado em mais cinco capítulos, cujos respectivos conteúdos são descritos, de forma resumida, a seguir.

O Capítulo 2 reporta, de maneira resumida, o levantamento bibliográfico de pesquisas realizadas e disponibilizadas em vários veículos, acadêmicos e/ou de divulgação, que trouxeram alguma contribuição relevante sobre a área de conhecimento que é foco da dissertação, *i.e.*, análise de dados e extração de conhecimento a partir de dados eleitorais. A principal contribuição do trabalho foi a proposta do uso do Índice de Desenvolvimento Humano Municipal, baseado no Índice de Desenvolvimento Humano, utilizado pela ONU.

O Capítulo 3 contextualiza a área de Aprendizado de Máquina (AM) como uma subárea da área de Inteligência Artificial (IA), apresenta os conceitos iniciais básicos associados à AM e introduz uma notação formal que é empregada no capítulo para a caracterização de dois grupos relevantes de algoritmos de AM, os supervisionados e os não-supervisionados.

O Capítulo 4 evidencia as fontes dos dados utilizados na pesquisa, informando suas origens e comentando sobre as características de veracidade e confiabilidade associadas aos dados delas obtidos. No capítulo são brevemente apresentados os dados governamentais disponibilizados publicamente, que estão sendo usados na realização de atividades relacionadas ao trabalho de pesquisa. De interesse majoritário ao trabalho de pesquisa são os dados relacionados aos resultados da eleição presidencial de 2018, os dados disponibilizados pelo Censo Demográfico de 2010 e os dados sobre os índices de desenvolvimento humano por município. Fórmulas para a análise estatística dos dados e para o cálculo de descritores utilizados no cálculo de índices são também apresentadas.

O Capítulo 5 descreve todo o processo de preparação dos dados, com destaque às etapas utilizadas, desde a importação dos dados até a geração de gráficos. O capítulo também apresenta uma breve introdução ao software Exploratory [Exploratory 2020], com foco em suas principais funcionalidades disponibilizadas e nas que foram utilizadas na fase inicial de preparação dos dados.

O Capítulo 6 apresenta uma descrição da metodologia utilizada na condução dos experimentos realizados ao longo da pesquisa, os resultados obtidos e uma discussão sobre os resultados, subsidiada por análises estatísticas. Também descreve os experimentos realizados com o uso de algoritmos de aprendizado supervisionado e não supervisionado, apresenta um conjunto de conclusões sobre o trabalho realizado e possíveis continuidades.

Capítulo 2

Revisão Bibliográfica sobre Análises e Previsões em Domínio de Dados Eleitorais

2.1 Considerações Iniciais

Diversas abordagens de pesquisa já foram utilizadas para conduzir análises e previsões de resultados de eleições, que aconteceram ou vão acontecer, em diversos países do mundo. Entre as mais populares estão a utilização de mensagens do *Twitter* [Tumasjan *et al*. 2010], a utilização de mensagens do *Twitter* agregadas com técnicas de análise de sentimento [Bermingham & Smeaton 2011], exibição de páginas da Wikipedia sobre artigos politicamente relevantes [Yasseri & Bright 2016], entre muitas outras. Neste capítulo serão abordadas, de forma sucinta, algumas dessas abordagens reportadas na literatura, de maneira a fornecer um contexto para a pesquisa realizada e descrita nesta dissertação, por meio da identificação junto à literatura, de métodos e técnicas utilizados em análises e, também, em previsão de resultados. A Tabela 2.1 apresenta um resumo dos trabalhos revisados.

Tabela 2.1 Resumo do levantamento bibliográfico realizado.

Ano	Referências	Proveniência dos dados	Comentários sobre técnicas utilizadas
2003	[Brender 2003]	Dados Fiscais e Eleitorais	Responsabilidade Fiscal
2009	[Antonakis & Dalgas 2009]	Imagens	Traços não-verbais e utilização de imagens
2010	[Tumasjan *et al.* 2010]	Twitter	Contagem de tweets
2011	[Bermingham & Smeaton 2011]	Twitter	Contagem de tweets e Análise de Sentimentos
2011	[Cantu & Saiegh 2011]	Dados Eleitorais	Algoritmos de AM
2012	[Sang & Bos 2012]	Twitter	Contagem de tweets e Análise de Sentimentos
2012	[Boutet *et al.* 2012]	Twitter	Contagem de tweets
2014	[Mattes & Milazzo 2014]	Imagens	Traços não-verbais e utilização de imagens
2015	[Coletto *et al.* 2015]	Twitter	Análise de Sentimentos e Algoritmos de AM
2016	[Bessi & Ferrara 2016]	Twitter	Robôs Sociais
2016	[Yasseri & Bright 2016]	Wikipedia	Visualização de páginas do Wikipedia
2017	[Ortiz-Ángeles *et al.* 2017]	Rede Social YouGov	Algoritmos de AM
2019	[Prabhu *et al.* 2019]	Twitter	Análise de Sentimentos

2.2 Abordagens Investigadas

A abordagem mais simples de previsão de resultado de eleições utilizando o *Twitter* é implementada por meio da contagem do número de vezes que um partido ou, então, um candidato específico, é mencionado nos *tweets* considerados. Os números finais obtidos da contagem são usados como preditores do resultado das eleições [Tumasjan *et al*. 2010]. Obviamente esse procedimento é muito simplista para receber qualquer crédito. É fácil intuir que a simples contagem de ocorrências de nomes de partidos ou nomes de candidatos nos *tweets* não é suficiente quando os comentários contidos no texto analisado não são considerados. Nomes de ambos, partidos e candidatos podem estar em um contexto positivo ou negativo.

Em [Boutet *et al.* 2012] os autores buscaram identificar as tendências políticas dos eleitores nas eleições gerais do Reino Unido em 2010 utilizando essa abordagem. A pesquisa analisou as características dos três principais partidos envolvidos no processo eleitoral, propondo um algoritmo simples e prático para identificar a tendência política dos usuários usando a quantidade de mensagens do *Twitter* que parecem relacionadas a partidos políticos. Os *tweets* foram coletados entre os dias 5 e 12 de maio de 2010. Foram mantidos apenas 419 tópicos com mais de 10.000 *tweets,* o que gerou um conjunto de dados com mais de 220.000 usuários responsáveis pelo envio de aproximadamente 1.150.000 *tweets*. Para esses usuários foram coletados os respectivos perfis e cerca de 79.000.000 de relacionamentos de seus seguidores, possibilitando a identificação manual das afiliações políticas e gerando um conjunto de dados verdadeiros. Sobre esse conjunto de dados foi utilizada uma técnica conhecida como método de propagação de rótulos [Raghavan, *et al*. 2007], o que possibilitou a detecção de 5.878 candidatos trabalhistas, 3.214 liberais e 2.356 conservadores, com alta precisão de 0,77; 0,78 e 0,90, respectivamente. A pesquisa identificou as duas principais formas de diferenciar os partidos políticos: (1) o gráfico de retuite apresentou uma estrutura partidária altamente segregada e (2) os membros do partido eram mais propensos a fazer referência ao seu próprio partido do que a outro. Os autores concluíram que o método de classificação proposto é capaz de atingir uma precisão de 86% sem nenhum treinamento, o que tornaria esse método de pesquisa uma solução perfeita para classificação em tempo real.

De maneira semelhante ao trabalho anteriormente abordado, a pesquisa descrita em [Tumasjan *et al*. 2010] utiliza o número de *tweets* que mencionam um partido político, como preditor dos resultados de uma eleição federal alemã. Durante a pesquisa foram examinados 104.003 *tweets* públicos relacionados à política, coletados entre 13 de agosto e 19 de setembro de 2008, semanas antes das eleições. Os *tweets* filtrados continham o nome dos 6 partidos políticos do parlamento alemão ou de políticos pertencentes a esses partidos que aparecem nas pesquisas de popularidade do instituto de pesquisa "Forschungsgruppe Wahlen". Essa consulta resultou na identificação de 70.000 *tweets* que mencionavam um dos partidos políticos e de 35.000 *tweets* referentes a nomes de políticos envolvidos em partidos. A extração dos sentimentos foi realizada com a utilização do software de análise de texto LIWC2007 (*Linguistic Inquiry and Word Count*) [Pennebaker *et al.* 2007] que, por meio do uso de um dicionário interno, analisa componentes emocionais, cognitivos e estruturais de amostras de texto. Para entender se a atividade no *Twitter* pode servir como um preditor do resultado da eleição, a pesquisa examinou dois aspectos. Primeiro, foram comparadas as parcelas de atenção que cada um dos partidos políticos recebeu no *Twitter*, com o resultado das eleições. Segundo, foram analisados se os *tweets* poderiam, de alguma forma, disponibilizar informação sobre os laços ideológicos entre partidos e possíveis coalizões políticas após a eleição. Respondendo às questões de pesquisa, foi descoberto que mais de um terço de todas as mensagens faziam parte de conversas, o que é um indicativo de que o *Twitter* não é usado apenas para espalhar opiniões políticas, mas também para discuti-las com outros usuários, e que os perfis de sentimentos de políticos e partidos refletem plausivelmente muitas nuances de uma campanha eleitoral. A pesquisa concluiu que o *Twitter* é realmente usado como uma plataforma para deliberação política e o mero número de *tweets* refletiam as preferências dos eleitores; os resultados apresentados eram próximos daqueles apresentados pelas pesquisas eleitorais tradicionais. Quanto aos sentimentos das mensagens no *Twitter*, correspondem de perto a programas políticos, perfis de candidatos e evidências da cobertura da mídia na campanha. A inclusão de um processo que realiza análise de sentimentos, associado às contagens de *tweets*, ofereceu uma maneira de melhorar a capacidade preditiva dos modelos subsidiados por contagens.

Em [Bermingham & Smeaton 2011] os autores preveem os resultados de uma eleição irlandesa por meio da contagem de *tweets* e análise de sentimentos. Foram coletados 32.578 *tweets*, entre os dias 8 e 25 de fevereiro de 2011, cujos respectivos textos faziam menção aos cinco principais partidos envolvidos na eleição. Os *tweets* relevantes foram identificados por meio da identificação do nome dos partidos e suas abreviações, juntamente com a *hashtag* das eleições. Para avaliar uma possível proposta, foi desenvolvido um aplicativo em parceria com uma empresa de notícias *online*. O objetivo do aplicativo "*Twitter Tracker*" foi possibilitar que usuários e jornalistas das empresas parceiras acessassem o conteúdo do *Twitter*, referente à eleição, bem como também tivessem acesso aos resumos e à visualização do volume e sentimento ao longo do tempo. Pesquisas anteriores haviam mostrado que o aprendizado supervisionado poderia fornecer uma análise de sentimentos mais precisa do que a fornecida por métodos não supervisionados, como o uso de léxicos de sentimentos. Com base nessas informações, os autores utilizaram anotadores *i.e.*, pessoas instruídas a anotar sentimentos em *tweets* relacionados a partidos e candidatos à eleição, que não representassem sentimentos positivos ou negativos, mas sim, de emoção, opinião, avaliação ou especulação. As categorias de anotações foram definidas como (1) três classes de sentimentos (positivo, negativo, misto), (2) uma classe de não-sentimento (neutro) e (3) três outras classes (não anotável, não relevante, pouco clara), de acordo com [Wilson *et al.* 2005].

Todas as anotações não relevantes, duplicadas, termos do tópico, nomes de usuários e URLs, foram desprezadas ou removidas. Os algoritmos Support Vector Machines e Multinomial Naive Bayes (MNB) foram então utilizados, tendo por entrada o conjunto de dados resultante; ambos produziram resultados insatisfatórios. Esse fato levou os autores a usarem o algoritmo Adaboost M1 [Freund & Schapire 1996], com 10 iterações de treinamento seguido do algoritmo Adaboost MNB que alcançou 65,09% de precisão em um processo de 10-validação cruzada. A pesquisa concluiu que o *Twitter* parece exibir uma qualidade preditiva que é marginalmente aumentada pela inclusão da análise de sentimentos.

Uma abordagem semelhante foi adotada no trabalho descrito em [Sang & Bos 2012], para as eleições no senado holandês de 2011. Os autores partiram do princípio de que o uso de análise de sentimentos melhoraria consideravelmente as previsões baseadas na contagem de entidades em *tweets*, tornando-se quase tão boas quanto os métodos

tradicionais de pesquisas eleitorais. Foram coletados 64.395 *tweets* na semana anterior à eleição entre 23 de fevereiro de 2011 a 1 de março de 2011. Os autores procuraram por mensagens que continham pelo menos uma, de uma lista de 100 palavras holandesas utilizadas com alta frequência, assim como *tags* de assuntos em holandês. Nesse processo foram encontradas palavras aparentemente holandesas, mas que eram escritas em outro idioma sendo necessário a aplicação do adivinhador de linguagem desenvolvido por Thomas Mangin [Mangin 2007], que classifica os idiomas comparando n-gramas de caracteres de um texto de entrada com modelos de n-gramas de textos em idiomas conhecidos. Os autores destacaram outros problemas que requereram análises prévias, tais como: vários *tweets* enviados por uma mesma pessoa; uma mesma pessoa pode ter enviado *tweets* sobre diferentes partidos políticos; nem toda mensagem contendo um partido é positiva; o *Twitter* é muito popular entre os adolescentes holandeses, mas eles não podem votar; idosos são sub-representados na Internet, mas têm grande participação nas eleições. Após as etapas de normalização, restaram 28.704 *tweets*. As ocorrências de nomes de partidos foram extraídas dos *tweets*, contadas e convertidas em assentos no Senado. Cada *tweet* que mencionava o nome de um partido passou a ter peso de um voto para aquele partido**.** Os números de assentos eleitorais previstos pelos *tweets* foram próximos aos resultados da eleição. Nas conclusões, os autores comentam que apenas contar os *tweets* que mencionam partidos políticos não é suficiente para obter boas previsões, porém, apesar de não terem dados de treinamento de alto padrão, o erro total final foi 29% maior do que aqueles obtidos por empresas de pesquisa experientes.

Em [Coletto *et al*. 2015] as abordagens de contagem de *tweets* e análise de sentimento são levadas um passo adiante, com a utilização de algoritmos de aprendizado de máquina para reduzir o viés da amostra de usuários do *Twitter*. Essa nova abordagem foi então avaliada nas eleições primárias do principal partido político italiano: o “Partido Democrático”. Esse estudo utilizou um conjunto de dados de 1,7 milhões de *tweets* coletados 10 dias antes e 5 dias após as eleições de 2013. Esse conjunto de dados foi reduzido a 95.627 por meio da exclusão de dados parciais e *tweets* irrelevantes, permanecendo os dados de *tweets* italianos com base no idioma declarado pelos usuários do *Twitter* e no idioma detectado por um classificador de aprendizado de máquina do *Twitter*. Os autores utilizaram vários estimadores ou preditores na tentativa de produzir

uma estimativa da parcela de votos que o candidato receberia, auxiliados por três medidas diferentes de avaliação, realizando uma análise por região, o que significa que uma previsão seria produzida para cada região explorando apenas os dados regionais cujos resultados abrangessem 20 regiões italianas.

O resultado foi um conjunto de treinamento usado para aprender os pesos associados a cada região, por meio de regressão linear e aplicados aos preditores que apresentaram melhor desempenho: UserShare e ClassTweetCountC. O UserShare foi capaz de fornecer a classificação correta dos candidatos em 15 das 20 regiões. A pesquisa propôs uma melhoria na estratégia de contagem de usuários, por entender que a relação de correspondência entre um usuário do *Twitter* e um eleitor não é satisfeita, pois usuários que mencionam mais de um candidato são levados em consideração várias vezes. Os autores utilizaram probabilidade para realizar a normalização dizendo que um usuário u $\in$ U provavelmente votará no candidato c $\in$ C, bem como para aprimorar a classificação da polaridade dos *tweets* para os candidatos, apresentando uma aproximação, com a suposição usual de que mencionar um candidato equivale a votar em um candidato. Os autores acreditam que grandes melhorias podem ser alcançadas por meio da integração de várias fontes de informação, tais como pesquisas tradicionais, múltiplas redes sociais, dados demográficos, dados históricos, análises de eventos relacionados, propriedades baseadas em conteúdo e em rede e concluem que essa riqueza de informações pode ser totalmente explorada por meio de abordagens de aprendizado de máquina.

Segundo o último Incapsula Bot Traffic Report [Imperva 2020], um estudo destinado ao levantamento de estatísticas do tráfego de contas automatizadas na Internet, divulgado em 2016, apenas 48% da atividade *online* vem de humanos, enquanto 52% é delegada a robôs. De forma geral os *bots,* termo derivado da palavra *"robot"*, são aplicações autônomas que rodam na Internet enquanto desempenham algum tipo de tarefa pré-determinada. Com a abertura de APIs que permitem aos desenvolvedores ampliarem as funcionalidades padrão de mídias sociais, tais como *Facebook* e *Twitter,* tem surgido os chamados robôs sociais, algoritmos que tem afetado as redes sociais de forma negativa e que podem ser utilizados para manipular discussões *online*, mudar a percepção do público sobre entidades políticas ou mesmo tentar afetar o resultado de eleições políticas.

Neste contexto [Bessi & Ferrara 2016] avaliam o impacto dos robôs sociais do *Twitter* nos resultados das eleições de 2016 nos EUA. Os autores elaboraram uma lista de *hashtags* e palavras-chave associadas a cada um dos principais candidatos, composta por 23 termos, incluindo cinco termos específicos para o candidato do Partido Republicano Donald Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), quatro termos para a candidata indicada pelo Partido Democrata, Hillary Clinton (#hillaryclinton, #imwithher, #nevertrump, #hillary) e vários termos relativos aos debates, bem como termos associados a outros dois candidatos.

Visando garantir resiliência e escalabilidade, a infraestrutura de coleta de dados foi executada dentro de uma instância da *Amazon Web Services* (AWS), onde os autores submeteram a lista de *hashtags* e palavras-chave para consultas realizadas com o auxílio da API de pesquisa do *Twitter* em intervalos regulares de 10 segundos, continuamente e sem interrupções em três períodos entre 16 de setembro e 21 de outubro de 2016. Foram coletados 20,7 milhões de *tweets* postados por quase 2,8 milhões de usuários distintos, dos quais foram testadas apenas as 50.000 contas principais classificadas por volume de atividades. A dificuldade em detectar se o controle de uma conta de mídia social é feito por um humano ou por um *bot* fez com que os autores utilizassem uma solução chamada BotOrNot [Davis *et al*. 2016], plataforma que avalia se uma conta do *Twitter* é controlada por humanos ou por máquina, analisando um conjunto de características como conteúdo e estrutura de rede, atividade temporal, dados de perfil de usuário e análise de sentimento, produzindo uma pontuação que sugere a probabilidade de que a conta inspecionada seja de fato um *bot* social *i.e.*, se a pontuação estiver acima de 50%.

Embora esses 50 mil usuários principais representem cerca de apenas 2% de toda a população, é importante notar que eles são responsáveis pela produção de mais de 12,6 milhões de *tweets*, o que é cerca de 60% do total de conversas. Um total de 7.183 usuários, responsáveis por 2.330.252 *tweets* foram classificados como *bots*. Os autores estimaram que cerca de 400.000 *bots* estiveram envolvidos na discussão política sobre a eleição presidencial de 2016 nos EUA, sendo responsáveis por cerca de 3,8 milhões de *tweets*, representando um quinto de toda a conversa. Os autores concluíram que a presença de *bots* sociais em discussões políticas *online* pode criar três questões tangíveis: primeiro, a influência pode ser redistribuída entre contas suspeitas que podem ser operadas com fins maliciosos; segundo, a conversa política pode se tornar ainda mais

polarizada; terceiro, a disseminação de desinformação e informação não verificadas pode ser aprimorada, como também a impossibilidade de determinar quem os opera.

Os autores em [Prabhu *et al*. 2019] propõem um método para selecionar um candidato com maior probabilidade de ser eleito com base, entre outros fatores, na análise de sentimentos dos *tweets* que se referem a ele. A pesquisa propôs determinar o candidato com base em sua popularidade nas redes sociais como o *Twitter*, coletando os *tweets* e montando um dicionário próprio com algumas palavras que correspondessem a esses *tweets*. O candidato receberia pontos positivos ou negativos de acordo com o significado das palavras, e sobre essa pontuação uma fórmula matemática seria aplicada para calcular a porcentagem de sua popularidade. Os autores sugerem que cada candidato seja avaliado de acordo com os seguintes parâmetros: qualificação educacional; registros criminais; trabalho social; status social e popularidade; registros eleitorais anteriores, onde uma ponderação adequada seria dada a cada parâmetro. Após a coleta de todas as informações relacionadas ao candidato, os autores apontam a utilização do algoritmo Naive Bayes para prever o candidato com maior mérito dentre todos os candidatos de um determinado distrito em particular e concluem afirmando que esse tipo de abordagem analítica pode ser usada para qualquer tipo de eleição na Índia e em outros lugares, bem como destacam o poder da análise de dados no campo da Ciência da Computação e a clareza que o uso de dados sociais pode oferecer para a análise de qualquer tipo de pessoa.

Outras fontes de dados para prever os resultados das eleições também têm sido utilizadas. Em [Yasseri & Bright 2016], os autores exploraram o potencial do uso de dados de exibição de páginas da Wikipedia, considerando estatísticas de leitores de artigos politicamente relevantes (como os de partidos políticos individuais), no período imediatamente anterior a uma eleição, como um preditor para os resultados da eleição, e testam esse modelo de análise usando como fonte de dados a base de dados das eleições do parlamento Europeu de 2009 e 2014.

Os autores coletaram e utilizaram dois tipos de dados sobre essas eleições: as estatísticas de visualização de página da página geral da Wikipedia sobre as eleições de 14 edições da Wikipedia em diferentes idiomas, cada uma representando um dos países que foram às urnas no dia da eleição, e um conjunto de dados de partidos políticos que

competiram em uma ou ambas as eleições de 2009 e 2014 nos cinco maiores países da Europa Ocidental: Reino Unido, França, Alemanha, Espanha e Itália, focando apenas nos partidos que obtiveram mais de 5% dos votos em qualquer um dos anos.

Nos experimentos também foram registradas variáveis de interesse tal como a quantidade de visualizações que a página do partido político em questão, na edição em idioma correspondente da Wikipedia, recebeu na semana antes da eleição. A partir do conjunto de dados, os autores examinaram a relação entre os padrões de tráfego da Wikipedia em torno da época das eleições e o comparecimento eleitoral geral, e desenvolveram um modelo para tentar prever os resultados da votação absoluta para diferentes partidos políticos e verificaram se o modelo em questão poderia ser desenvolvido para mudanças na participação de votos. Os autores perceberam que o número de visualizações da página geral da Wikipedia sobre a eleição poderia oferecer uma estimativa razoável da mudança relativa na participação nas eleições analisadas e que um modelo teoricamente subsidiado por resultados nacionais anteriores, visualizações de páginas da Wikipedia, menções à mídia de notícias e informações básicas, poderia oferecer uma boa previsão da parcela geral de votos do partido de interesse. Os autores, entretanto, concluem que tais informações não podem oferecer resultados confiáveis sobre o comportamento individual.

Já o trabalho realizado e descrito em [Ortiz-Ángeles *et al.* 2017] usou dados da Rede Social YouGov para determinar as preferências eleitorais para as eleições presidenciais primárias dos EUA, em 2016. Os conjuntos de dados utilizados na pesquisa foram compilados a partir de informações *online* fornecidas por 1.200 usuários da rede social YouGov e pelas respostas fornecidas por esses usuários a um questionário *online* aplicado pelo American National Eleection Studies (ANES) entre 22 e 28 de fevereiro de 2016. O questionário contém duzentas e treze questões relacionadas a economia, raça, violência, tendências mundiais e outros temas políticos, bem como a intenção de voto para as eleições primárias de 2016. Durante a pesquisa foram analisados os comportamentos de desempenho de 25 algoritmos supervisionados disponíveis na plataforma WEKA, na busca da intenção de voto para quatro cenários: a) Prever o candidato democrata em que o eleitor votaria; b) Prever o candidato republicano em que o eleitor votaria; c) Prever se o eleitor votaria em um candidato democrata ou não; d) Prever se o eleitor votaria em um candidato republicano ou não. Após os testes

estatísticos, os autores puderam verificar que os algoritmos BFTree e CART foram os melhores para prever as intenções de voto para um único candidato e para o voto democrata/republicano, respectivamente. A análise de teste de hipóteses dos resultados experimentais indicou que prever as intenções de voto a favor de um candidato democrata ou republicano é mais simples do que prever o candidato em particular, visto que os desempenhos de previsão para um candidato democrata ou republicano (melhores desempenhos de 80% e 78%, respectivamente) são melhores do que os dados na previsão de um candidato específico (70% para candidatos democratas e 56% para candidatos republicanos).

Além disso, os autores constataram que em ambas as situações avaliadas, 9 dos 25 algoritmos ofereceram resultados significativamente piores, enquanto 16 classificadores ofereceram bons resultados, que também não são significativamente diferentes entre eles. Por fim, os autores também constataram que os atributos retirados do perfil do usuário YouGov são suficientes para prever um voto democrata, enquanto os dados retirados do perfil YouGov e do questionário ANES são necessários para prever uma preferência de voto para um candidato republicano. No entanto, as informações do estudo da ANES dão o melhor resultado, ao preverem a preferência de voto para um determinado candidato, seja ele democrata ou republicano.

Embora não seja baseado em dados coletados a partir de redes sociais, o trabalho descrito em [Brender 2003] explora o impacto que a responsabilidade fiscal dos prefeitos em Israel teve em suas chances de reeleição durante o período de 1989-1998, e discute sobre as condições sob as quais os eleitores recompensariam uma política financeira prudente em nível local. O autor utilizou um banco de dados único contendo dados financeiros de diversas fontes, resultados eleitorais e informações sobre o desempenho do sistema educacional, em busca de uma relação entre prudência fiscal e reeleição. Não foram encontrados efeitos estatisticamente significativos das variáveis fiscais sobre os resultados das campanhas eleitorais em 1989 e 1993.

Quanto à eficácia da “economia do ano eleitoral”, o autor descobriu que em nenhuma das três campanhas as políticas expansionistas, refletidas por um maior acúmulo de dívida *per capita* durante o ano eleitoral, importavam. Os coeficientes em todas as três campanhas não são significativos e são negativos em 1989 e 1993,

indicando que grandes déficits durante um ano eleitoral estão associados a uma probabilidade menor de reeleição. O único coeficiente significativo é da proporção de votos recebidos nas últimas eleições, como preditor de reeleição, sendo que as equações não têm sucesso em prever os resultados das eleições, adicionando apenas marginalmente uma previsão baseada na suposição de que todos os prefeitos seriam reeleitos.

O autor concluiu que embora o desempenho fiscal não tenha sido um fator relevante para as decisões dos eleitores nas campanhas de 1989 e 1993, os eleitores em 1998 parecem ter sido substancialmente afetados por ele. Também pareceu ao autor que o progresso em três áreas críticas para uma supervisão eleitoral eficaz contribuiu para essa mudança: a) tendência do eleitor a se concentrar em questões locais; b) melhor disponibilidade de informações; c) a imposição de uma restrição orçamentária mais dura por parte do governo.

Outras propostas baseadas em características não verbais também foram utilizadas nas previsões de eleições. O projeto apresentado em [Antonakis & Dalgas 2009] descreve um experimento em que adultos e crianças suíços são convidados a selecionar um vencedor entre dois candidatos concorrentes, no segundo turno das eleições parlamentares francesas em 2002. Os autores usaram 57 pares (3 mulheres, 54 homens) de fotos de rostos de candidatos às eleições. Nenhum dos participantes tinha conhecimento prévio dos candidatos. Em um primeiro experimento, 684 estudantes universitários públicos suíços, dos quais 43,71% eram mulheres, avaliaram qual dos dois candidatos era mais competente, mais inteligente e melhor líder através de um questionário composto por um par de rostos.

No segundo experimento, 681 crianças com 13 anos de idade ou menos e 160 participantes mais velhos com idade média de 30,49 anos, também suíços, receberam um questionário com um par de rostos. Esses participantes foram submetidos a um jogo experimental que reencenava a viagem de Odisseu de Tróia a Ítaca com o objetivo de voltar para casa o mais rápido possível, e então foi pedido aos participantes que imaginassem que iriam repetir a viagem naquele momento e precisavam indicar um capitão para seu barco. Tanto adultos quanto crianças escolheram o vencedor real em

mais de 70% dos casos, usando como entrada apenas a confiabilidade percebida a partir da exibição de fotos dos rostos dos candidatos.

Da mesma forma [Mattes & Milazzo 2014] exploram o impacto de traços positivos e negativos nas imagens dos candidatos no resultado das eleições parlamentares britânicas. Os procedimentos da pesquisa foram realizados em novembro de 2010 na Universidade de Iowa, com a participação de 153 estudantes graduados sendo 88 mulheres e 65 homens, entre eles, alguns estudantes dos Estados Unidos, utilizados para atenuar os efeitos de familiaridade e/ou partidarismo.

Os participantes do estudo receberam estímulos apresentados em um monitor dentro do laboratório de informática com o auxílio do software DirectRT. Primeiro foram apresentadas as instruções que explicavam a tarefa e enfatizavam a importância de dedicar tempo suficiente para tomar decisões precisas. Depois foi solicitado aos participantes que julgassem as imagens faciais dos candidatos britânicos levando em consideração duas características (atratividade e competência) e dissessem em qual candidato eles teriam maior probabilidade de votar. A apresentação das imagens seguiu o protocolo TED [Kim *et al.*, 2007], que mostra as fotografias dos candidatos, uma de cada vez, em vez de simultaneamente, forçando assim uma codificação do rosto na memória para a comparação. Foram exibidas imagens de dois candidatos em cada análise. As imagens foram mostradas a cada participante por 75 milissegundos, de maneira alternada, com um intervalo entre imagens de um segundo. Os participantes indicavam sua escolha de acordo com as características que estavam sendo julgadas. As imagens continuaram sendo mostradas a cada 60 segundos, e as próximas imagens apenas eram mostradas depois que o participante já tivesse escolhido uma imagem entre os dois candidatos anteriores. Como os estudos em [Spezio *et al.* 2012] demonstram que os aspectos não faciais do *headshot* de um candidato podem influenciar na tomada de decisão do eleitor, os autores mostraram apenas imagens com fundos neutros.

Por fim os autores concluem que usando apenas julgamentos de atratividade e competência relativas, puderam prever com sucesso os resultados em aproximadamente 60% dos candidatos incluídos no estudo e que levando em conta a marginalidade eleitoral, a precisão das previsões aumentaram consideravelmente, tendo como efeito mais notável, que candidatos considerados mais atraentes têm uma probabilidade

significativamente maior de vencer as eleições atingindo 72%, também observaram que de forma mais geral, o fato de descobrirem que os atributos não relacionados a políticas dos candidatos podem ser usados nos resultados das eleições sugere a importância de compreender as respostas dos eleitores sobre as imagens dos candidatos e, no futuro, compreender a capacidade dos candidatos de manipular tais imagens.

Os autores em [Cantu & Saiegh 2011] apresentam um modelo para detectar resultados fraudulentos de eleições e testam o modelo com as contagens de votos em nível distrital na província de Buenos Aires entre 1931 e 1941, período em que as eleições fraudulentas eram conhecidas. Segundo os autores a principal inovação foi a do uso de dados sintéticos para desenvolver e treinar um protótipo de detecção de fraude eleitoral. Devido a problemas de disponibilidade de dados, os autores utilizaram os métodos de Monte Carlo para gerar grandes quantidades de dados eleitorais que mantinham as propriedades estatísticas de um conjunto selecionado de dados autênticos usados como base. Em seguida, empregaram um classificador Naive Bayes como algoritmo de aprendizado. Os autores definiram um procedimento com 4 etapas: (1) foi criado um conjunto de eleições simuladas. Este conjunto de treinamento era composto por dois subconjuntos disjuntos: um contendo contagens de votos que seguem uma distribuição com propriedades conhecidas e outro em que os dados eram propositalmente “manipulados”; (2) os valores foram selecionados através de análise digital e, em seguida, os valores de adesão das eleições simuladas foram limpos e manipulados usando regressão logística; (3) foram recuperadas as densidades condicionais de classe usando as frequências relativas do conjunto de treinamento; (4) foi avaliada a capacidade de detecção do classificador usando dados autênticos extraídos de um novo conjunto de dados de contagens de votos em nível de distrito na província de Buenos Aires (Argentina) entre 1931 e 1941. Os autores afirmam que os resultados do processo evidenciam a viabilidade do uso de dados sintéticos para treinar e testar um sistema de detecção de fraude eleitoral e que o estudo realizado forneceu evidências indiscutíveis do escopo e intensidade da fraude eleitoral durante a "década infame" da Argentina. Por fim, concluem que as descobertas confirmam que a fraude eleitoral, em vez de uma mudança nas preferências dos eleitores, levou a mudanças eleitorais dramáticas durante o período em questão.

2.3 Comentários sobre a Pesquisa Bibliográfica Realizada

A pesquisa bibliográfica realizada teve como objetivo encontrar trabalhos com estratégias semelhantes ao trabalho desenvolvido. Uma quantidade substancial de pesquisas pode ser encontrada na literatura relacionadas a eleições, mesmo que o objetivo não seja prever seus resultados. Como também podemos encontrar na literatura uma ampla gama de abordagens que trata da previsão dos resultados de eleições. Os trabalhos descritos nos artigos revisados, comentados anteriormente, apresentam estratégias diversificadas, na tentativa de encontrar soluções que apresentassem resultados próximos ou melhores do que as metodologias tradicionais de previsão do resultado de eleições, sendo elas: (a) contagem de *tweets*; (b) análise de sentimentos; (c) uso de algoritmos de aprendizado de máquina; (d) impacto das contribuições de campanha sobre os votos dos parlamentares; (e) traços não-verbais e utilização de imagens; (f) impacto da responsabilidade fiscal nas chances de reeleição. Porém, fica claro a existência de severas limitações para os tipos de previsões eleitorais que podem ser feitas com base no que as pessoas dizem nas redes sociais. Desafios relacionados à análise de sentimento, viés populacional, *bots* sociais e spam, apenas para citar alguns, acabam gerando resultados mistos, aparentemente funcionando bem em alguns casos e produzindo resultados decepcionantes em outros.

Capítulo 3

Aprendizado de Máquina

Aprendizado de Máquina (AM) é uma área de conhecimento com ampla literatura e que tem como foco aspectos teóricos e práticos de procedimentos que implementam aprendizado automático por computadores. AM é considerada uma subárea da área de Inteligência Artificial (IA). Algoritmos de AM e técnicas para o tratamento de dados são dois aspectos importantes a serem considerados, quando do desenvolvimento de um ambiente computacional que viabiliza aprendizado automático.

Esse capítulo apresenta algumas definições e conceitos básicos sobre AM que são relevantes ao trabalho realizado. As seções desse capítulo se propõem a explicar algumas das principais etapas de uma metodologia que tem por objetivo viabilizar um ambiente automático de AM, em uma área de aplicação (*i.e.*, área de conhecimento associada aos dados a serem utilizados). Também são apresentadas características de algoritmos de AM supervisionados e não supervisionados, considerando que ambos os tipos de algoritmos foram usados nos experimentos apresentados no Capítulo 6.

3.1 Considerações Iniciais

É fato que a capacidade de aprender, por parte de seres vivos e/ou sistemas computacionais, é essencial para um comportamento inteligente em qualquer área de conhecimento. O processo de aprendizado, como um todo, contempla a aquisição de diferentes formas de conhecimento, entre elas, atividades de memorização, observação e exploração das situações. Por meio do uso de instruções fornecidas por uma fonte externa ou, então, de situações observadas pelo aprendiz, novos fatos podem ser descobertos e incorporados pelo aprendiz, o que resulta em um melhoramento de seu desenvolvimento motor e cognitivo.

3.2 Aprendizado de Máquina

Com a crescente complexidade dos problemas e, consequentemente, com os sistemas computacionais implementados para solucioná-los e, também, considerando o grande volume de dados gerados, tornou-se evidente a necessidade de abordagens mais autônomas que permitissem a redução de intervenção humana. Esse fato provocou a necessidade do desenvolvimento de técnicas capazes de, a partir de experiências passadas, induzir hipóteses por meio das quais o problema tratado pudesse ser resolvido. Esse processo de indução de uma hipótese levando em consideração experiências passadas caracteriza o processo realizado por um algoritmo de aprendizado de máquina.

Mitchell, em [Mitchell 1997] define Aprendizado de Máquina (AM) como a área de pesquisa que visa desenvolver programas computacionais capazes de melhorar seu desempenho por meio da experiência. Considerando o estágio atual da área de AM, apesar do objetivo final de AM ser o desenvolvimento de sistemas inteligentes que simulam o aprendizado humano, muito ainda precisa ser pesquisado e elaborado, tanto na área de AM quanto nas áreas que a subsidiam, para viabilizar tal objetivo.

Similarmente à linha de pensamento adotada por Mitchell, Alpaydin, em [Alpaydin 2010], define Aprendizado de Máquina como a programação de computadores para otimizar um critério de desempenho, usando como dados, exemplos ou experiências anteriores. De acordo com o autor, a ideia principal é que técnicas utilizadas em AM sejam capazes de aprender e resolver problemas, se adaptando ou mudando seus comportamentos com base nesses exemplos, sendo assim capazes de induzir informações, com base nos dados fornecidos.

Deve ser ressaltado que os processos de aprendizado automático não ocorrem instantaneamente; ocorrem por meio de processos iterativos que, em determinas situações, exigem algum tipo de interação com o usuário do sistema. Essa interação, em que o conhecimento prévio, juntamente com as novas observações podem levar a novos conhecimentos, está relacionada a métodos de inferência lógica.

Dentre as muitas subáreas de pesquisa em AM, merecem destaque: (1) métodos de busca, particularmente, busca cega, busca heurística e busca contraditória, mais comumente conhecida, como busca min-max; (2) diferentes formas para a representação

de conhecimento, tanto existentes quanto induzidas, tais como lógica, probabilidade, conjunto de instâncias de dados, vetores de dados, árvores de decisão, agrupamento de dados, redes neurais, etc.., (3) proposta de um número considerável de algoritmos indutivos que, a partir de um de um conjunto de instâncias de um determinado conceito a ser aprendido, induz uma expressão geral que presenta tal conceito. De uma maneira simplista, tais algoritmos podem ser abordados como algoritmos simbólicos (e.g., aquele que produz como expressão induzida do conceito uma árvore de decisão, que pode ser 'traduzida' como um conjunto de regras lógicas proposicionais) ou algoritmos conexionistas (*e.g.*, aquele que produz como expressão induzida do conceito uma rede neural).

Como discutido em [Nicoletti 1994], com exceção de algumas propostas de AM que fazem tradução de linguagens formais em alto nível para alguma de nível mais básico e, para tanto, usam procedimentos dedutivos, toda outra forma de AM empregada é, basicamente, indutiva. Dedução pode ser entendida como a conclusão obtida a partir de um conjunto de premissas iniciais, supostamente verdadeiras, à semelhança quando do uso da lógica formal, para a prova de argumentos lógicos. Devido às características da pesquisa proposta neste documento, métodos dedutivos de AM não se adequam ao trabalho pretendido e, portanto, não foram considerados.

Usualmente o processo de aprendizado de máquina pode ser caracterizado como um processo indutivo que envolve duas fases. Na primeira fase do processo, que é conhecida como fase de *treinamento*, a partir de um conjunto de instâncias de treinamento (ou exemplos de treinamento), que representam um determinado conceito (ou classe), o processo induz uma descrição geral do conceito que, em muitas referências, é denominado de modelo ou expressão do conceito. Tal descrição pode ser uma estrutura de árvore de decisão, um conjunto de regras, um conjunto de instâncias de dados, etc..., na dependência do algoritmo utilizado. Na segunda fase, nomeada de fase de *classificação*, o modelo induzido na primeira fase é utilizado para a categorização de novas instâncias de dados (identificadas como conjunto de teste), como pertencentes ou não ao conceito representado pela estrutura induzida (modelo). Como as instâncias pertencentes ao conjunto de teste têm já uma classe associada, o procedimento de validação do modelo consiste em avaliar quanto o modelo categoriza corretamente as

instâncias de dados do conjunto de teste. A Figura 3.1 exibe um diagrama do processo de indução do modelo, a partir do conjunto de treinamento.

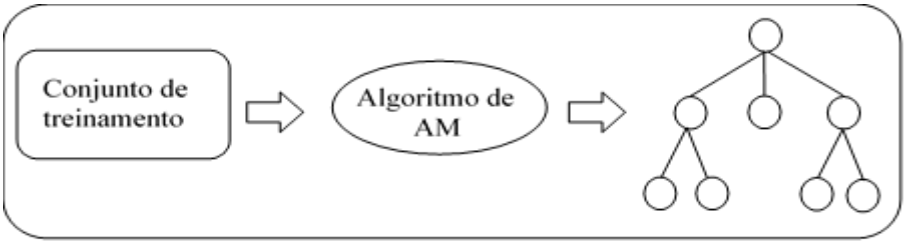


Figura 3.1 Esquema geral de um processo de AM, em que o modelo (classificador) induzido é uma árvore de decisão.

Na literatura relativa a AM podem ser encontradas diversas taxonomias que buscam organizar algoritmos de AM em grupos, em função de algumas de suas características. Como informado em [Nicoletti 2018], uma taxonomia geral e abrangente, em que algoritmos podem ser caracterizados como: (1) supervisionado (2) não supervisionado e (3) semi-supervisionado é usualmente adotada, como mostra o diagrama da Figura 3.2.

Dentre os principais algoritmos supervisionados estão os que induzem classificadores e aqueles que realizam o aprendizado por regressão. Algoritmos não supervisionados são considerados organizadores do conjunto de dados que lhes é fornecido, sendo os algoritmos de agrupamento (*clustering*) os mais populares desse grupo de algoritmos de AM. Uma breve descrição sobre algoritmos de agrupamento é apresentada na Seção 3.4 e o uso do algoritmo k-Means, um dos mais populares algoritmos de agrupamento, é também mostrado na seção. O k-Means foi também utilizado em alguns dos experimentos realizados durante a pesquisa descrita nesta dissertação que estão descritos no Capítulo 6.

Como comentado em [Nicoletti 2018], algoritmos que implementam aprendizado semi-supervisionado são adequados para situações em que o conjunto de treinamento é formado por dois grupos de instâncias com as seguintes características: (a) um grupo, geralmente com um número pequeno de instâncias, em que as instâncias do grupo têm, associada a elas, a classe à qual pertencem; (b) um segundo grupo de instâncias, geralmente com um número alto de instâncias, em que as instâncias do grupo não têm uma classe a elas associada. Algoritmos que implementam aprendizado semi-supervisionado geralmente utilizam o grupo (1) como conjunto de treinamento para induzir uma expressão geral do conceito (modelo) e, então, utilizam esse modelo para

classificar as instâncias do grupo (2). Os algoritmos Self-Training [Rosenberg *et al.* 2005] e o Co-Training [Blum & Mitchell 1998] são considerados algoritmos semi-supervisionados.

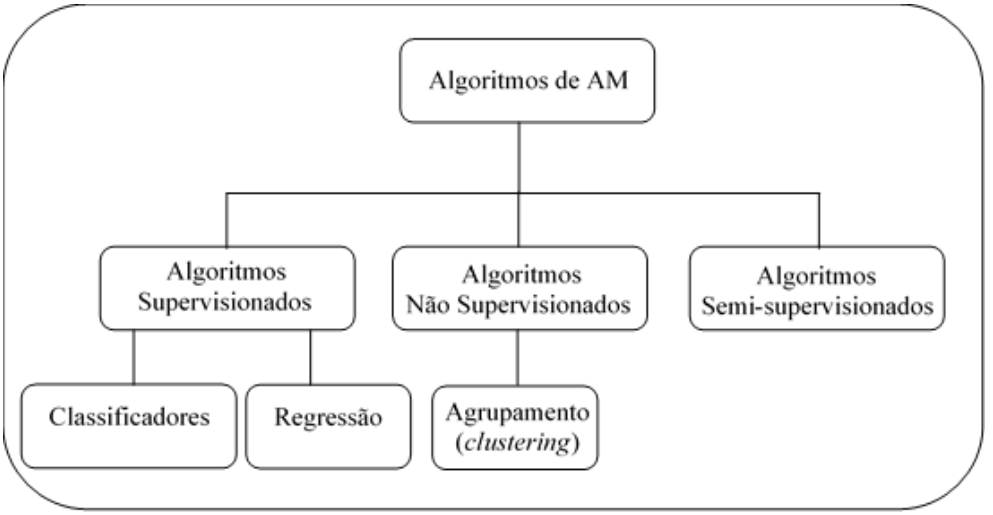


Figura 3.2 Organização simplificada de famílias de algoritmos de AM.

3.3 Etapas de um Projeto de Aprendizado de Máquina

O uso de um algoritmo de AM para a indução da expressão de conceito requer um trabalho prévio de pré-processamento de dados, etapa importante do processo de aprendizado automático. Como informado em [Fayyad *et al.* 1996] tal processo é responsável por identificar e corrigir problemas com os dados originais, com o objetivo de corrigir eventuais problemas que podem estar presentes nos dados. As subseções a seguir explicam de forma sucinta as principais fases desse processo.

3.3.1 Obtenção dos Dados

Uma enorme quantidade de dados é gerada por diferentes fontes todos os dias, dentre as quais podem ser citadas transações financeiras, monitoramento ambiental, dados clínicos e genéticos, navegação em sítios Internet, entre outras. Dados podem também estar disponibilizados em diferentes formatos, tais como textos, gráficos, imagens, vídeos e áudios. Existe uma estimativa de que a cada 20 meses, o volume de dados armazenados nos bancos de dados do mundo dobra [Witten *et al.* 2011].

Diante disso, o trabalho de identificação do conjunto de dados que serão utilizados para atingir um determinado objetivo é fundamental, pois o volume e a qualidade desse

conjunto de dados obtido é que vão determinar a eficiência de predição do modelo a ser induzido, a partir deles, por um algoritmo de AM.

Os dados pertencentes a um conjunto de dados usualmente representam informações sobre objetos, instâncias ou situações. Cada instância de dado do conjunto, via de regra, é descrita por um conjunto de atributos e, muitas vezes, por um atributo diferenciado, chamado de classe, que representa o conceito que tal instância de dado representa.

Como exemplo, considere o conjunto de dados Iris, que é considerado bem simples e que está disponibilizado no UCI Machine Learning Repository [Dua & Graff 2019]. Tal conjunto contém 150 instâncias de dados, cada uma delas descrita por cinco atributos, sendo que quatro deles são relativos às medidas associadas a uma flor íris e o quinto (classe), indica qual tipo de íris a instância de dado em questão representa.

Os quatro primeiros atributos, cujos valores são números reais, são relativos às quatro seguintes dimensões associada a uma particular flor íris: comprimento da sépala, largura da sépala, comprimento da pétala, largura da pétala, respectivamente. As três possíveis classes representam os três possíveis tipos de flores íris *i.e.*, setosa, versicolor e virginica.

O conjunto Iris contém 50 instâncias de dados de cada uma das três classes sendo, pois, um conjunto de dados balanceado, em que os números de instâncias de dados das classes representadas são equilibrados, entre as diferentes classes. A Tabela 3.1 mostra um exemplo de cada uma das classes, extraídas do conjunto com 150 instâncias disponibilizado junto ao UCI Machine Learning Repository [Dua & Graff 2019].

Apesar do grande volume de dados disponíveis, na maioria das vezes é necessária a utilização de técnicas de pré-processamento para prepará-los para serem analisados por métodos estatísticos ou, então, para serem usados como dados de treinamento de algoritmos de AM.

Tabela 3.1 Três instâncias de dados do domínio de conhecimento Iris, cada uma delas representa uma das três classes representadas do domínio Iris [Dua & Graff 2019].

Comprimento da sépala	Largura da sépala	Comprimento da pétala	Largura da pétala	Tipo de íris (classe)
4,6	3,1	1,5	0,2	setosa
6,0	2,2	4,0	1,0	versicolor
7,7	2,8	6,7	2,0	virginica

Usualmente dados chamados de dados brutos *i.e.*, que não sofreram qualquer tratamento após terem sido coletados, podem apresentar inúmeros problemas: valores espúrios, ausência de valores com relação a atributo(s) que o(s) descreve(m), valores de atributos não compatíveis com o atributo que estão representando, instâncias idênticas com classes diferentes (dados contraditórios), etc. A próxima seção aborda esses problemas e possíveis soluções.

3.3.2 Preparação dos Dados

Usualmente empresas, órgãos governamentais e outras organizações armazenam seus dados em mais de uma base ou conjunto de dados; isso significa que os dados são originários de diferentes fontes. Nesse caso, para serem utilizados por um algoritmo de AM, os dados precisam ser agrupados em um único conjunto ou tabela, o que pode levar a inconsistências e redundâncias, principalmente quando o volume de dados é muito grande. Vários outros problemas podem estar presentes também, tais como ruídos, dados incompletos e/ou inconsistentes, ausência de valores ou erros de digitação. A Figura 3.3 mostra alguns dos tipos de problemas que ocorrem em dados, e que devem ser tratados por processos de preparação de dados.

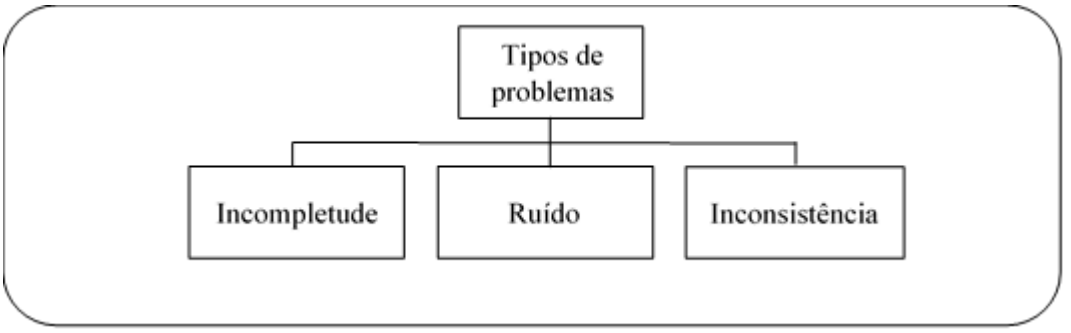


Figura 3.3 Alguns dos tipos de problemas que ocorrem em dados [Castro & Ferrari 2016].

A incompletude ocorre quando da ausência de valores para alguns atributos que são utilizados para a descrição de instâncias do conjunto de dados. Problemas dessa natureza podem ser resolvidos com a utilização de práticas como: (a) remoção das instâncias que têm atributos com valores ausentes, (b) preenchimento dos valores faltantes com a média ou moda dos valores do mesmo atributo nas demais instâncias do conjunto, (c)

preenchimento com a mediana dos valores do mesmo atributo nas demais instâncias do conjunto, (d) uso de algoritmos de AM que lidam internamente com valores ausentes.

Ruído pode ser definido como uma variância ou erro aleatório no valor gerado ou medido associado a um atributo [Han *et al.* 2000] *i.e.*, apresentam valores diferentes do esperado. A detecção de dados com ruídos é a principal dificuldade no processo de preparação dos dados, pois tais ruídos podem ser causados de diversas maneiras, tais como: problemas nos equipamentos que realizam a coleta, transmissão e armazenamento dos dados ou causados por seres humanos no momento do preenchimento ou entrada dos dados.

Várias técnicas de pré-processamento podem ser aplicadas na detecção e remoção de ruídos. Os autores em [Barnett e Lewis 1994] explicam que em estatística é comum resolver esse problema por meio de técnicas baseadas em distribuição. Já os autores em [Barnett e Lewis 1994] [Nuts e Rousseeuw 1996] também destacam que ruídos podem ser detectados por técnicas que utilizam o conceito de profundidade, organizando os dados em camadas. Na literatura podem ser encontrados trabalhos que investigam outras técnicas tais como as associadas a: (a) agrupamento de dados, (b) técnicas baseadas em distâncias, (d) regressão ou classificação [Xu & Wunsch 2005] [Berkhin 2006].

Problemas de inconsistência normalmente se referem a valores que se encontram fora do domínio do atributo ou, então, apresentam distorções em relação a valores que comparecem nas descrições de outras instâncias do mesmo conjunto. A existência de inconsistências influencia na validade e utilidade do conjunto de dados, e detectar sua existência pode ser uma tarefa difícil. Diferentes unidades de medidas ou notação, como é o caso de peso apresentado em quilos (kg) ou em libras (£), e distâncias apresentadas em metros ou em quilômetros, são exemplos comuns de inconsistências [Castro & Ferrari 2016]. Uma das formas para se resolver problemas de inconsistência é a realização de uma análise auxiliada por rotinas específicas que verificam, por exemplo, se os valores de todos os atributos pertencem a domínios específicos, conhecidos *a priori*.

Para todas essas possíveis situações, em se tratando de dados reais, existem técnicas de pré-processamento capazes de auxiliar no tratamento dos dados brutos; essa etapa de pré-processamento deve ser feita de maneira estruturada e cuidadosa. Ao aplicar uma

dessas técnicas deve-se entender sua função e seu efeito sobre o conjunto de dados. As principais tarefas associadas a um processo de pré-processamento estão mostradas na Figura 3.4, como sugerido em [Castro & Ferrari 2016] e brevemente descritas na sequência.

- **Limpeza**: técnica para imputação de valores ausentes de atributos, remoção de ruídos e correção de inconsistências;
- **Integração**: visa unir dados de múltiplas fontes em um único local, como em um dispositivo de armazenamento de dados;
- **Redução**: promove uma representação reduzida do conjunto de dados, que possui um volume menor, porém que mantém a integridade dos dados originais. As estratégias de redução incluem reduzir a dimensionalidade e/ou a quantidade dos dados reduzindo a dimensão do conjunto de dados;
- **Transformação**: visa padronizar e deixar os dados em um formato passível de utilização pelos muitos algoritmos de AM;
- **Discretização**: reduzir o número de valores de um atributo contínuo, divisão da amplitude dos atributos em intervalos iguais. Os rótulos dos intervalos substituem os valores originais do atributo.

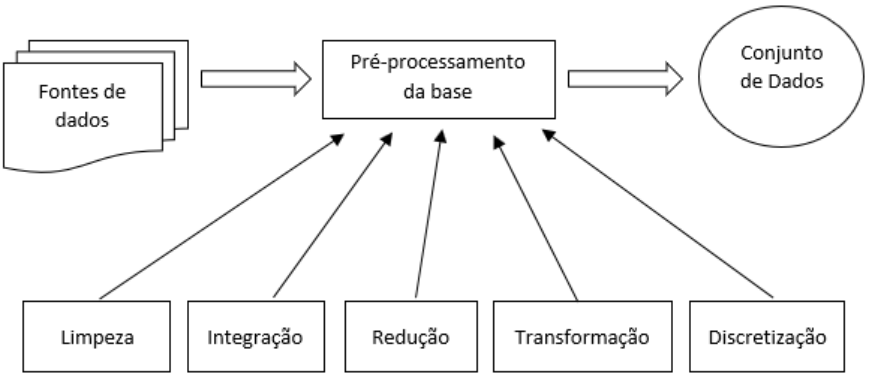


Figura 3.4 - Etapas do processo de preparação do conjunto de dados.

3.3.3 Treinamento e Teste do Modelo

Como comentado anteriormente, uma das muitas maneiras de avaliar o poder preditivo de um modelo induzido a partir de um conjunto de instâncias de dados (conjunto de treinamento), é utilizar um conjunto de dados referenciado como conjunto de teste. Os autores Arlot e Celise, em [Arlot & Celise 2010], explicam um processo de validação conhecido como validação cruzada (*cross-validation*), que é um método de validação amplamente adotado pela comunidade de AM para a avaliação de modelos

preditivos. O método de validação cruzada pode ser considerado um processo de reamostragem dos dados em que os treinamentos e testes parciais são repetidos várias vezes de forma sistemática, tendo como objetivo avaliar o desempenho do modelo para um novo conjunto de dados, permitindo selecionar o modelo preditivo com menor estimativa de erro.

A validação cruzada particiona o conjunto de dados em *k* subconjuntos de tamanhos aproximados, onde *k*-1 são utilizados como conjuntos de treinamento e o subconjunto restante como teste. Esse processo de treinamento e teste é repetido *k* vezes para cada modelo preditivo a ser validado, dessa forma cada um dos *k* subconjuntos é usado uma vez como conjunto de teste. A Figura 3.5 ilustra um processo de 4-validação cruzada (*i.e.*, *k*=4) para um modelo preditivo.

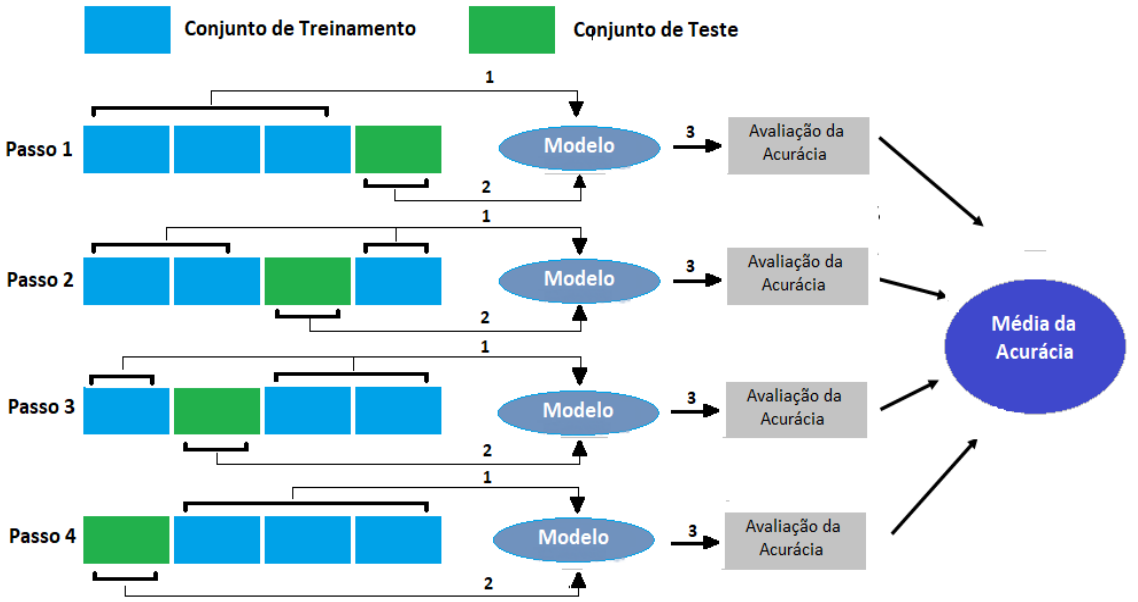


Figura 3.5 - Processo de *k*-validação cruzada (*k*=4), em que a ordem dos processos envolvidos é representada por: (1) Indução, (2) Teste do modelo, (3) Resultado da avaliação da acurácia.

3.4 Aprendizado Não-supervisionado e Algoritmos de Agrupamento

3.4.1 Considerações Iniciais

No aprendizado supervisionado cada instância de dado do conjunto de treinamento é descrita por um conjunto de atributos e por um atributo associado chamado de classe. A classe de cada instância de dado é normalmente determinada *a priori* por uma fonte externa, que pode ser um especialista humano da área de conhecimento em questão.

Porém, existem casos no mundo real em que a classe à qual cada instância pertence não é conhecida, o que pode ser consequência de (1) inexistência de um especialista humano com conhecimento suficiente para estabelecer a classe de cada instância; (2) do processo para determinação da classe ser extremamente custoso e difícil de ser aplicado, (3) da coleta de dados ter acontecido muito tempo atrás e a informação da classe não ter sido considerada na ocasião. Nesses casos os algoritmos de aprendizado não-supervisionado devem ser considerados.

3.4.2 Algoritmos de Agrupamento

Como comentado na Seção 3.3, diferentemente do método supervisionado, o método não-supervisionado não se utiliza da classe associada as instâncias de dados durante o processo de treinamento do algoritmo, uma vez que essa informação pode não estar incorporada ao conjunto de treinamento. A estratégia utilizada por algoritmos de aprendizado não-supervisionados é o particionamento de um conjunto de instâncias de dados, em grupos apropriados, com base em semelhanças (ou diferenças) entre as instâncias.

De forma simplista, algoritmos de agrupamento podem ser caracterizados como processos que particionam um conjunto de dados X em grupos de instâncias de dados, de acordo com algum critério, de tal forma que cada grupo contenha instâncias de dados com características semelhantes entre si e características distintas em relação às instâncias dos demais grupos, ou seja, particiona todo o conjunto de dados em grupos (*clusters*) em que a similaridade entre as instâncias pertencentes a cada um dos grupos é maximizada e a similaridade entre instâncias pertencentes a grupos diferentes é minimizada.

Formalmente, considere um conjunto X com N instâncias de dados, notado por $X = \{I_1, I_2, ..., I_N\}$, e que cada instância de dados é descrita por M atributos notados por $A_1, A_2, ..., A_M$.

Atributos podem assumir valores quantitativos em intervalos contínuos ou em um conjunto finito discreto, como por exemplo, no conjunto {0,1} ou, então, valores qualitativos (categóricos: nominal ou ordinal). Como apontam vários trabalhos, a caracterização do tipo de atributo direciona o processo de seleção da medida de

similaridade que será utilizada [Gowda & Diday 1992], [Jain *et al* 1999], [Kaufman & Rousseeuw 2005].

Considerando o conjunto de instâncias X e um número inteiro K, um agrupamento AG de X é definido com uma partição de X em k grupos (*clusters*), G_1, G_2, ..., G_k ou seja, AG = {G_1, G_2, ..., G_k}. De acordo com a definição matemática de partição de um conjunto, as seguintes condições devem, pois, ser satisfeitas:

(1) $G_i \neq \emptyset$, $\forall$ i = 1, ..., k (cada um dos grupos é não vazio)

(2) A união de todos os grupos recompõe o conjunto inicial, ou seja:

$$X = \bigcup_{i=1}^{k} G_i$$

(3) $G_i \cap G_j = \emptyset$, i ≠ j e i, j = 1, ..., k. (grupos são dois a dois disjuntos)

Como sugerido em [Xu & Wunsch 2005], algoritmos de agrupamento podem ser divididos em hierárquicos e particionais (não-hierárquicos).

Os algoritmos-hierárquicos realizam uma decomposição hierárquica das instâncias de dados iniciais e, de acordo com o processo de decomposição realizado, podem ser classificados em (a) aglomerativos e (b) divisivos.

Algoritmos aglomerativos: produzem uma sequência de agrupamentos com um número decrescente de grupos. Como descrito em [Theodoridis & Koutroumbas 2009], o agrupamento produzido em um passo p é baseado no agrupamento produzido no passo p−1. Para a indução do agrupamento no passo p, dois grupos do agrupamento produzido no passo p−1 são unidos, o que faz com que o agrupamento produzido no passo p tenha um grupo a menos do que o número de grupos do agrupamento do passo p−1. O processo é repetido até que o agrupamento final tenha apenas um grupo, constituído do conjunto inicial das instâncias a serem agrupadas.

Já um algoritmo de agrupamento divisivo considera o conjunto inicial de dados como um único grupo e, em um processo iterativo escolhe, a cada iteração, um grupo e o divide em dois, induzindo assim, um novo agrupamento, com um grupo a mais do que o número de grupos do agrupamento anterior. O processo continua até que uma condição de parada seja satisfeita ou, então, o critério de parada *default*, em que cada grupo do agrupamento tem apenas um elemento.

Algoritmos de agrupamento não-hierárquicos usualmente particionam um conjunto com N instâncias de dados, em um conjunto de k grupos (*clusters*), em que K é um parâmetro fornecido ao algoritmo, como informado anteriormente. O conjunto formado pelos K grupos é referenciado como agrupamento, notado por AG = {G_1, G_2, ..., G_k}. Algoritmos não-hierárquicos induzem uma partição inicial no conjunto de instâncias de dados considerado e, usando uma estratégia de realocação de instâncias, realiza o deslocamento das instâncias de seu grupo para outro grupo no qual as instâncias estariam melhor atribuídas, de acordo com o critério adotado.

Esse processo de realocação é repetido até que não se obtenha mais nenhuma melhoria com os deslocamentos.

O algoritmo conhecido como k-Means [MacQueen 1967] é um dos algoritmos mais populares dentre os algoritmos particionais, e foi o algoritmo de agrupamento utilizado em alguns dos experimentos descritos na Seção 6.5 desta dissertação. A Seção 3.4.5 apresenta o pseudocódigo do k-Means, com breves comentários sobre seus passos, e a Seção 3.4.6 um exemplo do seu uso em um pequeno conjunto de instâncias de dados.

3.4.3 Medidas de Similaridade e Distâncias

Lembrando o que foi informado na Seção 3.4.2, métodos de agrupamento têm por objetivo, a partir de um conjunto inicial de instâncias de dados, agrupar instâncias similares entre si e dissimilares a instâncias pertencentes a outros grupos. Desta forma, o conceito de similaridade é parte integrante de um processo de agrupamento, sendo a definição de uma medida de similaridade entre duas instâncias de dados, essencial a qualquer procedimento que implemente um algoritmo de agrupamento. Devido a variedade de tipos de atributos e de suas respectivas unidades de medida, a medida de distância deve ser cuidadosamente escolhida [Jain *et al.* 1999].

A qualidade dos resultados obtidos por métodos de agrupamento depende tanto da medida de similaridade ou dissimilaridade usada pelo método, quanto da forma como é implementada. Como comentado anteriormente, um método de agrupamento é considerado bom quando induz grupos com alta similaridade intragrupos e baixa similaridade intergrupos.

Nas referências [Xu & Wunsch 2005] e [Jain & Dubes 1988] é informado que a métrica de dissimilaridade mais comum para mensurar a distância de atributos contínuos

é a distância Euclidiana. A distância Euclidiana entre duas instâncias de dados M-dimensionais, I_a e I_b, é apresentada na Equação (3.1).

$$d(I_a, I_b) = \sqrt{\sum_{j=1}^{M}(I_{a_j} - I_{b_j})^2} \quad (3.1)$$

Algumas outras medidas de similaridade e distância para atributos quantitativos, tais como a distância *Manhattan*, distância de *Mahalanobis* dentre outras podem ser encontradas com mais detalhes em [Jain & Dubes 1988], [Duda *et al.* 2000], [Xu & Wunsch 2005], [Theodoridis & Koutroumbas 2009].

3.4.4 Um Exemplo de Agrupamento

A Tabela 3.2 apresenta a descrição de um conjunto de dez pessoas, por meio dos valores de um conjunto de seis atributos. A identificação associada a cada um dos seis atributos são: lentes, cabelo, barba, cor, estatura, sexo. Cada linha da tabela descreve uma pessoa em particular e é considerada uma instância do conjunto. A primeira coluna está na tabela apenas como nomeação de cada uma das instâncias, com o objetivo de facilitar a referência a qualquer uma delas.

Se apenas o atributo lentes for utilizado para organizar o grupo de dez pessoas, o agrupamento formado (de maneira automática ou manual) AG_{lentes}, contém dois grupos: o grupo das pessoas que usam lentes *i.e.*, $G_1 = \{I_1, I_5, I_6, I_9\}$ e o grupo das pessoas que não usam lentes *i.e.*, $G_2 = \{I_2, I_3, I_4, I_7, I_8, I_{10}\}$ e, portanto, o agrupamento formado é $AG_{lentes} = \{G_1, G_2\} = \{\{I_1, I_5, I_6, I_9\}, \{I_2, I_3, I_4, I_7, I_8, I_{10}\}\}$.

Se o atributo utilizado for cor, o conjunto das dez pessoas será dividido em três grupos, $G_1 = \{I_1, I_3, I_5, I_8, I_9\}$, $G_2 = \{I_2, I_7\}$, $G_3 = \{I_4, I_6, I_{10}\}$, resultando no agrupamento $AG_{cor} = \{G_1, G_2, G_3\} = \{\{I_1, I_3, I_5, I_8, I_9\}, \{I_2, I_7\}, \{I_4, I_6, I_{10}\}\}$.

Utilizando a composição dos atributos barba e sexo como critério para a criação do agrupamento, os grupos formados seriam $G_1 = \{I_1, I_7, I_8\}$, $G_2 = \{I_2, I_5, I_6, I_9\}$, e $G_3 = \{I_3, I_4, I_{10}\}$ e o agrupamento formado seria $AG_{sexo\&barba} = \{\{I_1, I_7, I_8\}, G_2 = \{I_2, I_5, I_6, I_9\}, \{I_3, I_4, I_{10}\}\}$.

Tabela 3.2 – Conjunto de dez pessoas descritas pelos valores os seis atributos que nomeiam cada uma das colunas, a partir da segunda, em que (S) Sim (N) Não; (B) Branca (P) Parda (N) Negra; (A) Alto (B) Baixo; (M) Masculino (F) Feminino.

Pesso	Lente	Cabe	Barb	Cor	Estat	Sexo
I1	S	S	N	B	A	M
I2	N	S	N	P	A	F
I3	N	N	S	B	A	M
I4	N	S	S	N	B	M
I5	S	S	N	B	B	F
I6	S	N	N	N	A	F
I7	N	N	N	P	B	M
I8	N	S	N	B	A	M
I9	S	N	N	B	B	F
I10	N	S	S	N	A	M

A Figura 3.6 ilustra o agrupamento AG_{lentes}, a partir do conjunto de instâncias de dados descritos na Tabela 3.2, contendo dois grupos: o grupo das pessoas que usam lentes $G_1 = \{I_1, I_5, I_6, I_9\}$ e o grupo das pessoas que não usam lentes $G_2 = \{I_2, I_3, I_4, I_7, I_8, I_{10}\}$ formando o agrupamento $AG_{lentes} = \{G_1, G_2\} = \{\{I_1, I_5, I_6, I_9\}, \{I_2, I_3, I_4, I_7, I_8, I_{10}\}\}$.

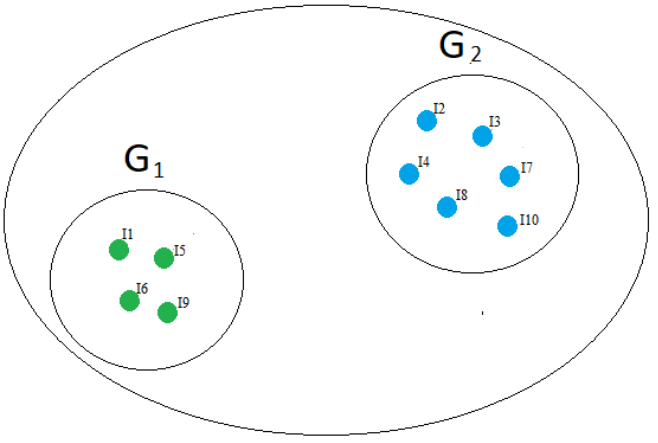


Figura 3.6 – Exemplo de agrupamento $AG_{lentes} = \{G_1, G_2\}$, em que $|G_1| = 4$ e $|G_2| = 6$, sendo $G_1 = \{I_1, I_5, I_6, I_9\}$ e $G_2 = \{I_2, I_3, I_4, I_7, I_8, I_{10}\}$.

3.4.5 O Algoritmo k-Means

O k-Means é o algoritmo de agrupamento de dados mais conhecido e utilizado entre todos os algoritmos de agrupamento existentes devido à sua simplicidade de implementação [Jain 2010] [Theodoridis & Koutroumbas 2009].

Como previamente comentado, o k-Means pode ser descrito como um algoritmo que particiona um conjunto de N instâncias de dados em k grupos, em que k é um parâmetro fornecido ao algoritmo, de forma que a similaridade intragrupo seja alta e a similaridade intergrupo seja baixa. A similaridade intragrupo é geralmente avaliada considerando o valor médio dos valores de atributos das instâncias em um grupo, definido como centroide, ou seja, cada instância vai pertencer ao grupo cujo centroide lhe for mais próximo.

O algoritmo k-Means utiliza uma técnica de refinamento iterativo onde os centroides iniciais de cada grupo são definidos aleatoriamente ou selecionados aleatoriamente a partir de algumas instâncias do próprio conjunto de dados. Após definidos os centroides iniciais, o algoritmo calcula a distância (geralmente a distância Euclidiana) entre as instâncias e cada um dos k centroides, e atribui cada instância ao grupo cujo centroide esteja mais próximo da instância. A escolha dos centroides iniciais é de extrema relevância, pois poderá interferir no resultado final do agrupamento. Em seguida, novos centroides são calculados, o que pode promover um reposicionamento dos centroides e uma realocação das instâncias de dados entre os grupos, sempre considerando o grupo cujo centroide é o mais próximo da instância. Esse processo é finalizado quando não há mais alterações nos centroides e nenhuma ou poucas mudanças no posicionamento das instâncias em relação aos grupos. Seguem os passos utilizados para realizar o agrupamento [Jain *et al.* 1999] [Kanungo *et al.* 2002] [Poteras *et al.* 2014].

(1) Selecionar aleatoriamente k instâncias de dados como centroides iniciais para os grupos. O valor de k deve ser informado pelo usuário como indicação da quantidade de grupos no agrupamento;
(2) Associar cada uma das instâncias de dados de entrada do algoritmo ao centroide mais próximo, onde cada conjunto de instâncias de dados associado a um centroide forma um grupo do agrupamento;
(3) Atualizar os centroides pertencentes a cada grupo para representar a média das instâncias que pertencem ao grupo;
(4) Repetir os passos 2 e 3 até que nenhuma instância mude de grupo, (ou os centroides não se alterem).

O algoritmo k-Means busca, a cada iteração, diminuir a distância entre as instâncias de dados e os centroides a elas associados, até que a distância entre as instâncias e seu

respectivo centroide resulte em um grupo o mais compacto possível (mínimo local possível) e a distância das instâncias com aquelas associadas os demais grupos sejam as maiores possível (maior local possível) [Kanungo *et al.* 2002].

O Algoritmo 3.1 apresenta um pseudocódigo simplificado do algoritmo k-Means.

```
procedure k-Means(X, k, AG)
   Entrada: X = {x1, x2, ..., xn}     %conjunto com N instâncias de dados a serem agrupadas
            k                          % número de grupos a serem criados
   Saída: AG = {G1, G2, ..., GK}     %agrupamento formado por k grupos
begin
   % Inicialização
   % cada grupo é definido apenas pelo centroide
   (1) escolha arbitraria de k instâncias do conjunto AG, como centroides dos grupos G1,G2,...GK
   % Indução do agrupamento AG
   (2) repeat
      (3) (re)atribuir cada instância x ϵ Gi (i=1, ..., N) ao grupo cujo centroide que lhe seja mais próximo;
      (4) atualizar os centroides de cada grupo, como a média dos valores das suas instâncias
   (5) until nenhuma alteração aconteça.
end.
return AG = {G1, G2, ..., GK}
end_procedure
```

Algoritmo 3.1 Pseudocódigo simplificado do algoritmo k-Means [MacQueen 1967].

3.4.6 Um Exemplo de Uso do Algoritmo k-Means

A Tabela 3.3 representa um conjunto de dados contendo 40 instâncias descritas por três atributos numéricos: Id, Renda e Score, cujos valores são fictícios e representam dados sobre clientes de *shopping*. Cada linha da tabela descreve um cliente em particular e é considerada uma instância do conjunto. A primeira coluna está na tabela para identificar cada uma das instâncias, com o objetivo de facilitar a referência a qualquer uma delas. Os valores do atributo Renda descrevem a renda anual de cada um dos clientes em R$(mil). O atributo *Score* descreve a pontuação obtida por cada cliente de acordo com o total anual de compras [0-100].

Tabela 3.3 – Dados fictícios referentes a clientes de *shopping*. A renda está representada em R$ (mil) e o Score [1-100].

Id	Renda	*Score*	Id	Renda	*Score*	Id	Renda	*Score*	Id	Renda	*Score*
1	15	39	11	19	14	21	24	35	31	30	4
2	15	31	12	19	99	22	24	73	32	30	73
3	16	6	13	20	15	23	25	5	33	33	4
4	16	77	14	20	77	24	25	73	34	33	92
5	17	40	15	20	13	25	28	14	35	33	14
6	17	76	16	20	79	26	28	82	36	33	81
7	18	6	17	21	35	27	28	32	37	34	17
8	18	94	18	21	66	28	28	61	38	34	73
9	19	3	19	23	20	29	29	31	39	37	26
10	19	72	20	23	98	30	29	87	40	37	75

O código do algoritmo k-Means utilizado nesse exemplo foi implementado na linguagem de programação Python 3.6 utilizando a biblioteca scikti-learn e, para a geração dos gráficos, foi utilizada a biblioteca matplotlib.

A Figura 3.7 ilustra as 40 instâncias, representadas como pontos, do conjunto em um gráfico de duas dimensões.

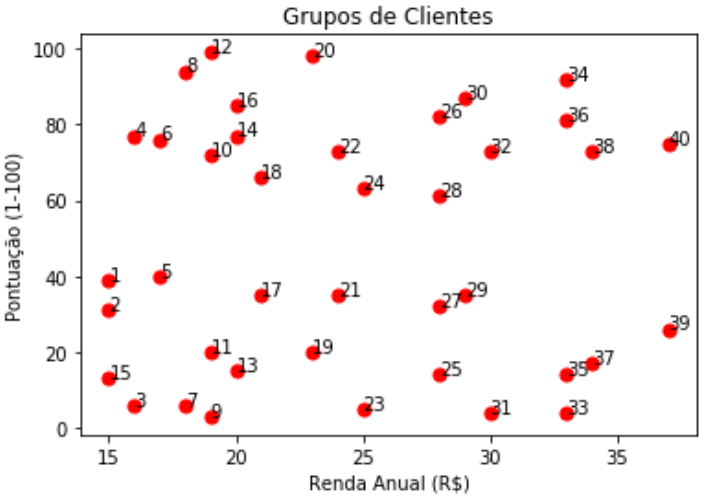


Figura 3.7 Representação das 40 instâncias do conjunto de dados da Tabela 3.3.

Os testes foram realizados com o valor inicial de k=3. Como pode ser visto, seguindo a descrição do k-Means, o algoritmo escolheu aleatoriamente k instâncias do conjunto de dados como centroides iniciais. A Figura 3.8 ilustra a escolha aleatória de k=3 instâncias do conjunto de dados como centroides iniciais, representados como estrelas amarelas no gráfico.

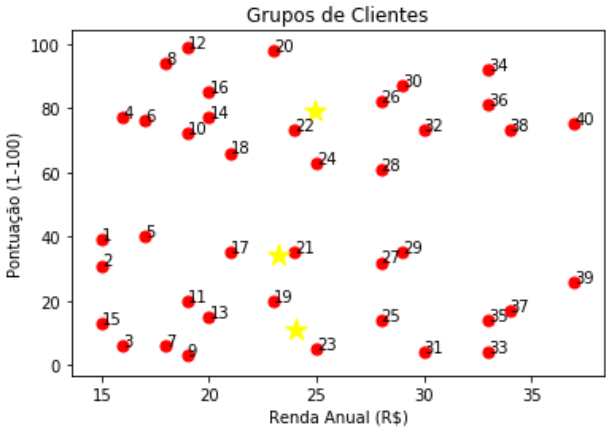


Figura 3.8 – Representação dos 3 centroides iniciais escolhidos, identificados como pontos amarelos na figura.

Após a escolha dos centroides, cada instância de dados é atribuída ao grupo cujo centroide estiver mais próximo (maior similaridade) da instância, assumindo, assim, cores diferentes para cada grupo. A Figura 3.9 ilustra os três centroides em suas posições iniciais e as instâncias que lhes foram atribuídas. As instâncias atribuídas a um centroide formam um grupo. Distingue-se pela cor (azul, verde e vermelho) o grupo ao qual pertencem as instâncias, mantendo-se os centroides na cor amarela.

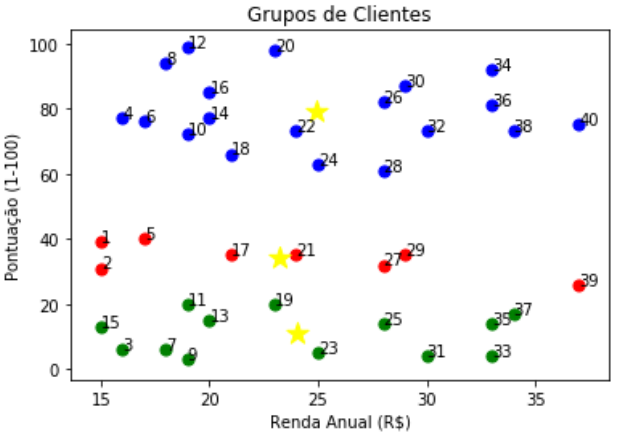


Figura 3.9 – Representação da atribuição das instâncias aos centroides mais próximos.

O Algoritmos k-Means continua atualizando os centroides pertencentes a cada grupo de acordo com a média das instâncias pertencentes ao grupo, até que nenhuma, ou poucas instâncias mudem de grupo, ou até que os centroides não sejam mais alterados**.**

A Figura 3.10 mostra o agrupamento final, no qual as instâncias permanecem nos grupos em que estavam e não há recálculo de centroides, encerrando assim a execução

do algoritmo, em que o agrupamento induzido é representado pelo conjunto dos últimos centroides obtidos.

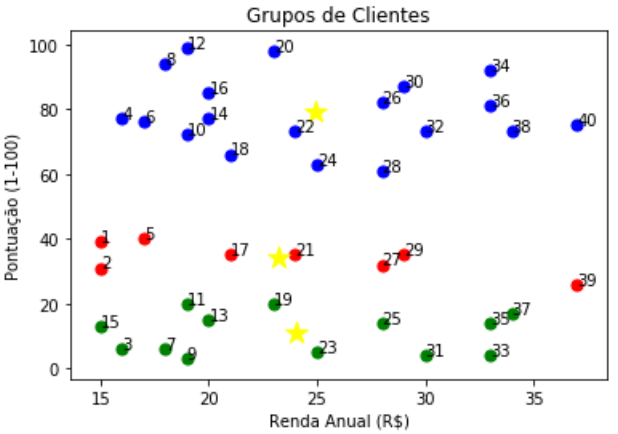


Figura 3.10 – Representação final do agrupamento.

3.5 Índices de Validação de Agrupamento

A validação é um processo essencial para a avaliação do resultado obtido a partir de algoritmos de agrupamento e sua utilização tem como intuito mostrar o quão representativo é o agrupamento induzido pelo algoritmo.

A avaliação de um agrupamento depende dos objetivos e do contexto da análise. Por exemplo, realizando uma análise exploratória em uma base de dados de imóveis, pode-se identificar perfis de moradores de determinado bairro ou município através de instâncias pertencentes ao mesmo grupo. Também é possível avaliar a saída de um algoritmo quanto à qualidade do agrupamento induzido, comparando a estrutura dos grupos encontrados com uma estrutura de agrupamento conhecida *a priori*, como também tentar determinar se a estrutura encontrada pelo algoritmo é apropriada aos dados.

Agrupamentos induzidos por um algoritmo podem ser avaliados quantitativamente por meio de índices de avaliação. Segundo [Halkidi *et al.* 2001] existem dois critérios para a avaliação da qualidade de um agrupamento: (a) compactação e (b) separação.

(a) Compactação: as instâncias pertencentes a cada grupo do agrupamento devem estar o mais próximo possível das instâncias do mesmo grupo.

(b) Separação: os grupos devem estar o mais distante possível uns dos outros.

De acordo com alguns estudos, os índices de avaliação podem pertencer a duas categorias principais: (a) avaliação externa, (b) avaliação interna [Theodoridis & Koutroubas 2009] [Kóvacs *et al.* 2001] [Halkidi *et al.* 2001].

(a) Avaliação externa: medem a semelhança entre o agrupamento gerado pelo algoritmo e um agrupamento de referência pré-definido, que reflete a intuição do especialista sobre qual seria a estrutura ideal. O cálculo dessas medidas requer o conhecimento prévio do grupo ao qual cada instância pertence.
(b) Avaliação interna: medem a qualidade de um agrupamento utilizando apenas informações inerentes às instâncias do agrupamento, baseando-se em medidas de similaridade e avaliando as distâncias intragrupos e/ou intergrupos.

Durante as experimentações envolvendo o algoritmo k-Means, foram utilizados o índice de avaliação interna conhecido como Silhouette [Rousseeuw 1987] e o índice de avaliação externa, especificamente o índice Rand [Rand 1971]. Uma breve descrição de cada um dos dois índices é apresentada a seguir.

3.5.1 Índice Silhouette

O índice Silhouette foi proposto por Rousseeuw em [Rousseeuw 1987], e sua principal contribuição está na interpretação e avaliação dos resultados da análise do agrupamento, observando a semelhança entre as instâncias de um mesmo grupo comparada com as instâncias dos outros grupos, ou seja, avalia o quão bem agrupados estão as instâncias ao grupo ao qual pertencem e a separação entre os grupos. O índice Silhouette é apropriado para medidas de proximidade que utilizam escala proporcional (*ratio scale*) como é o caso da distância Euclidiana. O cálculo do índice Silhouette (SIL) de um agrupamento AG é dado pela Equação (3.2).

$$\mathrm{SIL(AG)} = \frac{1}{k}\Sigma_i \left\{ \frac{1}{|G_i|} \Sigma_{P_i \in G_i} \frac{b(P_i) - a(P_i)}{\max[a(P_i), b(P_i)]} \right\} \quad (3.2)$$

em que k é o número de grupos; $|G_i|$ é o número de instâncias no i-ésimo grupo G_i; $a(P_i)$ é a distância média entre a instância P_i e as instâncias pertencentes ao mesmo grupo que P_i; e $b(P_i)$ é a menor distância média entre a instância P_i e as instâncias pertencentes a cada um dos grupos aos quais P_i não pertence. A função $a(P_i)$ está definida na Equação (3.3) e a função $b(P_i)$ está definida na Equação (3.4).

$$a(P_i) = \frac{1}{|G_i| - 1} \sum_{P_j \in G_i} dist(P_i, P_j) \quad (3.3)$$

$$b(P_i) = \min_{j, j \neq i} [\frac{1}{|G_j|} \sum_{p_j \in G_j} dist(P_i, P_j)] \quad (3.4)$$

O valor do índice Silhouette varia no intervalo de -1 a +1 e, quanto mais próximo de 1 for o valor do índice obtido para um determinado agrupamento, melhor é o agrupamento.

3.5.2 Índice Rand

O índice de Rand, proposto por Rand em [Rand 1971], é um dos índices mais conhecidos e utilizados para avaliação externa, podendo ser entendido como uma medida de similaridade entre dois agrupamentos. O índice Rand pode ser representado formalmente da seguinte forma descrita a seguir.

Considere X um conjunto de dados a ser agrupado. Considere dois agrupamentos de X notados como G = {G_1, G_2, ... G_{NG}} e P = {P_1, P_2, ... P_{NP}}. Para determinar o índice Rand associado aos dois agrupamentos, os valores a, b, c e d devem ser inicialmente calculados e, com base neles, determinar o valor do índice Rand associado. A forma de cálculo dos valores assim como a equação para determinar o índice Rand estão descritos a seguir.

- a: número de pares de instâncias de dados de X que estão em um mesmo grupo no agrupamento G e no mesmo grupo no agrupamento P;
- b: número de pares de instâncias de dados de X que estão em grupos diferentes no agrupamento G e em grupos diferentes no agrupamento P;
- c: número de pares de instâncias de dados de X que estão em um mesmo grupo no agrupamento G e em grupos diferentes no agrupamento P e
- d: número de pares de instâncias de dados de X que estão em grupos diferentes no agrupamento G e no mesmo grupo no agrupamento P.

O índice Rand é dado pela Equação (3.5).

$$R = \frac{a+b}{a+b+c+d} \quad (3.5)$$

O valor obtido pelo índice Rand varia no intervalo [0,1], onde 0 indica que os agrupamentos são totalmente diferentes entre si, e 1 indica que os agrupamentos são iguais.

3.6 Considerações Finais

Este capítulo teve por objetivo contextualizar a área de AM como uma subárea da área de Inteligência Artificial, apresentando algumas definições e conceitos básicos relevantes ao trabalho realizado. Também foram apresentadas características de algoritmos de AM supervisionados e não supervisionados, ilustrados por meio de exemplos, considerando que ambos os tipos de algoritmos foram usados nos experimentos que serão apresentados no Capítulo 6.

Capítulo 4

Escolha e Coleta de Dados

4.1 Considerações Iniciais

Como comentado anteriormente, o governo brasileiro disponibiliza um grande volume de dados relativos a diversos domínios de atuação governamental, de forma aberta *i.e.*, cidadãos interessados em tais dados podem utilizá-los para análises, projeções, estimativas, etc.

Considerando a disponibilidade de utilização desses dados para o trabalho de pesquisa cogitado, na fase inicial do trabalho foram feitas pesquisas bibliográficas e buscas na Web, com o objetivo de identificar *sites* com dados relevantes ao trabalho e, também, que tivessem um alto grau de confiabilidade.

Nas seções que seguem são abordadas as várias atividades realizadas, com vista à execução da coleta de dados. A Seção 4.2 apresenta brevemente o PNUD, que é um dos principais fomentadores dos dados produzidos e de índices para a caracterização de desenvolvimento de países, utilizados na pesquisa em andamento. A Seção 4.3 tem por objetivo a apresentação do Atlas do Desenvolvimento Humano no Brasil [AtlasBrasil 2020], um repositório de dados utilizados neste trabalho de pesquisa. A Seção 4.4 descreve em detalhes os três sub-índices utilizados, quando do estabelecimento do Índice de Desenvolvimento Humano Municipal (IDHM) e, finalmente, a Seção 4.5 aborda os resultados das eleições de 2018, uma vez que se trata do domínio de dados utilizado na pesquisa.

4.2 Programa das Nações Unidas para o Desenvolvimento (PNUD)

O PNUD [PNUD 2020] é uma rede global de desenvolvimento da Organização das Nações Unidas (ONU). Presentemente o PNUD está estabelecido em 170 países e territórios, com a missão de alinhar seu trabalho às necessidades de cada país, colaborando no desenvolvimento de políticas, habilidades de liderança, capacidades

institucionais, resiliência e, especialmente, erradicação da pobreza e redução de desigualdades e exclusão social.

O PNUD busca promover o desenvolvimento humano, e contemplar os objetivos de desenvolvimento sustentável nos países em que atua, por meio de investimentos em três áreas principais:

(1) desenvolvimento sustentável;

(2) apoio à política democrática e construção da paz e

(3) resiliência ao clima e desastres.

O programa procura também ajudar países a atrair ajuda de maneira efetiva e, em suas atividades, promove igualdade de gênero bem como a proteção dos direitos humanos. A estratégia de atuação do PNUD no Brasil contempla, entre várias outras metas, dentre elas, a de investir em áreas vulneráveis e populações de baixo Índice de Desenvolvimento Humano (IDH).

4.3 Atlas do Desenvolvimento Humano

O repositório de informações, conhecido AtlasBrasil [AtlasBrasil 2020], é uma plataforma de consulta do Índice de Desenvolvimento Humano Municipal (IDHM), relativo a todos os municípios brasileiros, que disponibiliza mais de 200 indicadores relacionados à demografia, educação, renda, trabalho, habitação e vulnerabilidade. Os dados que o Atlas disponibiliza foram extraídos dos Censos de 1991, 2000 e 2010, sendo que o último Censo foi realizado em 2010.

O objetivo primeiro do AtlasBrasil é o de disponibilizar dados concretos relativos a municípios e regiões brasileiras, com vista a promover e unificar dados confiáveis para que, com base neles, possam ser feitas análises e projeções realísticas, com o objetivo de monitoramento do desenvolvimento humano no Brasil.

O AtlasBrasil foi desenvolvido e é mantido pelo PNUD, em colaboração com o Instituto de Pesquisa Econômica Aplicada (IPEA) [IPEA 2020] e a Fundação João Pinheiro (FJP). A IPEA é uma fundação pública federal que é vinculada ao Ministério da Economia e foi criada em 1967. Como informa o site do IPEA, o Instituto divulga, de forma espontânea, dados de interesse coletivo ou geral, com o objetivo de facilitar o

acesso à informação pública, conforme determinam a Lei de Acesso à Informação (Lei no 12.527/2011) e o decreto que a regulamenta no âmbito do Executivo federal (Decreto no. 7.724/2011). Já a Fundação João Pinheiro (FJP) [FJP 2020], criada em 1969, é o órgão oficial de pesquisa em políticas públicas, estatísticas e ensino em administração pública do Governo do Estado de Minas Gerais e está vinculada à Secretaria do Estado de Planejamento e Gestão de Minas Gerais [FJP 2020].

O AtlasBrasil foi planejado e projetado para ser uma ferramenta de fácil utilização e de disponibilização de informações, com o objetivo de facilitar o manuseio dos dados e, assim, incentivar e promover análises. A Figura 4.1 mostra uma das telas da plataforma para consulta de dados.

Figura 4.1 Tela de consulta dos dados disponibilizados no *site* do AtlasBrasil. [AtlasBrasil 2020].

A plataforma disponibiliza os dados, permitindo a seleção de dados de interesse por meio de filtros, e os mantém subdivididos em Estados e Municípios. A Figura 4.2 mostra o processo de seleção dos dados referentes ao IDHM associado a um estado do país, por meio da filtragem de todos os municípios do estado selecionado.

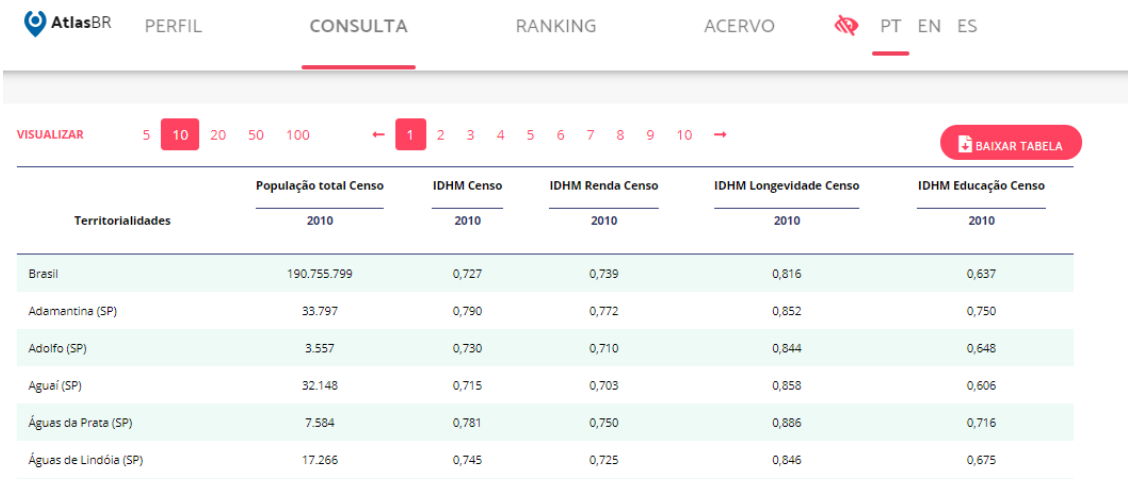
Figura 4.2 Filtragem do IDHM por município [AtlasBrasil 2020].

Além dos IDHM, a plataforma também permite a extração de outra informação relevante para o trabalho de pesquisa, que é a população total de cada um dos municípios selecionados**,** como mostra a Figura 4.3.

Figura 4.3 Filtragem da população total por município [AtlasBrasil 2020].

Além da seleção de dados, a plataforma oferece a opção de exportação dos dados selecionados para diversos formatos, o que possibilita uma integração rápida e fácil com as ferramentas de análise. A Figura 4.4 mostra os dados resultantes da pesquisa, após aplicação dos filtros, com a possibilidade de exportação dos dados formatados como

"CSV" (*Comma Separated Values*), formato comumente utilizado para a integração com as demais ferramentas de análise.



Territorialidades	População total Censo 2010	IDHM Censo 2010	IDHM Renda Censo 2010	IDHM Longevidade Censo 2010	IDHM Educação Censo 2010
Brasil	190.755.799	0,727	0,739	0,816	0,637
Adamantina (SP)	33.797	0,790	0,772	0,852	0,750
Adolfo (SP)	3.557	0,730	0,710	0,844	0,648
Aguaí (SP)	32.148	0,715	0,703	0,858	0,606
Águas da Prata (SP)	7.584	0,781	0,750	0,886	0,716
Águas de Lindóia (SP)	17.266	0,745	0,725	0,846	0,675

Figura 4.4 Resultado da consulta de indicadores por município e exportação para um arquivo csv [AtlasBrasil 2020].

4.4 Índices de Desenvolvimento por Município

Segundo o PNUD, desenvolvimento humano pode ser conceituado como um processo que visa à ampliação das possibilidades de escolha disponibilizadas às pessoas, para que elas tenham liberdade de serem aquilo que desejam ser. Na visão do PNUD, o crescimento econômico e a renda são importantes, mas como meios de desenvolvimento e não como seu fim, sugerindo uma mudança de perspectiva para aquela em que o ser humano é o alvo do desenvolvimento humano.

De acordo com o PNUD, o IDH é uma medida condensada do progresso a longo prazo, em três dimensões básicas do desenvolvimento humano *i.e.*, (1) saúde, (2) educação e (3) renda. O IDH tem por objetivo apresentar uma outra visão em relação ao índice comumente usado, dado pelo Produto Interno Bruto (PIB) *per capita*, que considera apenas a dimensão econômica como fator de desenvolvimento. O IDH tem o objetivo de ser uma média geral e sintética do desenvolvimento humano. É importante ressaltar, entretanto, que o IDH não abrange todos os aspectos que impactam o desenvolvimento humano de países, uma vez que não leva em consideração aspectos tais como: sustentabilidade, democracia, equidade e participação.

O PNUD Brasil, o IPEA e a Fundação João Pinheiro assumiram o desafio de adaptar a metodologia do IDH Global para calcular o IDH Municipal (IDHM) dos 5.565

municípios brasileiros. Esse cálculo foi realizado a partir das informações dos três últimos Censos Demográficos do IBGE – 1991, 2000 e 2010 – e conforme a malha municipal existente em 2010.

O IDHM brasileiro considera as mesmas três dimensões do IDH Global, ou seja (1) saúde, (2) educação e (3) renda. O IDHM, entretanto vai um pouco além, ao adequar a metodologia global ao contexto brasileiro e à disponibilidade de indicadores nacionais, criando assim índices mais adequados para a avaliação dos municípios brasileiros. As subseções a seguir explicam como são calculados cada um dos índices utilizados nesse trabalho.

4.4.1 Sobre o IDH Geral

O PNUD explica que, desde 2010, quando o Relatório de Desenvolvimento Humano completou 20 anos, novas metodologias foram incorporadas ao cálculo do IDH. Atualmente, os três aspectos *i.e.*, (1) saúde, (2) educação e (3) renda, que são considerados para o cálculo do IDH, são avaliados como descrito a seguir.

- saúde: uma vida longa e saudável, que é associada à longevidade, é medida pela expectativa de vida;
- educação: abordada como acesso ao conhecimento, que é medida por:

•• adultos: média do número de anos de educação recebida durante sua vida, por pessoas a partir de 25 anos;

•• crianças na idade de iniciar a vida escolar: expectativa do número de anos de escolaridade, dado pelo número total de anos de escolaridade que uma criança na idade de iniciar a vida escolar pode esperar receber, se os padrões prevalecentes de taxas de matrículas específicas por idade permanecerem os mesmos durante a vida da criança;

- renda: abordada como padrão de vida, é medida pela Renda Nacional Bruta (RNB) *per capita* expressa em poder de paridade de compra (PPC) constante, em dólar, tendo 2005 como ano de referência**.**

Vale ressaltar que o ano de 2005 ficou marcado por uma campanha mundial sem precedentes, dedicada a relegar a pobreza para o passado. Nesse ano foi redigido o

Relatório de Desenvolvimento Humano, reunindo propostas para os próximos 10 anos, com enfoque na cooperação entre os países ricos de transformar as promessas em ações concretas para ajudar a erradicar a pobreza extrema do nosso mundo. Os três pilares de cooperação foram: a) ajuda ao desenvolvimento; b) comércio internacional; c) segurança [PNUD 2020b].

As subseções que seguem apresentam os três subíndices que são considerados para o cálculo do IDHM. Como um deles *i.e.*, o cálculo do IDHM Educação, bem como o cálculo do IDHM fazem uso do conceito de média geométrica, a Subseção 4.2.2 relembra o conceito de média geométrica e discute o objetivo de sua utilização em ambos, IDHM Educação e IDHM.

4.4.2 Justificativas para o Uso de Média Geométrica no IDHM

A média geométrica (Mg) dos N dados de um conjunto $\{x_1, x_2, ..., x_N\}$ é calculada como a raiz N-ésima da multiplicação desses N dados. Assim como a média aritmética, a média geométrica também é uma medida de tendência central e pode ser classificada em simples e ponderada. A média geométrica simples é formalmente representada pela expressão (4.1).

$$M_g = \sqrt[N]{x_1 \times x_2 \times x_3 \times \ldots \times x_N} \tag{4.1}$$

Já para o cálculo da média geométrica ponderada (M_{gp}) considere que cada elemento do conjunto de dados $\{x_1, x_2, ..., x_N\}$ têm um peso associado, notado por pi, i = 1, ..., N. A média geométrica desses N dados ponderados pelos respectivos pesos é dada pela expressão (4.2), em que $K = \sum_{i=1}^{N} pi$.

$$M_{gp} = \sqrt[K]{x_1^{p1} \times x_2^{p2} \times x_3^{p3} \times \ldots \times x_N^{pN}} \tag{4.2}$$

De acordo com o AtlasBrasil, embora o IDHM tenha sido inspirado pelo Índice de Desenvolvimento Humano (IDH) global, foram necessários alguns ajustes para melhor adequá-lo à realidade brasileira, adaptando-o às bases de dados do Censo e às

características próprias dos municípios. A construção da metodologia de cálculo do IDHM teve como objetivo adequar a metodologia do IDH global para:

- ajustar a metodologia ao contexto brasileiro, buscando indicadores mais adequados para avaliar as condições de núcleos sociais menores – os municípios;

- adaptar a metodologia do IDH global aos indicadores disponíveis nos Censos Demográficos brasileiros, de forma a garantir uma mesma fonte de dados e comparabilidade entre todos os municípios brasileiros.

O uso da média geométrica nos cálculos relacionados ao IDHM Educação e IDHM está justificado junto ao United Nations Development Programme – Human Development Reports – Frequently Asked-Questions (http://hdr.undp.org/en/faq-page#t292n2880), como segue:

Em 2010, a média geométrica foi introduzida para calcular o IDH. O mau desempenho em qualquer dimensão é refletido diretamente na média geométrica. Em outras palavras, um baixo desempenho em uma dimensão não é linearmente compensado por um desempenho mais alto em outra dimensão. A média geométrica reduz o nível de substitutibilidade entre as dimensões e, ao mesmo tempo, garante que uma queda de 1% no índice de, digamos, expectativa de vida tenha o mesmo impacto no IDH que uma queda de 1% no índice de educação ou renda. Assim, como base para comparações de realizações, esse método também respeita mais as diferenças intrínsecas entre as dimensões do que uma média simples.

As subseções a seguir explicam os conceitos relacionados a cada subíndice do IDHM e as fórmulas de cálculo utilizadas.

4.4.3 IDHM de Longevidade

Conforme informado no AtlasBrasil, o IDHM de Longevidade é calculado por um único indicador medido pela expectativa de vida de uma pessoa ao nascer, por método indireto, a partir dos dados dos Censos Demográficos do IBGE [IBGE 2020]. Esse indicador mede as taxas de mortalidade para cada faixa etária no município levando em consideração tanto as doenças quanto as causas externas, como acidentes e violência em

geral. Depende de alguns fatores constantes que indicam os valores mínimo e máximo de referência considerados para o indicador de expectativa de vida ao nascer.

A expressão (4.3) apresenta o cálculo do valor do índice $IDHM_{longevidade}$, em que LE é a expectativa de vida (em anos) ao nascer e V_{Max} e V_{Min} são estimativas dos limites estabelecidos para expectativa de vida de um indivíduo em um determinado município.

$$\mathrm{IDHM_{Longevidade}} = \frac{\mathrm{LE} - \mathrm{V_{Min}}}{\mathrm{V_{Max}} - \mathrm{V_{Min}}} \tag{4.3}$$

Os limites seguem os mesmos valores adotados pelo IDHM, em que o $Valor_{Max}$ = 85 (anos) e $Valor_{Min}$ = 25 (anos). A seguir é mostrado o cálculo do $IDHM_{Longevidade}$, de um município em que a expectativa de vida é de 72 anos.

$$\mathrm{IDHM_{Longevidade}} = \frac{72 - 25}{85 - 25} = \frac{47}{60} = 0{,}784$$

4.4.4 IDHM de Educação

Como informado no AtlasBrasil, o IDHM da Educação é medido por meio de dois indicadores. O primeiro deles é dado pela escolaridade da população adulta ($escolaridade_{adulta}$), medida pelo percentual de pessoas com idade igual ou superior a 18 anos com ensino fundamental completo. Esse indicador tem peso 1. O segundo indicador é dado pelo fluxo escolar da população jovem ($escolaridade_{jovem}$), avaliado como a média aritmética dos percentuais de (1) crianças entre 5 e 6 anos frequentando a escola, (2) jovens entre 11 e 14 anos frequentando os anos finais do ensino fundamental, (3) jovens entre 15 e 17 anos com ensino fundamental completo e (4) jovens de 18 a 20 anos com ensino médio completo. Esse indicador tem peso 2.

O indicador fluxo escolar da população jovem acompanha a população em idade escolar em quatro momentos importantes da sua formação. Isso facilita aos gestores identificar se crianças e jovens estão nas séries adequadas, nas idades certas.

O índice IDHM Educação ($IDHM_{Educação}$) é calculado como a média geométrica dos dois indicadores, escolaridade da população adulta, ponderado pelo valor 1 e o fluxo escolar da população jovem, ponderado pelo valor 2. Os dados a partir dos quais

indicadores e índice são calculados, são captados do Censo Demográfico do IBGE. A expressão (4.4) apresenta a fórmula do cálculo do índice $IDHM_{Educação}$.

$$\mathrm{IDHM}_{\mathrm{Educação}} = \sqrt[3]{(\mathrm{escolaridade}_{\mathrm{adulta}} \times 1) \times (\mathrm{escolaridade}_{\mathrm{jovem}} \times 2)} \qquad (4.4)$$

4.4.5 IDHM de Renda

O AtlasBrasil informa que o IDHM de Renda é medido pela renda municipal *per capita*, ou seja, a renda média dos residentes de um determinado lugar (município, estado, região, etc.). É a soma da renda de todos os residentes, dividida pelo número de residentes no município, inclusive crianças e pessoas sem registro de renda. Esse indicador mede a capacidade média de aquisição de bens e serviços por parte dos habitantes do município de referência evidenciando se eles possuem condições de arcar financeiramente com suas necessidades básicas, tais como água, alimento e moradia. Os dados são captados dos Censos Demográficos do IBGE e estão expressos em reais.

O índice possui uma limitação, que é a de não considerar a desigualdade de renda entre os habitantes da área de referência. Isso pode refletir uma falsa realidade, pois, ao mesmo tempo em que o município apresenta uma renda *per capita* elevada, pode ter uma grande parcela da população vivendo na pobreza.

O cálculo do índice de Renda do IDHM ($IDHM_{Renda}$) é realizado por meio do uso da expressão (4.5), cujo objetivo é aproximar os maiores valores de renda *per capita* dos menores valores, reduzindo a desigualdade de renda existente. Porém, esse procedimento considera que, à medida que a renda *per capita* (RPC_local_de_referência) se eleva, o retorno desse acréscimo, em termos de desenvolvimento humano, diminui.

$$\mathrm{IDHM}_{\mathrm{Renda}} = \frac{[\ln(\mathrm{RPC\_local\_de\_referência}) - \ln(\mathrm{VRMin})]}{[\ln(\mathrm{VRMax}) - \ln(\mathrm{VRMin})]} \qquad (4.5)$$

O exemplo a seguir tem como base o Valor de Referência Máximo (VRMax) de R$4.033,00 que corresponde ao valor da menor renda *per capita* entre os 10% mais ricos residentes no Distrito Federal, localidade com maior renda média do país, e o Valor de

Referência Mínimo (VRMin) de R$8,00 correspondente a aproximadamente US$100 PPC que é o limite adotado para o cálculo do IDH Global. O PPC é uma métrica comparativa entre moedas de diferentes países que utiliza um índice para medir o poder de compra. Essa medida considera a quantidade em moeda necessária para adquirir um conjunto de produtos e serviços em um país, podendo ser comparada com a medida de outros países. Na análise do PPC podem ser utilizados o preço de bens, o Produto Interno Bruto (PIB), a renda *per capita* ou índices de preço utilizados em cálculos de medição da inflação [Dicionário Financeiro].

O cálculo do $\text{IDHM}_{\text{Renda}}$ para um município com renda *per capita* de R$ 832,45, por exemplo, é feito como segue:

$$\text{IDHM}_{\text{Renda}} = \frac{[\ln(832{,}45) - \ln(8{,}00)]}{[\ln(4033{,}00) - \ln(8{,}00)]} = 0{,}746$$

4.4.6 Calculando o Índice de Desenvolvimento Humano Municipal (IDHM)

O Índice de Desenvolvimento Humano Municipal (IDHM) é um ajuste metodológico ao IDH Global, e foi publicado em 1998, a partir dos dados do Censo de 1970, de 1980 e de 1991 e publicado em 2003, a partir dos dados do Censo de 2000. Segundo os órgãos responsáveis [PNUD 2020], o IDHM tem importância relevante na avaliação do desenvolvimento dos municípios e regiões metropolitanas brasileiras pelos seguintes motivos:

- cria um contraponto ao PIB, popularizando o conceito de desenvolvimento centrado nas pessoas, e não na visão de que desenvolvimento se limita a crescimento econômico.
- viabiliza a comparação entre os municípios brasileiros ao longo do tempo sintetizando uma realidade complexa em um único número.
- estimula formuladores e implementadores de políticas públicas no nível municipal a priorizar a melhoria da vida das pessoas em suas ações e decisões.

Como comentado anteriormente, o IDHM compreende indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda. O índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano. A Figura 4.5, extraída de [AtlasBrasil 2020] mostra o mapa do Brasil colorido com cinco cores, em que cada cor

representa uma faixa de valores de IDHM, associados às regiões do território nacional mostradas no mapa, coloridas por ela.

O IDHM é calculado como a média geométrica dos três índices associados, *i.e.*, $\text{IDHM}_{\text{Longevidade}}$, $\text{IDHM}_{\text{Educação}}$ e $\text{IDHM}_{\text{Renda}}$, como mostra a expressão (4.6).

$$\text{IDHM} = \sqrt[3]{\text{IDHM}_{\text{Longevidade}} \times \text{IDHM}_{\text{Educação}} \times \text{IDHM}_{\text{Renda}}} \tag{4.6}$$

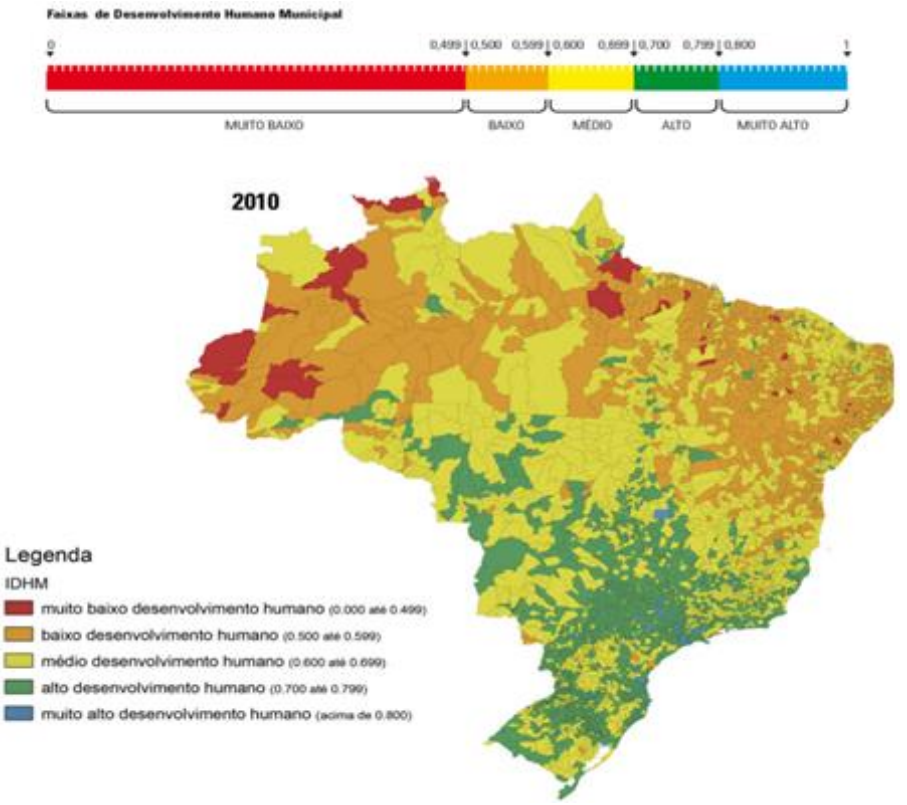


Figura 4.5 Mapa do IDHM do Brasil em 2010 [AtlasBrasil 2020]. http://atlasbrasil.org.br/.

4.4.7 População

O indicador final utilizado como um preditor potencial do vencedor das eleições em um município específico é o tamanho da população. Este é um dos indicadores comumente utilizados na análise eleitoral, uma vez que os grandes municípios tendem a se comportar de forma bem diferente dos pequenos em eleições.

Para facilitar a análise**,** os municípios serão classificados de acordo com seu número de habitantes. Não existe uma classificação única comumente aceita e por essa razão será utilizada a classificação adotada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em (IBGE 2020), mostrado na Tabela 4.1.

Tabela 4.1 Classificação dos municípios de acordo com o tamanho da população.

Classificação	Número de habitantes
Pequeno	0 – 20.000
Medio	20.000 – 100.000
Medio-Grande	100.000 – 500.000
Grande	500.000 – 1,000.000
Muito Grande	> 1.000.000

4.5 Resultados das Eleições de 2018

Segundo as estatísticas do TSE, as eleições de 2018 no Brasil levaram às urnas 117.366.956 de eleitores, ou seja, 79,86% dos eleitores aptos, conforme mostra a Tabela 4.2, cujos dados foram extraídos do *site* do TSE [TSE 2020].

Tabela 4.2 Total de eleitores aptos, de comparecimentos e de abstenções na eleição de 2018 (1º. Turno) [TSE-Estatísticas 2020].

Eleitores Aptos (%)	Comparecimentos (%)	Abstenções (%)
147.306.275 (100)	117.366.956 (79,68)	29.939.319 (20,32)

A Tabela 4.3 mostra o grau de instrução dos eleitores que compareceram às urnas nas eleições de 2018 (1º Turno), como apresentado em [TSE-Estatísticas 2020].

Tabela 4.3 Eleitores na eleição de 2018 (1º. Turno) por grau de instrução [TSE-Estatísticas 2020].

Grau de Instrução	Total	%
Ensino Fundamental Incompleto	38.064.617	25,84
Ensino Médio Completo	33.678.197	22,86
Ensino Médio Incompleto	24.864.650	16,88
Superior Completo	13,576.583	9,22
Lê e Escreve	13.147.331	8,93
Ensino Fundamental Completo	10.030.422	6,81
Superior Incompleto	7.313.928	4,97
Analfabeto	6.574.188	4,46
Não Informado	56.359	0,04

A Figura 4.6 mostra os resultados eleitorais no segundo turno das eleições de 2018, por unidade da federação. As cores verde e rosa permitem a visualização da divisão de opiniões existente entre os eleitores brasileiros.

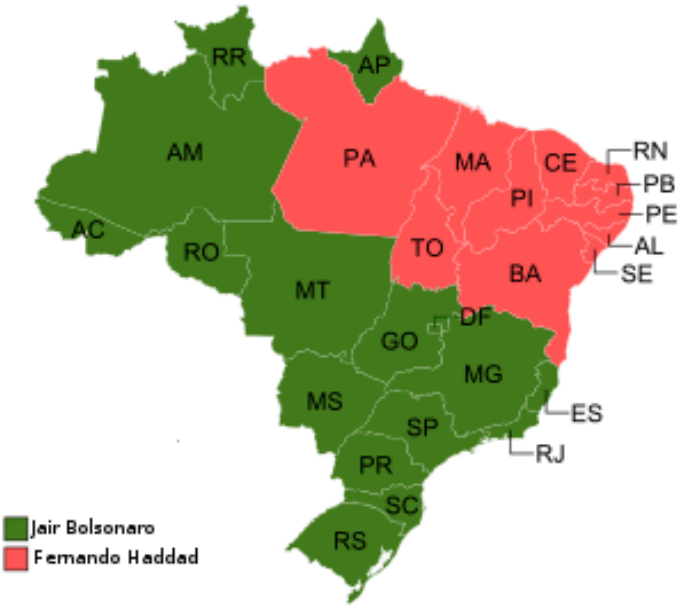


Figura 4.6 Mapa da distribuição de votos por candidato [TSE 2020].

4.6 Considerações Finais

O capítulo abordou várias das atividades de coleta de dados e apresentou em detalhes o Atlas do Desenvolvimento Humano no Brasil, um dos importantes repositórios de dados utilizados na pesquisa. O texto buscou descrever, de forma breve, o PNUD, incluindo um pequeno comentário sobre o Relatório de Desenvolvimento Humano de 2005, que foi um marco para as políticas mundiais de desenvolvimento e erradicação da pobreza. O capítulo também explicou de forma detalhada, os três subíndices do IDHM utilizados na pesquisa e abordou os resultados das eleições de 2018, por se tratar do domínio dos dados nos quais essa pesquisa se baseia.

Capítulo 5
Preparação dos Dados Usados na Pesquisa

Após terem sido devidamente selecionados, os dados passaram por processos de preparação e importação. Durante a preparação foi utilizada uma ferramenta capaz de realizar as etapas necessárias para que os dados pudessem ficar em condições de serem processados. Este capítulo descreve o processo de preparação e importação dos dados e faz uma breve introdução à ferramenta de software utilizada durante esse processo.

5.1 O Software Exploratory

Atualmente existem diversos softwares que auxiliam na tarefa de mineração e análise de dados. O Exploratory [Exploratory 2020] é uma ferramenta de visualização baseada em R, linguagem de programação para computação estatística [R Project 2020], que disponibiliza um ambiente computacional e experimental avançado, interativo e simplificado para análise de dados. A ferramenta possui recursos que permitem, de uma forma relativamente fácil e ágil, extrair, visualizar e interagir com os dados.

O Exploratory dispõe uma ampla variedade de tipos de visualização, o que ajuda muito na exploração dos dados e na descoberta de padrões ocultos de maneira rápida. Disponibiliza a implementação dos mais populares algoritmos de código aberto, o que agiliza o uso de tais algoritmos nos mais diversos tipos de aplicação.

A extração e integração das mais diversas fontes de dados como CSV, Excel, PostgreSQL, MySQL, Oracle, SQL Server, Google Analytics, entre outras, podem ser feitas de forma rápida e simples.

A Figura 5.1 mostra o d*ashboard* da ferramenta, com informações sumarizadas, referente a uma pesquisa sobre atrasos de voos em companhias aéreas americanas, durante o mês de agosto de 2016. São apresentados três quadros de informações em que o primeiro quadro contém a relação das 600 empresas de transporte aéreo analisadas durante a pesquisa, o segundo quadro apresenta o tempo de atraso mínimo e máximo geral em minutos e o terceiro quadro apresenta os voos com alcance de tempo inferior a

500 minutos. Um outro recurso disponibilizado no *Dashboard* permite que as medidas de tendência central possam ser calculadas de forma automatizada pela ferramenta Exploratory [Exploratory 2020].

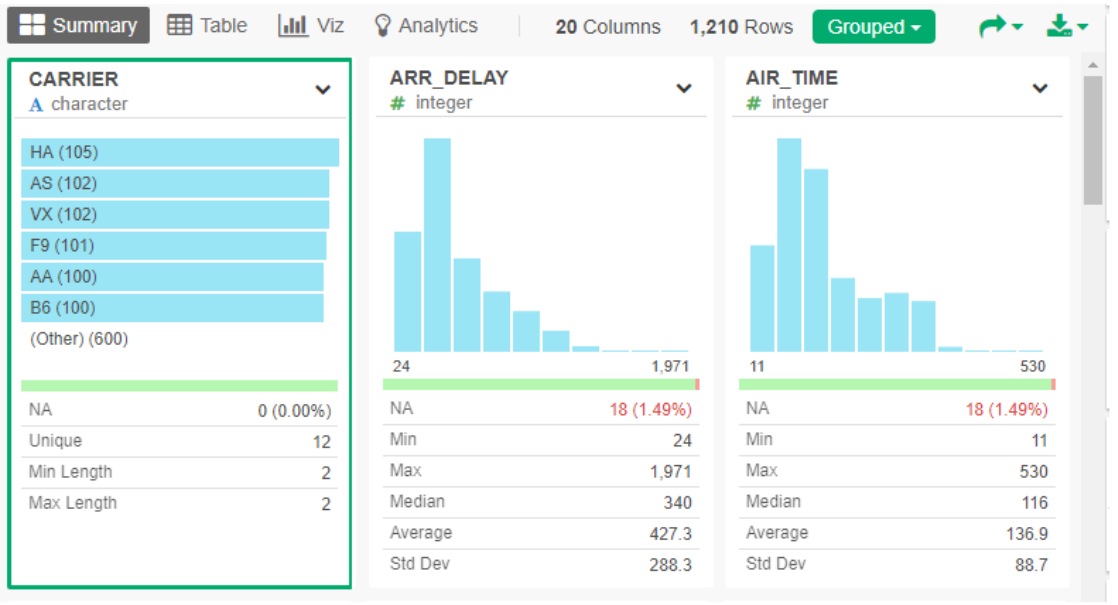


Figura 5.1 *Dashboard* da ferramenta Exploratory [Exploratory 2020].

5.2 O Processo de Importação dos Dados Originais

Os dados relacionados ao processo eleitoral, utilizados para o trabalho sendo realizado provêm de duas fontes distintas, (1) o AtlasBrasil [AtlasBrasil 2020] e (2) o repositório do TSE [TSE 2020]. Foi definido um processo para a importação dos dados, que se repetiu para ambas as fontes, e foi dividido em quatro etapas: (a) *download* dos dados originais; (b) tratamento das inconsistências; (c) seleção dos atributos (realizado manualmente); (d) importação dos dados.

Na primeira etapa foram realizados os *downloads* dos arquivos contendo as informações de ambas as fontes de dados.

Na segunda etapa foi realizado o trabalho de identificação das inconsistências e o tratamento inicial dos dados, deixando os arquivos preparados para a etapa de importação. Nessa etapa foram considerados como inconsistências, no contexto do trabalho os formatos de codificação diferentes de caracteres, diferenças na forma de escrita do nome dos municípios e os municípios ausentes. As inconsistências foram identificadas com o auxílio da ferramenta Exploratory durante os primeiros testes de importação e visualização dos dados.

A terceira etapa teve por objetivo a análise e seleção dos atributos relevantes para descrever as instâncias de dados, ao objetivo do trabalho, com vista a reduzir o volume de dados. Essa etapa foi realizada com o auxílio da ferramenta Exploratory.

Na quarta etapa os dados selecionados foram importados com o auxílio da ferramenta Exploratory. Durante essa etapa, e com o auxílio do Exploratory, os dados receberam o tratamento final eliminando as inconsistências identificadas.

A Figura 5.2 ilustra a metodologia utilizada no processo de importação.

Figura 5.2 Metodologia utilizada no processo de importação dos dados originais.

5.2.1 – Descrição dos Dados Originais

Como explicado no capítulo anterior, o AtlasBrasil disponibiliza mais de 200 indicadores relacionados à demografia, educação, renda, trabalho, habitação e vulnerabilidade associados aos municípios com dados provenientes dos Censos de 1991, 2000 e 2010. Para fins da pesquisa sendo realizada, foram selecionados indicadores que representam a população e o desenvolvimento humano de cada um dos municípios brasileiros, considerando como fonte os dados coletados pelo Censo realizado em 2010. Para compor o indicador de desenvolvimento humano foram selecionados os seguintes índices: (a) índice de desenvolvimento geral; (b) índice de renda; (c) índice de longevidade e (d) índice de educação. A seleção dos índices foi feita manualmente durante o processo de exportação dos dados, ilustrado na Seção 5.2.2. Um total de 27 arquivos foram gerados, representando os Estados brasileiros e o Distrito Federal. Cada arquivo é composto pelo conjunto dos municípios de cada Estado, representados por seus códigos e nomes, pela população total de cada município e pelos índices de desenvolvimento humano por município citados anteriormente. O total de registros de cada arquivo associado a um Estado é proporcional ao número de municípios existentes em cada Estado.

Os dados referentes aos índices de desenvolvimento humano por município são representados por valores no intervalo [0 – 1000], em que valores mais próximos de 1000 representam um melhor desempenho do município para aquele índice. A estrutura completa do arquivo e a descrição de cada atributo estão mostradas na Tabela 5.1 e a Tabela 5.2 mostra uma parte do arquivo com dados originais do AtlasBrasil.

Tabela 5.1 – Estrutura completa do arquivo com dados do AtlasBrasil de 2010.

Atributos	Dados	Descrição	Status
CODIGO	1200013	Código do Município	Não
ESPACIALIDADE	Acrelândia	Nome do Município	Selecionado
POPULACAO	12538	População do Município	Selecionado
IDHM 2010	604	IDHM Geral por Município	Selecionado
IDHM RENDA	584	IDHM de Renda por Município	Selecionado
IDHM	808	IDHM de Longevidade por	Selecionado
IDHM	466	IDHM de Educação por	Selecionado

Tabela 5.2 – Dados originais do AtlasBrasil 2010, relativos às seguintes espacialidades: Acrelândia, Assis Brasil, Brasiléia e Bujari.

Codigo	Espacialidades	Populacao Total	IDHM 2010	IDHM Renda	IDHM Longevidade	IDHM Educacao
1200013	ACRELÂNDIA	12538	604	584	808	466
1200054	ASSIS BRASIL	6072	588	578	77	456
1200104	BRASILÉIA	21398	614	619	77	485
1200138	BUJARI	8471	589	603	772	439

Os dados referentes ao segundo turno das eleições federais de 2018 foram baixados do repositório do TSE e, a exemplo dos dados do AtlasBrasil, também foi gerado um arquivo para cada Estado brasileiro e Distrito Federal, totalizando 27 arquivos. Cada um dos arquivos contém em média 120.000 registros, em que cada registro é descrito por 42 atributos, com informações detalhadas da eleição, de acordo com o Boletim de Urna (documento gerado durante a apuração dos votos). O número de registros em cada arquivo varia com o número de zonas eleitorais e municípios de cada Estado.

Após análise dos valores de cada atributo, quatro atributos de maior relevância foram selecionados para o trabalho, entre eles: o estado, o nome do município, os partidos políticos e a quantidade de votos. A seleção dos atributos foi realizada manualmente durante o processo de importação dos dados com o auxílio da ferramenta Exploratory. A Tabela 5.3 mostra a estrutura completa do arquivo com dados de exemplo e uma coluna adicional denominada "*Status*" que representa os atributos selecionados para o trabalho.

Tabela 5.3 – Estrutura completa do arquivo com os dados originais do TSE.

Atributos	Valores	Descrição	Status
DT_GERACAO	30/10/2018	Data da Geração do Arquivo	
HH_GERACAO	17:44:27	Hora da Geração do Arquivo	
ANO_ELEICAO	2018	Ano da Eleição	
CD_PLEITO	229	Código do Pleito	
DT_PLEITO	28/10/2018	Data do Pleito	
NR_TURNO	2	Número do Turno	
CD_ELEICAO	296	Código da Eleição	
DS_ELEICAO	Eleição Geral Federal	Descrição da Eleição	
DT_ELEICAO	28/10/2018	Data da Eleição	
SG_UF	AC	Sigla do Estado	Selecionado
CD_MUNICIPIO	1007	Código Município	
NM_MUNICIPIO	BUJARI	Nome Município	Selecionado
NR_ZONA	9	Número da Zona Eleitoral	
NR_SECAO	1	Número da Seção	
NR_LOCAL_VOTACAO	1104	Número Local Votação	
CD_CARGO	1	Código do Cargo	
DS_CARGO	Presidente	Descrição do Cargo	
NR_PARTIDO	(13, 17)	Número do Partido	
SG_PARTIDO	(PT, PSL)	Sigla do Partido	Selecionado
NM_PARTIDO	(Partido Trabalhista, Partido Social Liberal)	Nome do Partido	
QT_APTOS	346	Qtde Eleitores Aptos	
QT_COMPARECIMENTO	275	Qtde Comparecimentos	
QT_ABSTENCOES	71	Qtde Abstenções	
CD_TIPO_URNA	1	Codigo do Tipo da Urna	
DS_TIPO_URNA	Apurada	Descrição Tipo da Urna	
CD_TIPO_VOTAVEL	(1,2,3)	Código dos Tipos Votáveis	
DS_TIPO_VOTAVEL	(Nominal, Branco,	Descrição dos Tipos	
NR_VOTAVEL	(13,17,95,96)	Números Votáveis	
NM_VOTAVEL	(Fernando Haddad, Jair Bolsonaro, Branco, Nulo)	Nomes Votáveis	
QT_VOTOS	1	Qtde Votos	Selecionado
NR_URNA_EFETIVADA	1762466	Número da Urna	
CD_CARGA_1_URNA_EFET	138.512.240.315.942.3	Código Carga de Urna 1	
CD_CARGA_2_URNA_EFET	174217	Código Carga de Urna 2	
CD_FLASCARD_URNA_EF	632C5110	Código Cartão de Urna	
DT_CARGA_URNA_EFETIV	29/09/2018 14:30:00	Data da Carga	
DS_CARGO_PERGUNTA_S	1-120	Descrição Pergunta Seção	
DS_AGREGADAS	#NULO#	Descrição Seções Agregadas	
DT_ABERTURA	28/10/2018	Data da Abertura da Votação	
DT_ENCERRAMENTO	28/10/2018	Data Encerramento Votação	
QT_ELEITORES_BIOMETRI	7	Eleitores com Biometria	
NR_JUNTA_APURADORA	#NULO#	Número Junta Apuradora	
NR_TURMA_APUR	#NULO#	Número Turma Apuradora	

Como a pesquisa se concentra em analisar a relação entre os índices de desenvolvimento humano por município e a opção de voto dos eleitores em cada município, os atributos selecionados se tornaram fundamentais. Os atributos estado e nome do município foram utilizados para identificar cada um dos municípios brasileiros,

como também o nome do município foi utilizado como atributo de ligação entre as fontes de dados, se apresentando como única alternativa existente para essa operação. Para identificar os candidatos que disputaram o segundo turno das eleições, foi selecionado o atributo sigla do partido, uma vez que possui quantidade menor de caracteres em relação aos atributos nome do partido e nomes votáveis, facilitando na classificação dos dados e nas análises a serem realizadas posteriormente. Por fim, o atributo quantidade de votos foi selecionado, por representar a votação efetiva que cada candidato obteve, e também por possibilitar a identificação do candidato ganhador em cada um dos municípios.

A escolha de apenas quatro atributos não diminui a importância e a utilização dos demais atributos em outros trabalhos que buscam, por exemplo, detalhes da eleição por zona eleitoral.

A Tabela 5.4, mostra um extrato dos dados originais do TSE, em que ESPACIALIDADE (AtlasBrasil) é referenciado como NM_MUNICIPIO.

Tabela 5.4 – Extrato dos dados originais do TSE.

DT_ELEICAO	SG_UF	CD_MUNICIPIO	NM_MUNICIPIO	NR_ZONA	SG_PARTIDO	QT_VOTOS
28/10/2018	AC	1007	BUJARI	9	PT	48
28/10/2018	AC	1015	CAPIXABA	2	PSL	51
28/10/2018	AC	1023	PORTO ACRE	1	#NULO#	62
28/10/2018	AC	1021	SANTA ROSA DO PURUS	3	PSL	105

5.2.2 – Exportação dos Dados

Os dados de ambas as fontes foram exportados no formato CSV e importados, inicialmente para análises estatísticas realizadas utilizando a ferramenta Exploratory.

O site do AtlasBrasil disponibiliza uma área de consulta que permite a seleção dos indicadores e a exportação dos dados. Figuras 5.3 e Figura 5.4 mostram o passo a passo do processo.

Figura 5.3 – Primeiro passo, escolha da territorialidade.

Figura 5.4 – Segundo passo, escolha dos indicadores e exportação dos dados.

De modo semelhante, o site do TSE disponibiliza um repositório com o resultado das eleições desde 1945 o que permite a escolha do ano eleitoral que se deseja obter os dados, como mostra a Figura 5.5.

Figura 5.5 – Escolha dos dados eleitorais.

Após a seleção do ano e do boletim de urna correspondente, o site direciona o usuário para o repositório com os resultados das eleições por Estado, permitindo o *download* dos arquivos por meio de links. A Figura 5.6 mostra a tela de seleção dos Boletins de Urna por Estado.

Figura 5.6 – Seleção do boletim de urna por Estado.

5.3 Tratamento de Erros e Dados Incompletos

Para a junção dos dados foi necessário identificar atributos com valores em comum e que pudessem ser utilizados no relacionamento entre ambas as fontes de dados. O atributo referente ao nome do município se apresentou como única alternativa confiável, passando a ser "chave" para essa operação de junção dos dados. Porém, durante a análise dos dados, processo realizado manualmente, foi observado a presença das seguintes inconsistências: (a) formatos de codificação de caracteres diferentes; (b) diferença na grafia do nome dos municípios; (c) municípios ausentes. Todas as inconsistências encontradas estavam diretamente relacionadas ao atributo nome do município, sendo tratadas individual e manualmente como explicado a seguir:

a) **Formatos de codificação de caracteres diferentes**: a língua portuguesa utiliza tanto a acentuação gráfica como o cedilha na escrita das palavras, o que ocasiona, quando se trata de computadores, diferenças no formato de codificação dos caracteres. Para resolver esse problema é necessário a utilização de um conjunto de caracteres (*charset*) adequado para a correta codificação (*encoding*) dos caracteres que formam as palavras. Em ambas as fontes de dados, foi utilizado o padrão ISO-8859-1 padronizando o formato de codificação dos caracteres.

b) **Diferença na grafia do nome dos municípios**: outra inconsistência encontrada durante a análise foi a diferença de grafia utilizada na escrita dos nomes dos municípios. Por se tratar de repositórios diferentes a escrita de nomes dos municípios, em alguns casos, apresentou uma grafia utilizando caracteres diferentes, o que tornaria inviável sua utilização no relacionamento entre as fontes de dados. A Tabela 5.5 mostra a diferença na grafia utilizada para a escrita do nome dos municípios.

Tabela 5.5 – Diferença na grafia do nome de municípios entre os arquivos provindos do Atlas Brasil e do TSE.

AtlasBrasil	TSE
Espacialidades	**NM_MUNICIPIO**
Bras**ó**polis	Braz**ó**polis
Pingo-D'água	Pingo D'água
Passa-Vinte	Passa Vinte
Restinga Seca	Restinga S**ê**ca
Vespasiano Correa	Vespasiano Corr**ê**a

Todas as instâncias de dados que apresentaram esse tipo de inconsistência foram corrigidas, substituindo manualmente os caracteres divergentes, durante a etapa de importação.

c) **Municípios ausentes**: para que a junção dos dados pudesse ser realizada de forma confiável, sem perda de informações, seria necessário que ambas as fontes possuíssem os mesmos municípios em comum, uma vez que o atributo principal de relacionamento entre elas era o nome do município. Porém, a análise dos dados identificou a ausência de municípios no conjunto de dados do AtlasBrasil em relação aos municípios do conjunto de dados do TSE. O motivo evidente foi a diferença de tempo entre as duas fontes de dados *i.e.*, os dados do AtlasBrasil são referentes ao ano de 2010 e os dados do TSE referentes ao ano de 2018. No intervalo de tempo de 8 anos aproximadamente, novos municípios foram criados deixando os dados do AtlasBrasil desatualizados. A Tabela 5.6 mostra cinco dos municípios criados no Brasil após o ano de 2010.

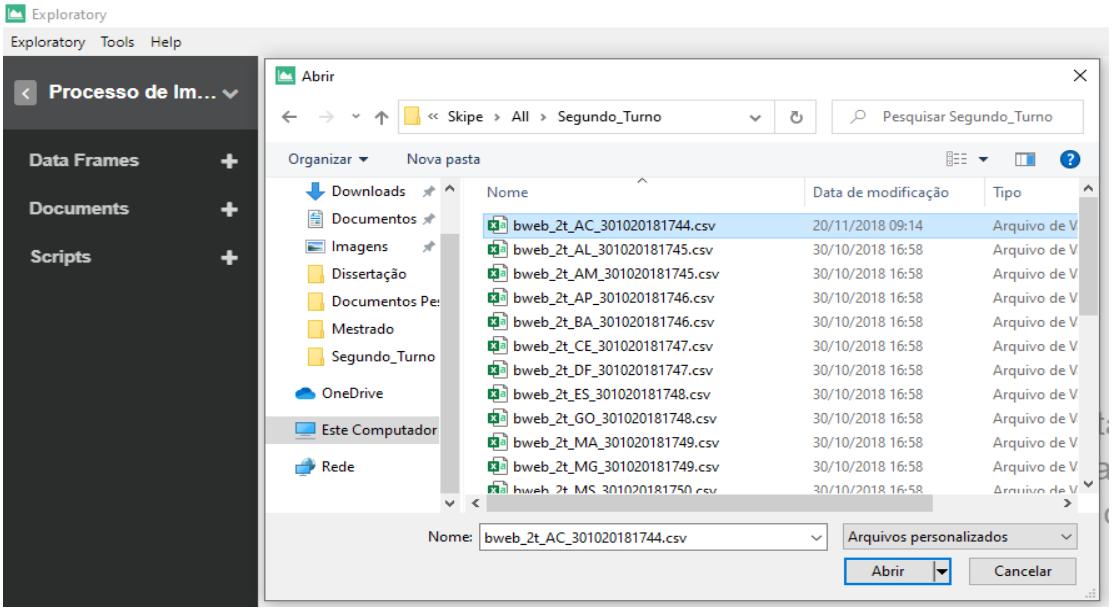
A solução adotada foi a exclusão dos registros dos municípios ausentes do arquivo do TSE. Os municípios excluídos representam 0,02% da população total.

Tabela 5.6 - Relação de cinco municípios criados depois do ano de 2010.

Estado	Município
SC	Pescaria Brava
SC	Balneário Rincão
PA	Mojuí dos Campos
RS	Pinto Bandeira
MT	Paraíso das Águas

5.4 Transformação dos Dados

Os arquivos com os dados originais foram importados individualmente, por meio do uso da ferramenta Exploratory, conforme mostram as Figuras 5.7 e 5.8.



. **Figura 5.7** – Importação dos arquivos CSV do TSE.

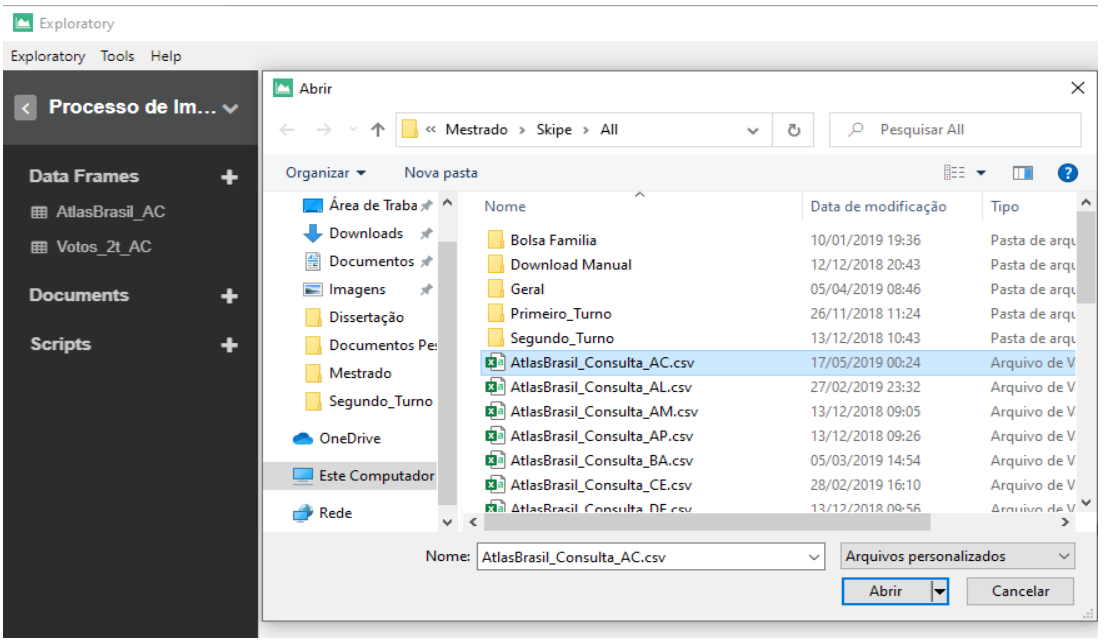


Figura 5.8 – Importação dos arquivos CSV do AtlasBrasil.

Durante o processo de importação foram realizados procedimentos para resolver inconsistências encontradas na etapa de análise. A Figura 5.9 mostra a seleção do Encoding ISO-8859-1 utilizado para padronizar o formato de codificação dos caracteres a todos os arquivos importados.

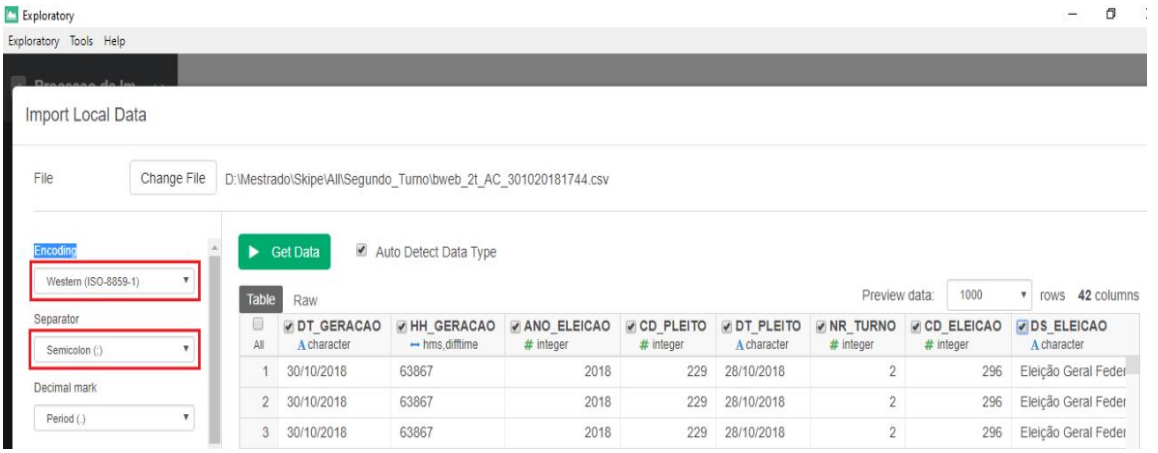


Figura 5.9 – Definição do *encoding* ISO-8859-1.

A etapa de seleção dos atributos foi realizada de forma manual durante o processo de importação dos dados. Essa etapa é mostrada na Figura 5.10, em que cada atributo é selecionado por meio do uso de uma caixa de seleção oferecida pela ferramenta Exploratory.

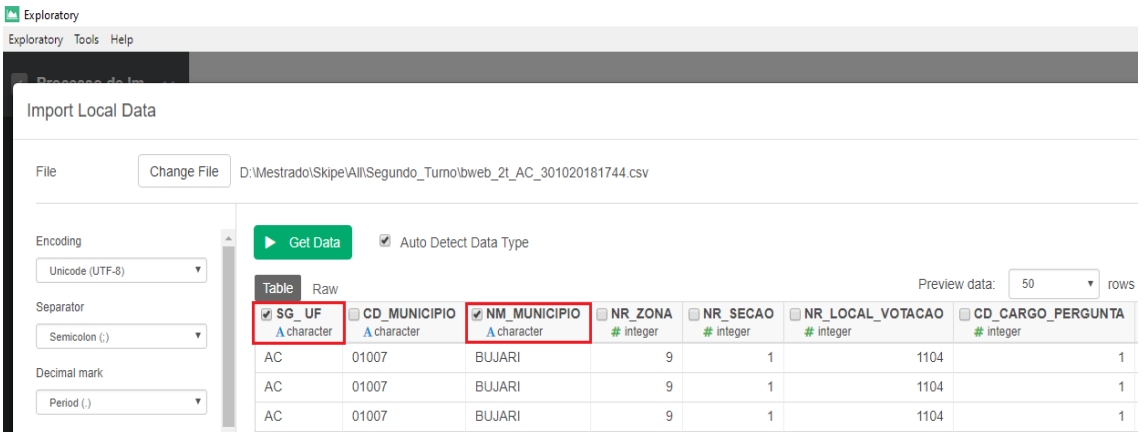


Figura 5.10 – Processo de seleção manual dos atributos do arquivo do TSE.

A Figura 5.11 mostra os dados de ambas as fontes após a finalização do processo de importação.

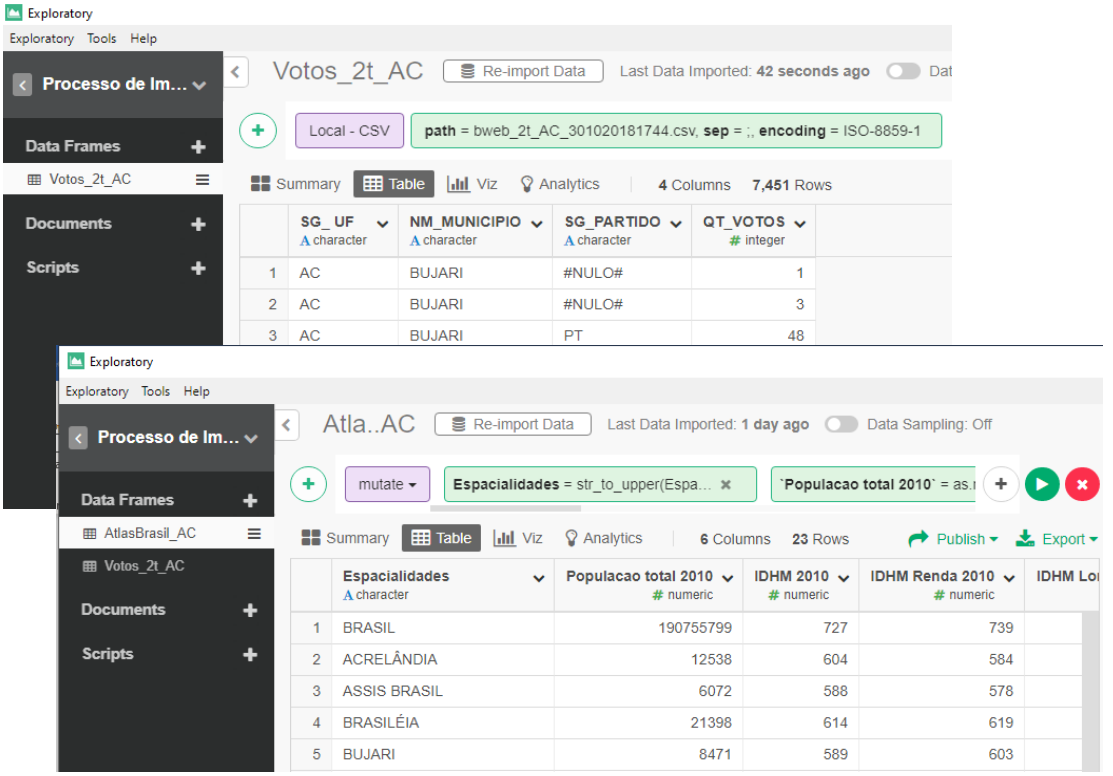


Figura 5.11 – Dados importados com o auxílio do Exploratory.

Após a importação dos dados foi iniciado o processo de organização dos dados, em que os dados foram agrupados por estado, município, partido político, como mostra a Figura 5.12.

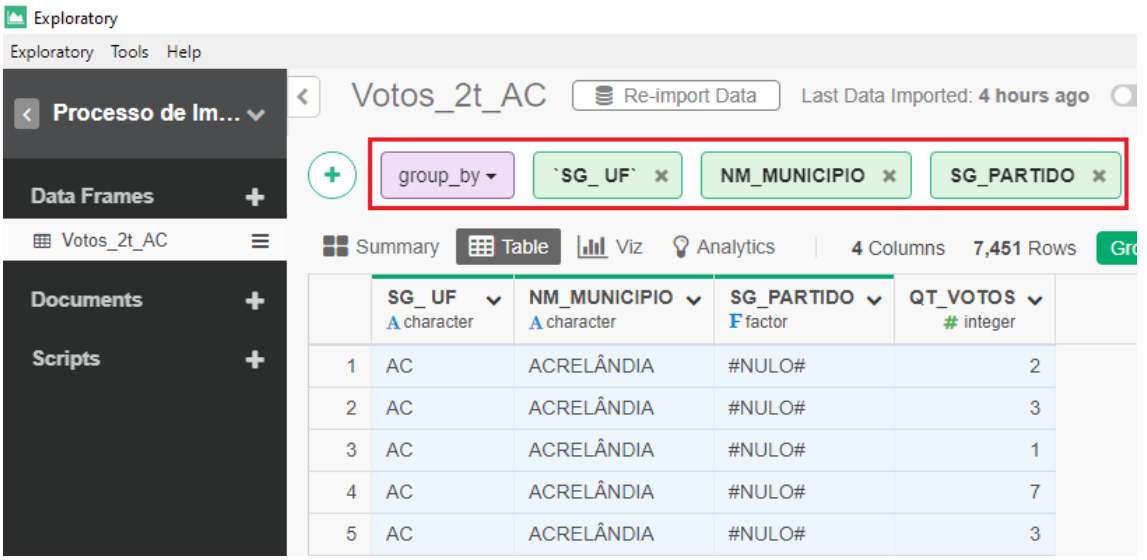


Figura 5.12 – Resultado do agrupamento dos dados utilizando os atributos estado, município e partido político.

Com os dados agrupados, o próximo passo foi a sumarização dos dados pelo número de votos, obtendo o total de votos nulos e por partido político em cada município, como mostrado na Figura 5.13.

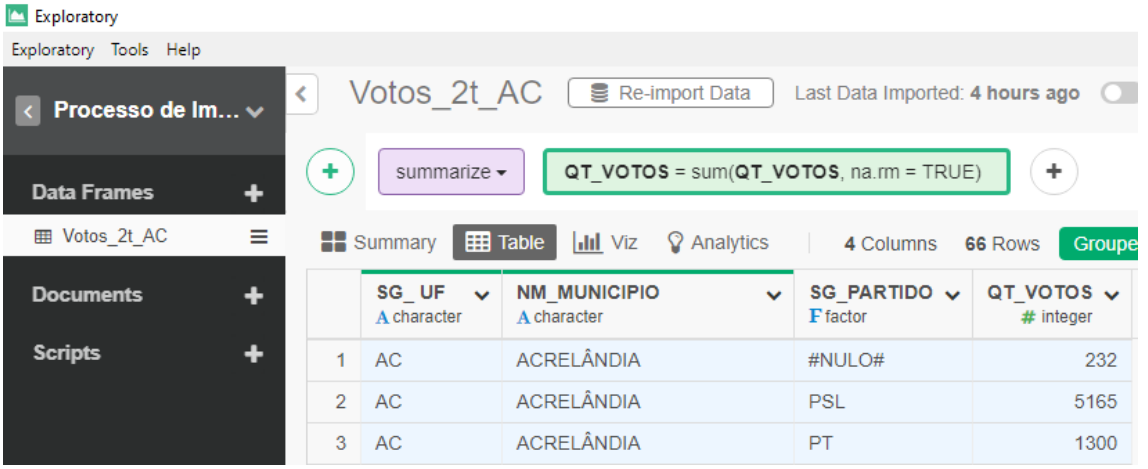


Figura 5.13 – Resultado do processo de sumarização dos dados utilizando a quantidade de votos.

No passo seguinte, os totais de votos foram transformados em atributos que representassem os partidos políticos permitindo a visualização clara dos votos de cada partido político por estado, conforme Figura 5.14.

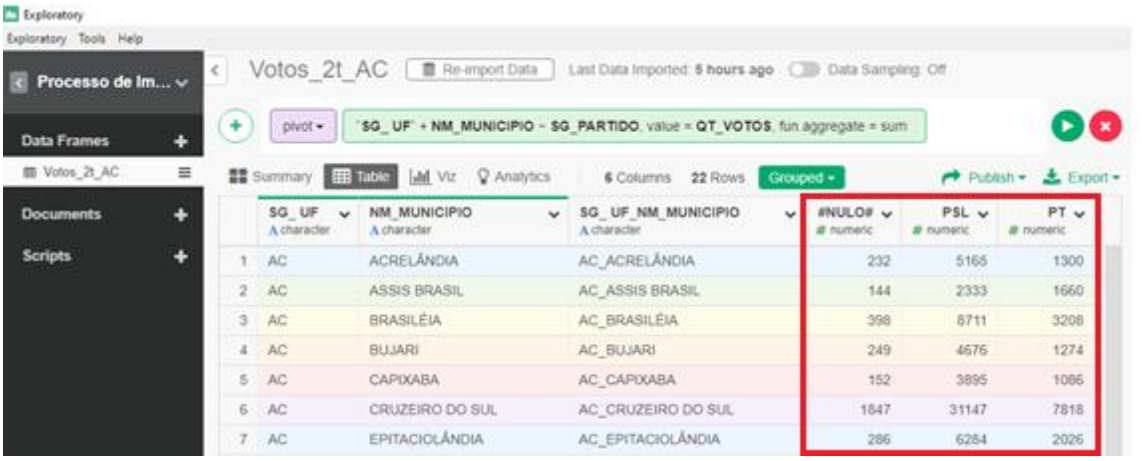


Figura 5.14 – Transformação dos totais de votos em atributos que representam os partidos políticos.

O conjunto de dados do AtlasBrasil também passou por transformações, tais como: (1) conversão do atributo ESPACIALIDADES, que contém o nome dos municípios, para letras maiúsculas (2) conversão dos atributos POPULACAO TOTAL, IDHM, IDHM RENDA, IDHM LONGEVIDADE, IDHM EDUCACAO para valores numéricos, como é mostrado na Figura 5.15. É valido lembrar que para a junção dos

dados os nomes dos municípios foram utilizados como atributo de ligação; por isso a importância da transformação do atributo ESPACIALIDADES para letras maiúsculas deixando os dados padronizados.

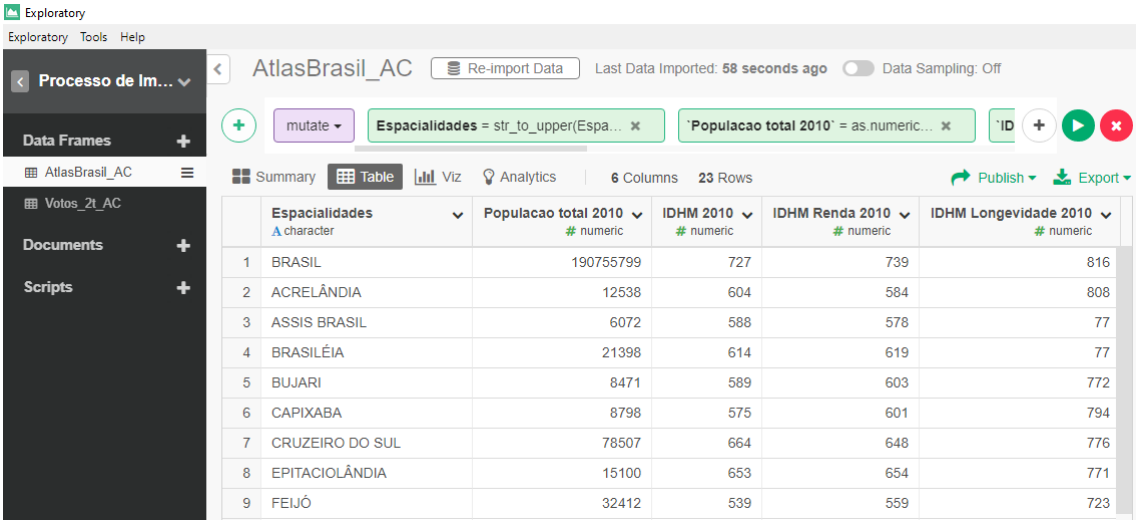


Figura 5.15 – Conversão do atributo espacialidades para letras maiúsculas e demais atributos em valores numéricos.

A Figura 5.16 mostra a etapa de junção das fontes de dados através do recurso de relacionamento (JOIN) oferecido pela ferramenta Exploratory. Essa etapa foi realizada utilizando o nome dos municípios que atuam como atributos de ligação entre as fontes de dados.

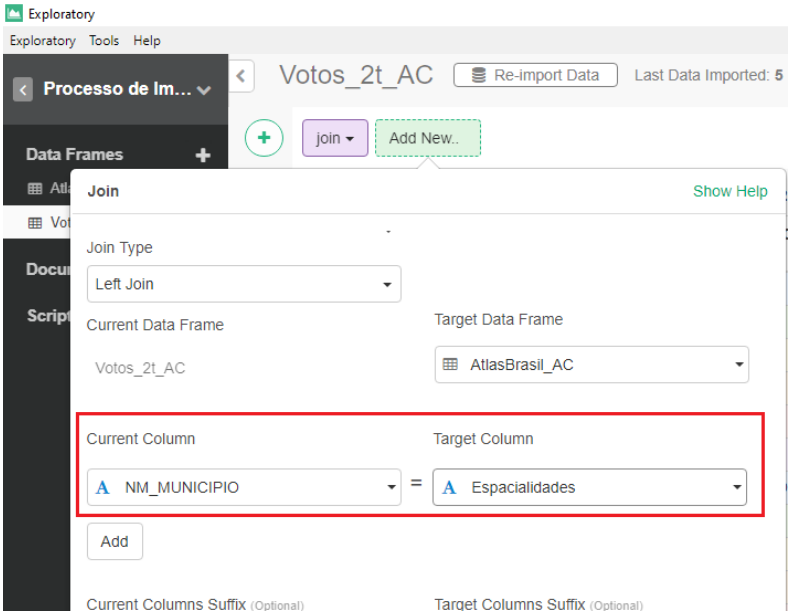
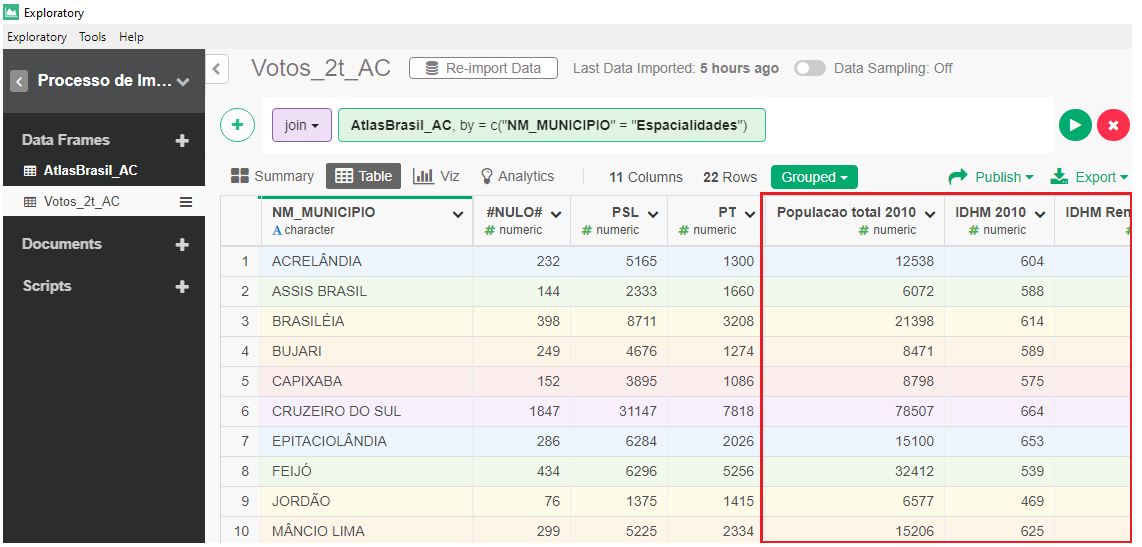


Figura 5.16 – Processo de relacionamento entre as fontes de dados.

A Figura 5.17 mostra o resultado do relacionamento entre as fontes de dados estabelecido na etapa anterior.



	NM_MUNICIPIO (character)	#NULO# (numeric)	PSL (numeric)	PT (numeric)	Populacao total 2010 (numeric)	IDHM 2010 (numeric)	IDHM Ren
1	ACRELÂNDIA	232	5165	1300	12538	604	
2	ASSIS BRASIL	144	2333	1660	6072	588	
3	BRASILÉIA	398	8711	3208	21398	614	
4	BUJARI	249	4676	1274	8471	589	
5	CAPIXABA	152	3895	1086	8798	575	
6	CRUZEIRO DO SUL	1847	31147	7818	78507	664	
7	EPITACIOLÂNDIA	286	6284	2026	15100	653	
8	FEIJÓ	434	6296	5256	32412	539	
9	JORDÃO	76	1375	1415	6577	469	
10	MÂNCIO LIMA	299	5225	2334	15206	625	

Figura 5.17 – Resultado do relacionamento entre as fontes de dados.

Na etapa seguinte, foi criada uma fórmula de cálculo para definir o partido político vencedor das eleições em cada um dos municípios. Como resultado, um novo atributo intitulado “ganhador”, com valores booleanos, foi gerado. Os registros de municípios em que o atributo “ganhador” foi rotulado como verdadeiro representaram vitória do partido político PSL e os registros rotulados como falso representaram vitória do partido político PT. Essa etapa foi fundamental para as análises preliminares. As Figura 5.18 e 5.19 mostram a definição da fórmula para a criação do atributo “ganhador” e o resultado final do processo de criação do atributo.

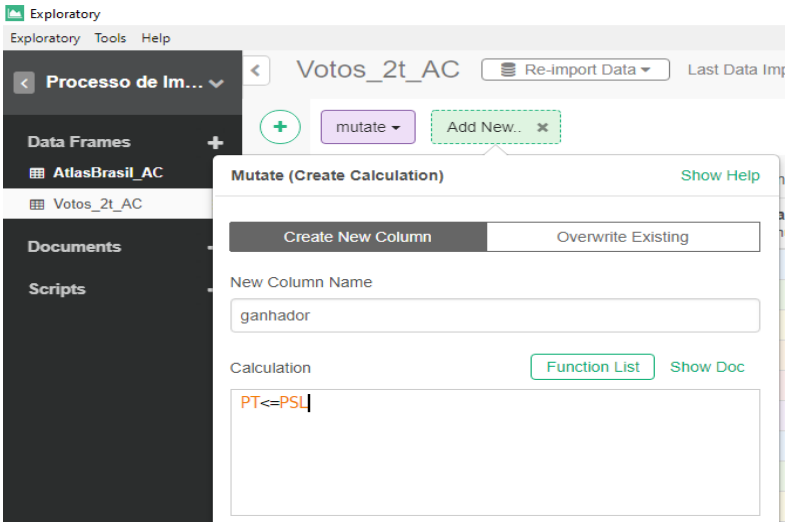


Figura 5.18 – Definição da fórmula de cálculo e criação do atributo "ganhador".

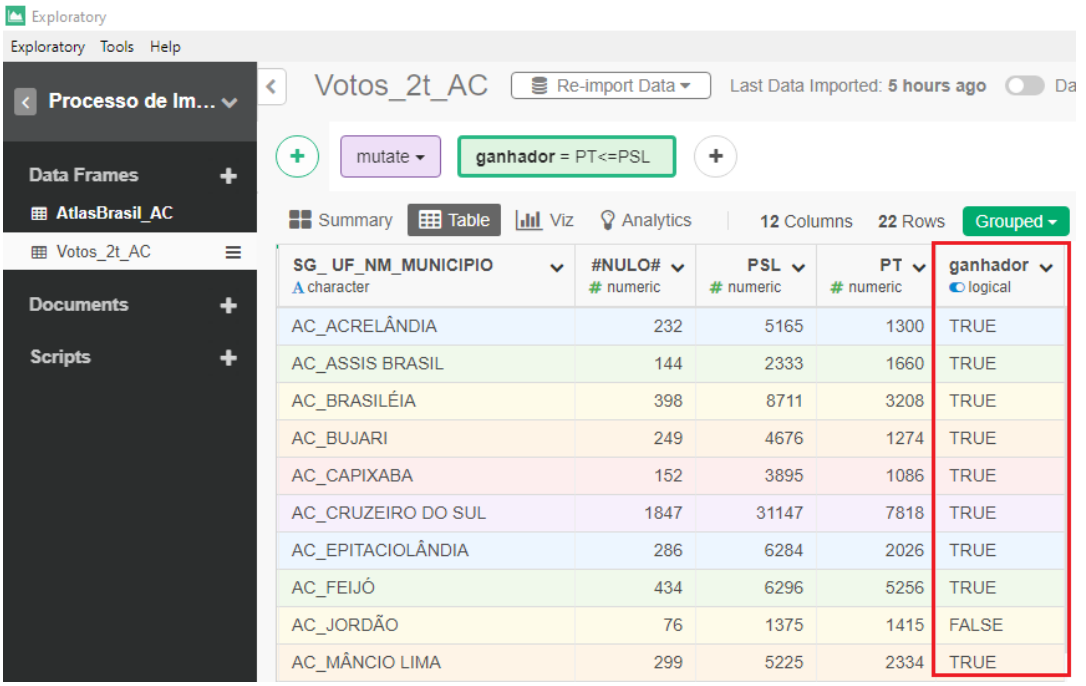


Figura 5.19 – Resultado do processo de criação do atributo "ganhador".

Vale ressaltar que o processo se repetiu para todos os arquivos de ambas as fontes de dados.

Com os rótulos definidos, foram realizadas algumas análises preliminares com base nos gráficos gerados pela ferramenta de análise, que auxiliaram na visualização do resultado das pesquisas – ver Figura 5.20. Os demais gráficos serão apresentados e discutidos no Capítulo 6.

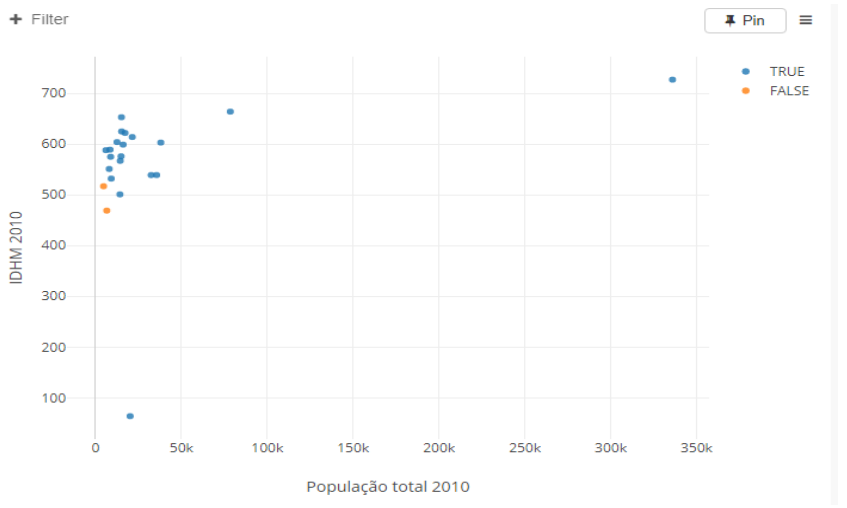


Figura 5.20 Gráfico Scatter – cruzamento entre o atributo IDHM e o atributo população total, utilizando os dados referentes ao estado do Acre.

A Tabela 5.7 mostra um extrato dos dados após a conclusão do processo de importação, com os valores mínimo e máximo dos indicadores de IDHM e População, classificados e ordenados por Estado (UF).

Tabela 5.7 – Extrato dos dados importados com os valores mínimo e máximo dos indicadores de IDHM e População classificados e ordenados por Estado.

UF	População Mínima	População Máxima	IDHM Mínimo	IDHM Máximo	RENDA Mínima	RENDA Máxima	LONG. Mínima	LONG. Máxima	EDUCACAO Mínima	EDUCACAO Máxima
AC	4.691	336.038	64	727	58	729	72	808	34	661
AL	2.866	932.748	57	721	52	739	7	799	4	635
AM	7.326	1.802.014	5	737	51	738	8	826	4	658
AP	3.793	398.204	64	733	61	723	76	801	55	692
BA	2.612	2.675.656	6	759	6	781	7	835	4	679
CE	4.164	2.452.185	6	754	7	749	72	832	5	695
DF	-	2,570.160	-	824	-	863	-	873	-	742
ES	4.516	414.586	7	845	63	876	81	864	54	805
GO	1.020	1.302.001	7	799	7	824	81	878	6	739
MA	4.020	1.014.837	5	768	4	741	7	813	4	752
MG	815	2.375.151	6	813	6	864	8	886	4	744
MS	2.928	786.797	7	784	64	758	8	873	45	724
MT	1.096	551.098	6	785	6	774	8	856	44	726
PA	3.431	1.393.399	6	746	47	751	8	828	28	673
PB	1.256	723.515	6	763	6	782	8	832	4	714
PE	2.630	1.537.704	6	788	53	798	8	839	4	748
PI	1.253	814.230	6	751	5	731	7	816	4	707
PR	1.409	1.751.907	6	823	6	806	8	869	6	768
RJ	5.269	6.320.446	7	837	7	887	8	855	7	773
RN	1.618	803.739	53	766	6	768	7	835	39	726
RO	2.315	428.527	7	736	63	764	8	825	6	659
RR	6.750	284.313	65	752	6	737	78	816	276	708
RS	1.216	1.409.351	7	805	7	867	81	888	5	754
SC	1.465	515.288	7	847	7	854	8	894	5	789
SE	2.275	571.149	6	664	52	784	7	823	43	708
SP	805	11.253.503	7	862	7	891	8	887	7	825
TO	1.037	228.332	5	788	6	789	7	847	6	749

Capítulo 6
Descrição dos Experimentos, Análise dos Resultados e Conclusões do Trabalho.

Este capítulo tem como objetivo apresentar uma descrição da metodologia utilizada na condução dos experimentos realizados ao longo da pesquisa, os resultados obtidos e uma discussão sobre os resultados subsidiada por análises estatísticas. A parte experimental do projeto envolveu também, em uma segunda etapa, o uso de algoritmos de aprendizado supervisionado e não supervisionado usando os mesmos dados que foram submetidos à análise estatística inicial. O capítulo é finalizado com a apresentação de um conjunto de conclusões sobre o trabalho realizado, particularmente as relacionadas ao domínio de dados que foi foco dos experimentos, as relacionadas às análises estatísticas realizadas e um sumário sobre os resultados dos experimentos com algoritmos de agrupamento no domínio de dados eleitorais considerado. Possíveis continuidades e extensões do trabalho já realizado e descrito nesta dissertação são também informadas.

6.1 – Considerações Iniciais

Como comentado anteriormente, a principal ideia que subsidiou a proposta e desenvolvimento da pesquisa descrita neste documento foi a de investigar possíveis relações entre características que descrevem eleitores, bem como características que descrevem municípios brasileiros, e o resultado obtido nas eleições federais à presidência, ocorridas em 2018.

Características escolhidas foram as que envolvem índices IDHM, bem como o número de habitantes de cada um dos municípios brasileiros. O IDHM, como detalhado na Seção 4.4, busca caracterizar numericamente o nível de desenvolvimento e a qualidade de vida oferecida por municípios brasileiros. O IDHM é derivado do Índice de Desenvolvimento Humano (IDH), um indicador global considerado uma das métricas mais amplamente aceitas para representar o status de desenvolvimento de um país.

O capítulo está organizado como segue. A Seção 6.2 descreve inicialmente algumas inspeções que foram realizadas por meio de gráficos e cálculos estatísticos básicos com objetivo de obter uma compreensão inicial dos dados. Na sequência, a Seção 6.3 apresenta os resultados das análises realizadas utilizando os cinco atributos selecionados, referentes aos índices IDHM e População. A Seção 6.4 apresenta os resultados preditivos obtidos com a utilização de algoritmos supervisionados de AM. A Seção 6.5 descreve experimentos relacionados ao uso do algoritmo de agrupamento conhecido como k-Means (ver Seção 3.4) no conjunto de dados alvo do trabalho, com o intuito de investigar a relevância dos atributos utilizados na descrição dos dados, no resultado da eleição considerada. Por fim, a Seção 6.6 apresenta as conclusões gerais do trabalho de pesquisa realizado e possíveis continuidades.

6.2 – Dados e Análise dos Resultados

Os dados utilizados na pesquisa foram importados e preparados como descrito no Capítulo 5, em que cada instância representa um município brasileiro. Para a condução dos experimentos e análises dos resultados, foi necessário a qualificação dos municípios de acordo com o tamanho populacional e de acordo com os valores dos índices de IDHM, isso foi feito com o intuito de caracterizar os municípios em cinco categorias. Tal qualificação foi realizada de acordo com os critérios apresentados pela tabela de qualificação de municípios por tamanho da população e pelo mapa do IDHM mostrados no Capítulo 4. A Tabela 6.1 apresenta os valores utilizados na qualificação dos municípios; com vista aos experimentos a serem realizados.

Tabela 6.1 – Qualificação dos municípios por população e índices de IDHM.

Qualificação dos municípios por população	Qualificação dos municípios por índices IDHM
Pequeno	Muito baixo
Médio	Baixo
Médio-grande	Médio
Grande	Alto
Muito grande	Muito alto

A Tabela 6.2 apresenta um extrato do conjunto de dados após a qualificação dos municípios, de acordo com a respectiva população e dos índices de IDHM. Vale

ressaltar que todas as análises e experimentos foram realizados utilizando um conjunto de dados único.

Tabela 6.2 - Conjunto de dados utilizados durante as etapas de análises.

Município	Votos PSL	Votos PT	População	IDHM	IDHM Renda	IDHM Long	IDHM Educação	Tamanho Municipio	Classificação IDHM
Acrelândia	5.165	1.300	12.538	604	584	808	466	Pequeno	Médio
Montadas	896	2.541	4.990	59	545	748	505	Pequeno	Muito baixo
São Paulo	3.694.834	2.424.125	11.253.503	805	843	855	725	Muito Grande	Muito alto
Rio de Janeiro	2.179.896	1.105.393	6.320.446	799	84	845	719	Muito Grande	Muito alto
São J.R.Preto	177.175	49.211	408.258	797	801	846	748	Médio Grande	Alto

A primeira etapa da análise buscou obter uma compreensão preliminar dos dados. Para tanto, foram realizados cálculos estatísticos básicos e produção de gráficos, com o objetivo de investigar a existência de possíveis relações entre as variáveis envolvidas. As análises foram realizadas com o auxílio da Ferramenta Exploratory. Os subitens a seguir descrevem as análises realizadas utilizando os atributos de tamanho da população e subíndices de IDHM.

6.2.1 – Tamanho da População

A Tabela 6.3 mostra o número de municípios em que cada um, PT ou PSL, foi o vencedor. Os municípios estão agrupados levando em conta o tamanho da respectiva população, pela categorização mostrada na Tabela 6.1.

Tabela 6.3 - Contagem dos municípios em que o PT ou PSL obteve maior número de votos. As colunas referentes a percentuais (%PSL e %PT), englobam os valores referentes a dois grupos: {Pequeno, Médio}, {Médio-grande, Grande, Muito-grande}.

Tamanho do Município	PSL	%PSL	PT	%PT	PSL + PT
Pequeno	1.870	48	2.039	52	3.909
Médio	662		706		1.368
Médio-grande	196	80	48	20	244
Grande	19		4		23
Muito-grande	11		4		15
Total	2.758		2.801		5.559

Os números da Tabela 6.3 mostram que para os municípios pequenos e médios os votos são quase uniformemente divididos, com um percentual de 52% associado ao PT e de 48% associado ao PSL. No entanto, à medida que o tamanho da população cresce, os resultados tornam-se desequilibrados, favorecendo o PSL, com os votos ficando

divididos aproximadamente em 80% a 20% quando são considerados municípios médios, médio-grandes, grandes e muito grandes.

A Figura 6.1 mostra a relação entre os percentuais de votos obtidos pelo PSL, levando em consideração o tamanho populacional de cada município. A porcentagem de votos obtidos pelo PT é, claramente, igual a porcentagem obtida pelo PSL. Municípios com mais de 1 milhão de habitantes foram suprimidos para reduzir a escala e ilustrar melhor as tendências dos dados.

Intuitivamente, os resultados indicam que o tamanho da população não é um atributo que caracteriza bem o fato de uma cidade ter votado PT ou PSL. Na Seção 6.3 essa situação é considerada novamente e confirmada estatisticamente.

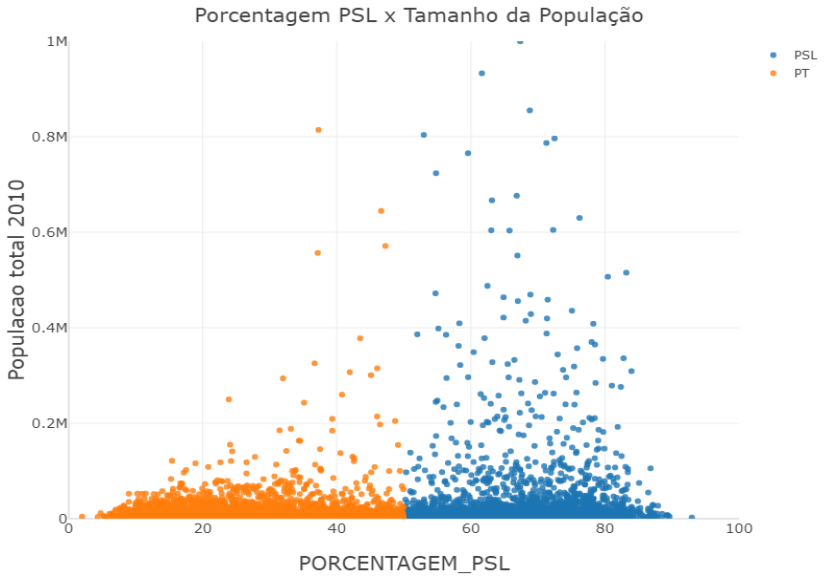


Figura 6.1. Relação entre os percentuais de votos obtidos pelo PSL e o tamanho da população.

6.2.2 IDHM de Educação

A Tabela 6.4 mostra o número de municípios que votaram em cada candidato, classificados pelo subíndice de $IDHM_{Educação}$. Diferente do tamanho da população, o subíndice de $IDHM_{Educação}$ parece produzir uma divisão clara entre os municípios. Do total de 3.760 municípios, cujos valores de $IDHM_{Educação}$ estão no conjunto {Muito baixo, Baixo} apenas em 31% deles o PSL teve mais votos, enquanto os 69% restantes preferiram o PT. Em 1.799 municípios, cujos valores de $IDHM_{Educação}$ estão no conjunto

{Médio, Alto, Muito alto}, o cenário é o inverso: com o PT recebendo apenas 12% dos votos e o PSL recebendo 88%.

Tabela 6.4 - Contagem de municípios por vencedor classificados por $IDHM_{Educação}$.

Classificação $IDHM_{Educação}$	PSL	%PSL	PT	%PT	PSL+PT
Muito baixo	397	31	1.554	69	1.951
Baixo	784		1.025		1.809
Médio	1.255		212		1.467
Alto	318	88	10	12	328
Muito alto	4		0		4

A Figura 6.2 representa a porcentagem de votos para o PSL em relação ao valor do subíndice de $IDHM_{Educação}$ para cada município. A tendência é nítida à medida que aumenta a porcentagem de votos do PSL, favorecendo claramente o PT em municípios em que o subíndice de $IDHM_{Educação}$ possui valores mais baixos e favorecendo o PSL em municípios em que tal subíndice possui valores mais altos.

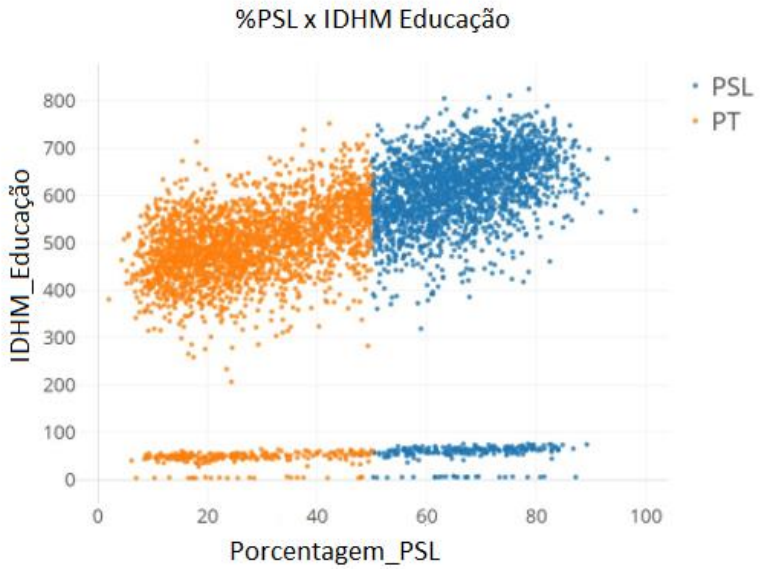


Figura 6.2. Porcentagem de votos PSL × $IDHM_{Educação}$.

6.2.3 IDHM de Renda

A Tabela 6.5 apresenta o número de municípios que elegeram um dos dois candidatos, abordados em seus respectivos subíndices $IDHM_{Renda}$. Assim como o $IDHM_{Educação}$, o $IDHM_{Renda}$ revela uma clara diferença entre os municípios com classificações muito baixas e baixas, e aqueles com classificações médias, altas e muito altas. Do total de 2.275 municípios cujos valores de $IDHM_{Renda}$ estão no conjunto

{Muito baixo, Baixo}, apenas 13% dos municípios preferiram o PSL, enquanto os 87% restantes apoiaram o PT. Em 3.284 municípios, cujos valores de $IDHM_{Renda}$ estão no conjunto {Médio, Alto, Muito alto}, 75% elegeram o PSL e apenas 25% preferiram o PT.

Tabela 6.5 - Contagem de municípios por vencedor classificados por $IDHM_{Renda}$.

Classificação por $IDHM_{Renda}$	PSL	%PSL	PT	%PT	PSL+PT
Muito baixo	266	13	404	87	670
Baixo	41		1.564		1.605
Médio	1.069	75	743	25	1.812
Alto	1.337		90		1.427
Muito alto	45		0		45

A Figura 6.3 representa a porcentagem de votos do PSL em relação ao subíndice de $IDHM_{Renda}$. Mais uma vez, verifica-se uma tendência clara de aumento de tal índice, à medida que aumenta a porcentagem de votos do PSL.

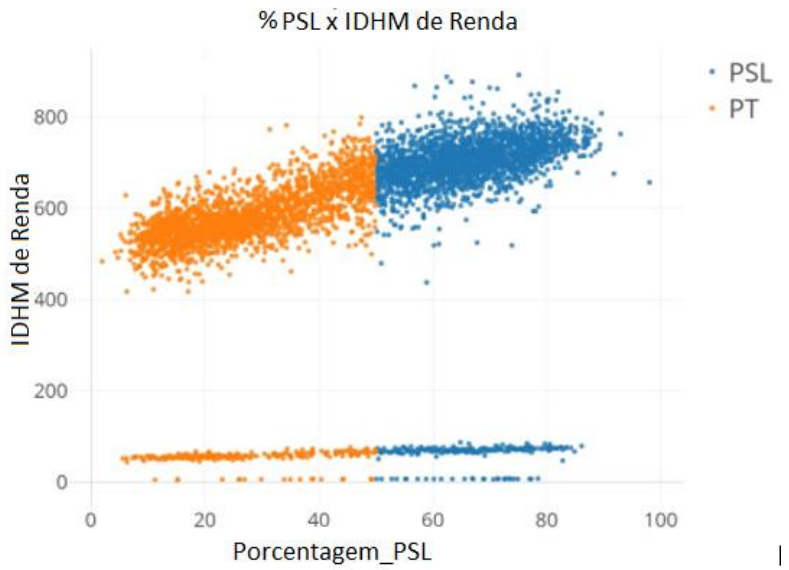


Figura 6.3. Porcentagem de votos PSL × $IDHM_{Renda}$.

6.2.4 IDHM de Longevidade

A Tabela 6.6 informa o número de municípios onde cada candidato venceu, dividido pela classificação do subíndice de $IDHM_{Longevidade}$.

Diferentemente dos subíndices de $IDHM_{Educação}$ e de $IDHM_{Renda}$, o uso do subíndice de $IDHM_{Longevidade}$ divide os vencedores entre os municípios no conjunto {Médio, Alto}, total de 2.149 municípios, dos quais 9% preferiram o PSL e 91% preferiram o PT,

aqueles que estão no conjunto {Muito alto}, total de 2.876 municípios, sendo que 79% optaram pelo PSL e 21% pelo PT e finalmente aqueles que estão no conjunto{ Muito baixo}, total de 534 municípios, onde os resultados são divididos quase uniformemente (54% PSL, 46% PT). Não há municípios no conjunto {Baixo}.

Tabela 6.6 - Contagem de municípios por vencedor classificados por $IDHM_{Longevidade}$.

Classificação por $IDHM_{Longevidade}$	PSL	%PSL	PT	%PT	PSL+PT
Muito baixo	288	54	246	46	534
Médio	0	9	67	91	67
Alto	199		1.883		2.082
Muito alto	2.271	79	605	21	2.876

A Figura 6.4 exibe a porcentagem de votos do PSL em relação ao índice $IDHM_{Longevidade}$. A forma do gráfico é mais linear do que as das figuras 6.2 e 6.3. Além disso, os valores parecem estar mais concentrados na parte superior e inferior da figura, com menos municípios no meio do espectro.

Intuitivamente, a forma mais linear do gráfico e o fato de que para municípios com valores muito baixos do subíndice de longevidade os resultados são divididos quase uniformemente, tornam este subíndice menos atraente como um preditor do candidato vencedor. Essa intuição é confirmada na Seção 6.3.

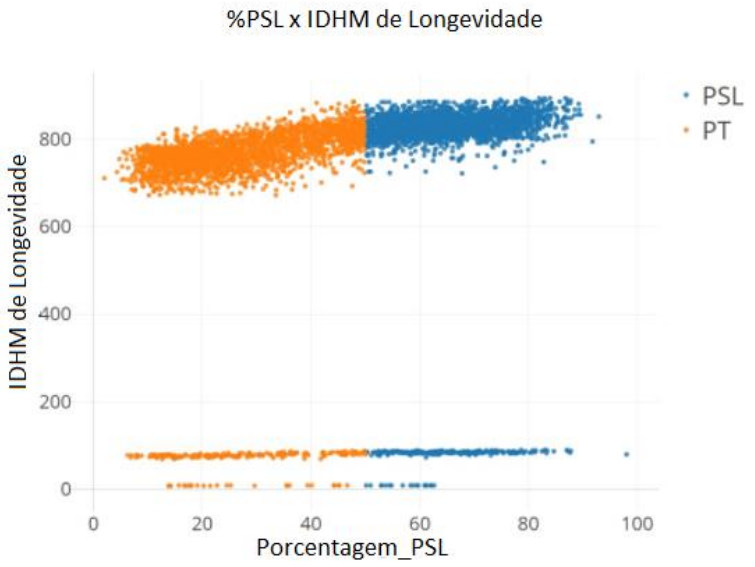


Figura 6.4 Porcentagem de votos PSL × $IDHM_{Longevidade}$.

6.2.5 IDHM

Por fim, a Tabela 6.7 mostra a contagem de municípios onde cada candidato venceu, dividida pela classificação de IDHM. Conforme mostrado no Capítulo 4, o índice de

IDHM é calculado como a média geométrica dos três subíndices *i.e.*, os associados à Educação, à Renda e à Longevidade.

O vencedor para os municípios cujos valores de IDHM estão no conjunto {Muito baixo}, total de 596 municípios, é dividido quase que igualmente, com 47% dos municípios elegendo o PSL e 53% dos municípios elegendo o PT. Municípios cujos IDHM estão no conjunto {Baixo, Médio}, total de 3.224 municípios, tendem a ser claramente mais favoráveis ao PT (63%, contra 37% do PSL), enquanto os municípios cujos IDHM estão no conjunto {Alto, Muito alto}, total de 1.739 municípios, são esmagadoramente favoráveis ao PSL (93%, contra 7% a favor do PT).

Tabela 6.7 - Contagem de municípios por vencedor classificados por IDHM.

Classificação IDHM	PSL	%PSL	PT	%PT	PSL+PT
Muito baixo	280	47	316	53	596
Baixo	38	37	1.176	63	1.214
Médio	827		1.183		2.010
Alto	1.574	93	126	7	1.700
Muito Alto	39		0		39

A Figura 6.5 representa os valores da porcentagem de votos do PSL em relação ao valor do índice de IDHM. Conforme sugerido pela Tabela 6.7, o comportamento é bastante simétrico para municípios com baixos valores de IDHM. Para valores mais elevados de IDHM, o gráfico adota uma tendência ascendente, indicando que quanto maior o valor de IDHM, maior o percentual de votos do PSL.

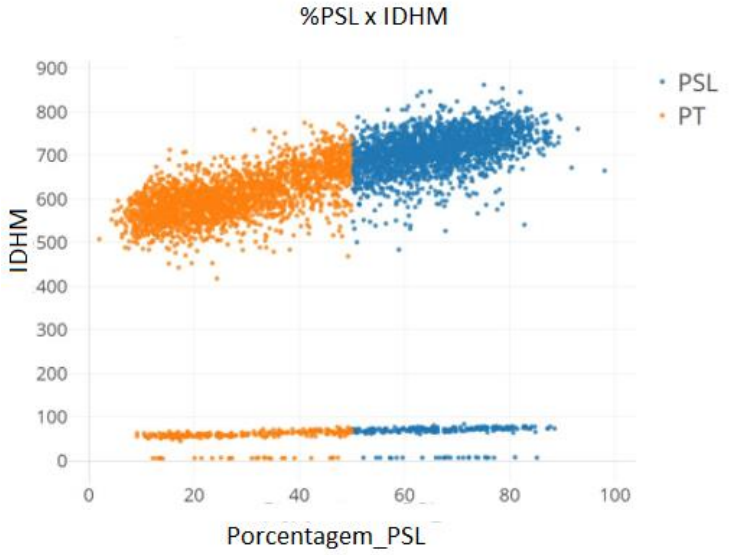


Figura 6.5. Porcentagem de votos PSL × IDHM.

6.3 Análises das Relações Existentes entre os Atributos Selecionados

A próxima etapa do trabalho foi selecionar os atributos relevantes, dentre aqueles que descrevem as instâncias de dados, a serem usados para induzir um modelo preditivo com base nas instâncias de dados descrita apenas por eles. Cinco atributos potenciais foram apresentados na seção anterior: (1) Tamanho da população, (2) $IDHM_{Educação}$, (3) $IDHM_{Renda}$, (4) $IDHM_{Longevidade}$ e (5) IDHM. Antes de abordar as medidas estatísticas, foi realizada uma análise intuitiva, com enfoque na relação entre cada um dos cinco atributos e o vencedor das eleições, em cada município. As figuras 6.6, 6.7, 6.8 e 6.9 mostram a relação considerando o tamanho da população e cada um dos índices de IDHM e os resultados eleitorais, respectivamente.

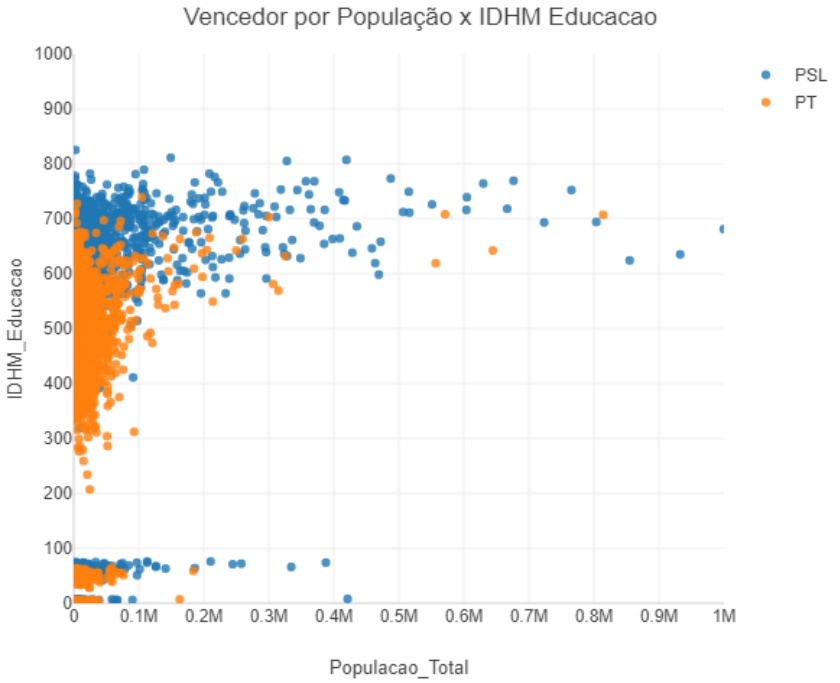


Figura 6.6 Vencedor por população e $IDHM_{Educação}$.

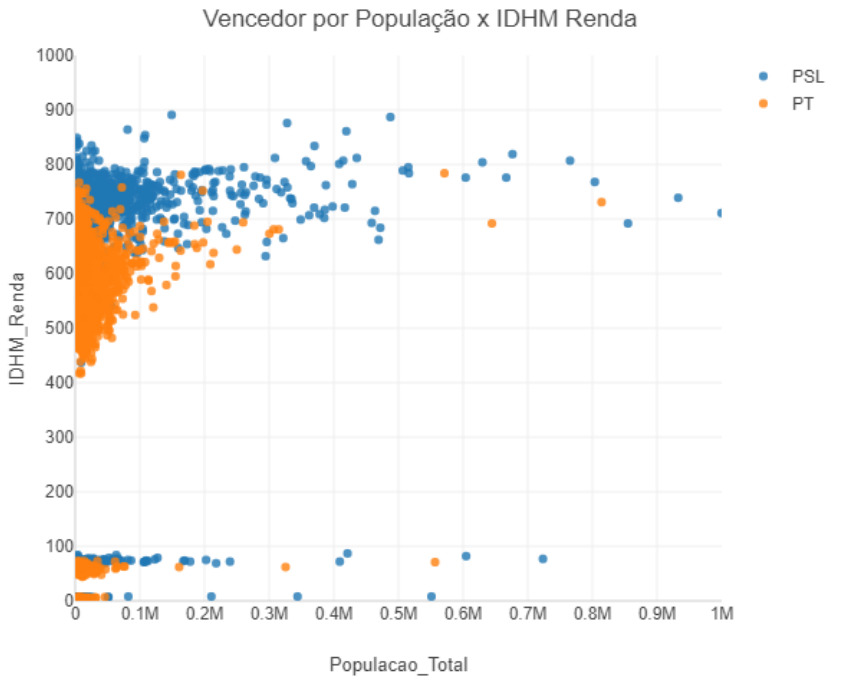


Figura 6.7 Vencedor por população e $IDHM_{Renda}$.

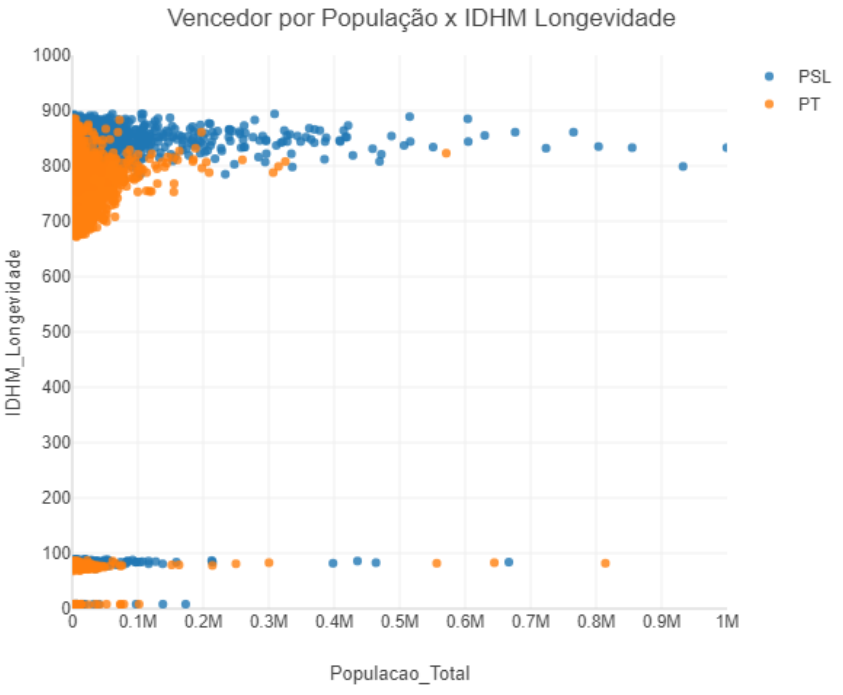


Figura 6.8 Vencedor por população e $IDHM_{Longevidade}$.

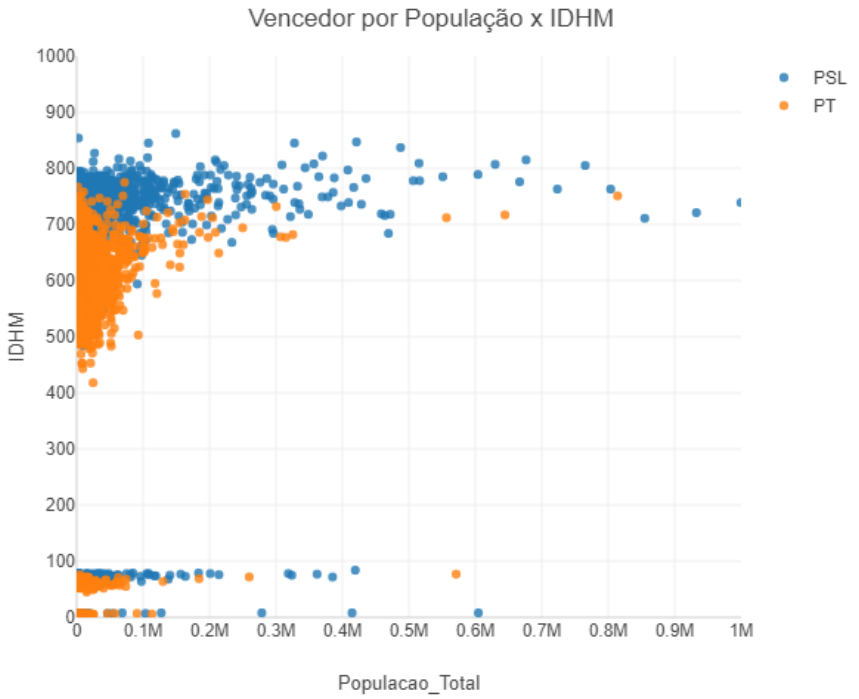


Figura 6.9 Vencedor por população e IDHM.

Os gráficos mostrados nas figuras anteriores sugerem que há pouca correlação entre o tamanho da população e qualquer um dos índices de IDHM, devido à forma como os dados se dispõem no plano cartesiano. Conforme mostrado na Seção 6.2, valores maiores para os índices de IDHM e para o tamanho da população parecem favorecer o PSL. A Figura 6.10 apresenta o vencedor das eleições em cada município relacionado a cada um dos 6 pares de índices de IDHM.

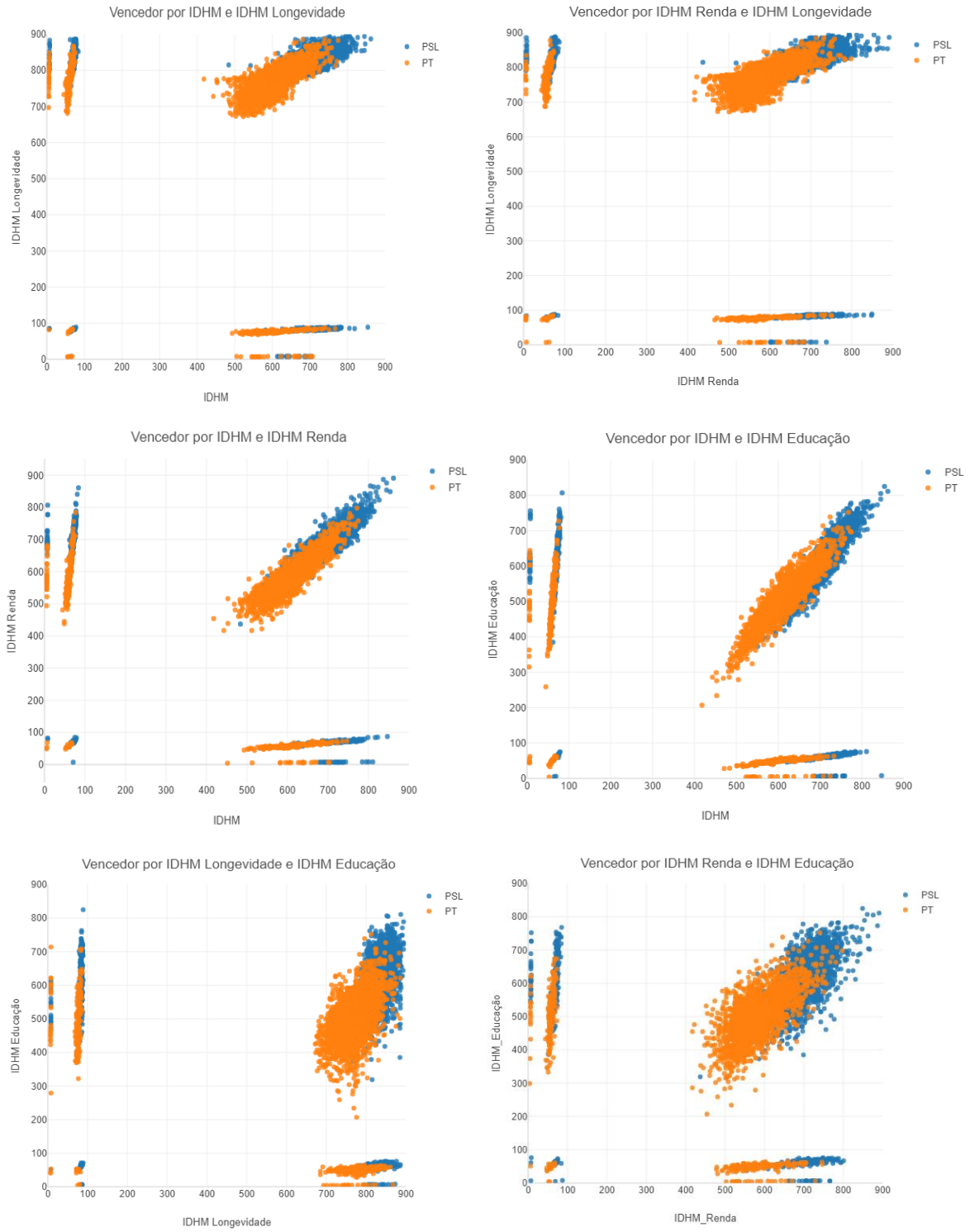


Figura 6.10. Demonstração do vencedor considerando todos os pares de índices de IDHM.

Os gráficos apresentados na Figura 6.10 confirmam a relação aparente entre o vencedor da eleição e cada um dos índices de IDHM, conforme previsto pela exploração inicial dos dados, descrita na Seção 6.2. Quanto mais alto o valor do índice, mais

provável é que o vencedor seja o PSL. Além disso, parece haver uma correlação entre cada par de índices, para a maioria dos municípios.

Cada gráfico da Figura 6.10 mostra um grande número de municípios em seu centro, com uma tendência clara de aumento, o que significa que o aumento em um índice geralmente implica em aumento no outro. Esse comportamento confirma a intuição de que, por exemplo, melhor educação deveria implicar em mais renda para o município. Há, no entanto, um número expressivo de municípios em cada gráfico, em que o valor de um índice é pequeno e o de outro é grande.

A Tabela 6.8 mostra a correlação estatística entre o IDHM, os subíndices de Educação, Renda e Longevidade, o tamanho da população e a porcentagem de votos obtidos pelo PT e PSL no município.

Tabela 6.8 - Correlação (Pearson) entre os atributos potenciais e a porcentagem de votos obtidos por partido.

Índices	IDHM	Educação	Renda	Long	População	%PSL	%PT
IDHM	1,00	0,15	0,12	0,02	0,05	0,26	-0,26
Educação	0,15	1,00	0,16	0,05	0,09	0,35	-0,35
Renda	0,12	0,16	1,00	0,06	0,04	0,33	-0,33
Longev.	0,02	0,05	0,06	1,00	0,02	0,13	-0,13
População	0,05	0,09	0,04	0,02	1,00	0,06	-0,06
PSL%	0,26	0,35	0,33	0,13	0,06	1,00	-1,00
PT%	-0,26	-0,35	-0,33	-0,13	-0,06	-1,00	1,00

As correlações para %PSL e %PT são exatamente opostas, pois quando um cresce o outro diminui. Confirmando algumas das informações intuitivas que derivaram dos gráficos apresentados na Seção 6.2, o subíndice de $IDHM_{Educação}$ tem o maior valor de correlação, seguido pelo de $IDHM_{Renda}$, pelo índice geral IDHM, de $IDHM_{Longevidade}$ e, finalmente, pelo tamanho da população.

Algumas informações interessantes se destacam na Tabela 6.8. Primeiro, o tamanho da população parece ter pouco impacto sobre o vencedor. Embora as análises anteriores tenham mostrado que para os grandes municípios, o voto tendia predominantemente a favor do PSL e para os pequenos e médios municípios o resultado era quase balanceado entre os dois.

Outro resultado relevante e um tanto surpreendente foi que os subíndices relacionados ao IDHM parecem ter pouca correlação entre eles. O maior valor de correlação encontrado ocorreu entre os subíndices de $IDHM_{Educação}$ e de $IDHM_{Renda}$, com

um valor de 0,16, o que é pouco significativo. O subíndice de $IDHM_{Longevidade}$ teve pouca correlação com qualquer um dos outros.

Para determinar a importância relativa de cada subíndice, foi usado o método do qui-quadrado, examinando cada um dos subíndices individualmente. A análise confirma que o subíndice de educação tem a maior influência no resultado, seguido por renda, IDHM, longevidade e, finalmente, tamanho da população, conforme mostrado na Tabela 6.9.

Tabela 6.9. Análise univariada de qui-quadrado.

	Valor Qui-Quadrado
$IDHM_{Educação}$	39,93
$IDHM_{Renda}$	35,79
IDHM	25,96
$IDHM_{Longevidade}$	4,31
População	2,14

Consistente com os resultados de correlação apresentados na Tabela 6.8, uma análise univariada baseada em qui-quadrado confirma que o subíndice de $IDHM_{Educação}$ teve a maior influência no resultado, seguido por $IDHM_{Renda}$, IDHM, $IDHM_{Longevidade}$ e, finalmente, tamanho da população.

6.4 Resultados Obtidos com o Uso de Algoritmos Supervisionados

Nesta seção serão apresentados os resultados obtidos com o uso de seis diferentes algoritmos de classificação de aprendizado de máquina supervisionado (ver Capítulo 3). Os algoritmos selecionados foram: (1) Decision Tree (DT), (2) K-Nearest Neighbors (k-NN), (3) Naive Bayes (NB), (4) Random Forest (RF), (5) SVC e (6) XGBoost [Duda *et al.*, 2000].

Cada um dos algoritmos foi utilizado como parte de um processo de k-validação cruzada (k=10), em que o conjunto alvo (no caso, o conjunto com 5.559 instâncias de dados descritas por cinco atributos numéricos *i.e.*, População, $IDHM_{Educação}$, $IDHM_{Renda}$, $IDHM_{Longevidade}$ e IDHM e uma classe associada (PT ou PSL)) é dividido em 10 partes, aproximadamente iguais, e um processo de treinamento-teste é conduzido 10 vezes, variando os conjuntos de treinamento e de teste. A variação é feita de maneira que o processo considera 9 dos conjuntos para o treinamento e o conjunto restante, como teste. Cada uma das 10 partes deve ser conjunto teste apenas uma vez.

Na Seção 6.3 experimentos iniciais com foco em relevância dos atributos que descrevem as instâncias de dados foram conduzidos, utilizando o método Qui-Quadrado.

Com base nos resultados obtidos, os cinco atributos, ordenados em ordem decrescente de relevância são: $IDHM_{Educação}$, $IDHM_{Renda}$, IDHM, $IDHM_{Longevidade}$ e População.

A Tabela 6.10 apresenta os resultados obtidos nos experimentos com algoritmos supervisionados de AM, utilizando o processo de 10-validação cruzada, considerando a seguinte metodologia. Os resultados apresentados na coluna Educação são relativos ao uso do conjunto de dados descrito apenas pelo atributo $IDHM_{Educação.}$ Os resultados apresentados na coluna $IDHM_{Renda}$, são relativos ao uso do conjunto de dados descrito pelos atributos: $IDHM_{Educação}$ e $IDHM_{Renda}$; Os resultados apresentados na coluna IDHM são relativos ao uso do conjunto de dados descrito pelos atributos: $IDHM_{Educação}$, $IDHM_{Renda,}$ IDHM. Os resultados apresentados na coluna $IDHM_{Longevidade}$, são relativos ao uso do conjunto de dados descrito pelos atributos: $IDHM_{Educação}$, $IDHM_{Renda,}$ IDHM, $IDHM_{Longevidade}$ e, os resultados apresentados na coluna População são relativos ao uso do conjunto de dados descrito pelos atributos: $IDHM_{Educação}$, $IDHM_{Renda,}$ IDHM e $IDHM_{Longevidade}$ e População.

A Tabela 6.10 apresenta os resultados obtidos por diferentes modelos, que foram induzidos usando as instâncias do conjunto alvo, descritas por subconjuntos distintos de atributos *i.e.*, os índices utilizados e, também, a população. Inicialmente modelos foram induzidos a partir de dados descritos apenas pelo subíndice $IDHM_{Educação}$ e, na sequência, foram gerados novos modelos a partir do conjunto de treinamento, progressivamente sendo descrito pela adição de um novo atributo ao conjunto anteriormente utilizado, obedecendo a ordem de relevância entre os atributos.

Tabela 6.10 - Precisão de cada modelo com diferentes combinações de atributos.

	Educação	+Renda	+IDHM	+Longevidade	+População
DT	78%	82%	82%	81%	81%
K-NN	77%	86%	86%	86%	75%
NB	76%	79%	81%	81%	67%
RF	78%	85%	86%	85%	86%
SVC	79%	87%	87%	85%	80%
XGBoost	80%	87%	87%	87%	87%

Os resultados mostram que os modelos criados tanto pelo SVC quanto pelo XGBoost foram capazes de prever corretamente o vencedor das eleições em 87% dos casos ao usar os subíndices de $IDHM_{Educação}$ e de $IDHM_{Renda}$ como variáveis preditoras. O resultado é razoavelmente bom para um algoritmo de classificação e evidencia o forte valor preditivo dos dois subíndices empregados. Os experimentos mostraram que a adição de

mais subíndices à descrição das instâncias de dados além deste ponto mantém o nível de precisão ou, então, o torna pior, indicando que o IDHM, o $IDHM_{Longevidade}$ e o tamanho da população não são preditores poderosos.

6.4.1 Considerações Sobre as Análises

Os resultados apresentados nas seções anteriores mostram que os eleitores de municípios com valores maiores para os subíndices de $IDHM_{Educação}$ e de $IDHM_{Renda}$, votaram de forma esmagadora no PSL, sendo o PT preferido nos municípios com valores mais baixos para estes índices.

Este trabalho de pesquisa não teve como objetivo considerar as influências que os panoramas político, econômico e sociológico tiveram no comportamento eleitoral. No entanto, os resultados aqui apresentados tendem a apoiar o conhecimento intuitivo de que nas eleições presidenciais de 2018 no Brasil, o PT foi o preferido pela parcela mais pobre e menos instruída da população, enquanto a parcela mais rica e instruída inclinou-se para o PSL.

A afirmação anterior é muitas vezes considerada como um conhecimento de senso comum, compartilhado por muitos da população e que foi, de certa forma, corroborada pelos resultados dos experimentos conduzidos, informando muito pouco sobre as preferências de um determinado indivíduo. Além disso, é importante notar que a análise se baseia nos dados disponíveis sobre a situação de desenvolvimento dos municípios brasileiros à época e, portanto, é suscetível a qualquer viés que esteja presente nos dados. De qualquer forma, acreditamos que os políticos em geral se beneficiariam com a compreensão dessas tendências e, com sorte, reconsiderariam seus focos e suas políticas, para melhor atingir uma parcela maior da população.

6.5 Resultados Obtidos com Uso do Algoritmo k-Means

Esta seção apresenta uma segunda abordagem à análise dos dados eleitorais utilizados neste trabalho, por meio do uso do algoritmo de agrupamento conhecido como k-Means (ver Seção 3.4). A motivação para a segunda abordagem foi a de usar um algoritmo de agrupamento para analisar a organização das instâncias de dados eleitorais, refletida no agrupamento induzido pelo algoritmo e, assim, poder contrapor os resultados de ambos os experimentos: o descrito na Seção 6.5 e o apresentado nesta.

Os experimentos foram conduzidos usando o conjunto de dados mostrado na Tabela 6.2. A metodologia adotada para a condução dos experimentos foi a de filtrar o conjunto de dados, descrevendo-os por meio de apenas um dos atributos discutidos anteriormente. Assim, foram produzidas cinco versões do conjunto original e, em cada uma das versões, o conjunto dos dados era descrito por: (1) índice de IDHM, (2) $IDHM_{Educação}$, (3) $IDHM_{Renda}$, (4) $IDHM_{Longevidade}$, e (5) população e, então, cada um dos cinco conjuntos foi entrada para o algoritmo k-Means, com a informação adicional de k=3. Foram adotados os rótulos Baixo, Médio e Alto para nomear os k=3 grupos participantes do agrupamento induzido em cada experimento.

Os resultados obtidos (*i.e.*, os cinco agrupamentos, cada um constituído de três grupos) foram então avaliados pelo uso do índice de validação interna Silhouette (ver Seção 3.5), cujos possíveis valores variam no intervalo [−1 +1] sendo que quanto mais próximo do valor + 1 for o valor do índice referente a um determinado agrupamento, melhor é o agrupamento. Lembrando, tal índice busca evidenciar o quão semelhantes são as instâncias de um mesmo grupo quando comparadas com as instâncias dos demais grupos.

Na sequência, cada um dos agrupamentos obtidos foi também avaliado por meio do cálculo do índice de validação externa Rand (ver Seção 3.5), cujos valores variam no intervalo [0 1]. Como discutido na Seção 3.5, para o cálculo do índice Rand são necessários dois agrupamentos. Para cada um dos cinco cálculos, um dos agrupamentos utilizados foi aquele induzido pelo k-Means (k=3), com dados descritos por um dos cinco atributos e o outro foi o agrupamento induzido pelo k-Means (k=2) em que os dados de entrada são descritos apenas pelo valor do atributo referente ao partido vencedor (*i.e.*, PT ou PSL). Note que esse último agrupamento é induzido apenas uma vez e é utilizado para o cálculo do índice Rand associado à avaliação de cada um dos agrupamentos induzidos referentes a cada um dos cinco atributos considerados.

Os resultados dos experimentos, assim como os resultados da avalição de cada experimento, em que o agrupamento resultante é avaliado por meio dos valores dos dois índices, Silhouette e Rand, nos agrupamentos induzidos, são apresentados nas próximas tabelas.

A Tabela 6.11 apresenta o resultado do agrupamento utilizando o atributo $IDHM_{Educação}$. Do total de 5.559 municípios brasileiros, 10% deles estão no grupo com

$IDHM_{Educação}$ Baixo, outros 45% estão no grupo com $IDHM_{Educação}$ Médio, e os 45% restantes estão no grupo com $IDHM_{Educação}$ Alto. O resultado mostra uma divisão equilibrada entre o número de municípios com $IDHM_{Educação}$ Médio e $IDHM_{Educação}$ Alto. Do total de 2.507 municípios no grupo com $IDHM_{Educação}$ Médio, em 80% deles o PT teve mais votos, enquanto o PSL foi preferido em apenas 20%. Em 2.527 municípios no grupo $IDHM_{Educação}$ Alto, os valores se invertem, com o PSL recebendo 79% dos votos e o PT recebendo 21%.

Tabela 6.11 – Resultado do agrupamento induzido, contendo três grupos, caracterizados como Baixo, Médio e Alto, em função dos valores do único atributo $IDHM_{Educação}$ que descreve as instâncias de dados de entrada para o algoritmo. As colunas PSL e PT informam o número de municípios em que o PSL e o PT foram vencedores, respectivamente, em cada um dos três grupos induzidos.

Grupos do agrupamento induzido por $IDHM_{Educação}$	PSL	% PSL	PT	% PT	centroides	PSL+PT	Índice Silhouette	Índice Rand
Baixo	256	49	269	51	52,27	525	0,7659	0,6463
Médio	492	20	2.015	80	482,38	2.507		
Alto	2.006	79	521	21	636,01	2.527		

A Tabela 6.12 apresenta o resultado do agrupamento utilizando o atributo $IDHM_{Renda}$. Do total de 5.559 municípios brasileiros, 10% deles estão no grupo com $IDHM_{Renda}$ Baixo, 40% estão no grupo com $IDHM_{Renda}$ Médio, e 50% estão no grupo com $IDHM_{Renda}$ Alto. O resultado mostra uma pequena diferença entre o número de municípios com $IDHM_{Renda}$ Médio e $IDHM_{Renda}$ Alto. Do total de 2.207 municípios no grupo com $IDHM_{Renda}$ Médio, em 93% deles o PT teve mais votos, enquanto o PSL foi preferido em apenas 7%. Em 2.802 municípios no grupo $IDHM_{Renda}$ Alto, os valores se mostram favoráveis ao PSL que recebeu 83% dos votos enquanto o PT recebeu 17%.

Tabela 6.12 – Resultado do agrupamento induzido, contendo três grupos, caracterizados como Baixo, Médio e Alto, em função dos valores do único atributo $IDHM_{Renda}$ que descreve as instâncias, de dados de entrada para o algoritmo. As colunas PSL e PT informam as colunas PSL e PT informam o número de municípios em que o PSL e o PT foram vencedores, respectivamente, em cada um dos três grupos induzidos.

Grupos do agrupamento induzido por $IDHM_{Renda}$	PSL	% PSL	PT	% PT	centroides	PSL+PT	Índice Silhouette	Índice Rand
Baixo	263	48	287	52	58**,**87	550	0,8259	0,7300
Médio	154	7	2.053	93	564**,**93	2.207		
Alto	2.337	83	465	17	705	2.802		

A Tabela 6.13 apresenta o resultado do agrupamento utilizando o atributo $IDHM_{Longevidade}$. Do total de 5.559 municípios brasileiros, 10% deles estão no grupo com $IDHM_{Longevidade}$ Baixo, 34% estão no grupo com $IDHM_{Longevidade}$ Médio, e 56% estão no grupo com $IDHM_{Longevidade}$ Alto. Comparado com o resultado dos atributos anteriores, o resultado para $IDHM_{Longevidade}$ mostra uma diferença significativa entre o número de municípios com $IDHM_{Longevidade}$ Médio e Alto. Do total de 1.888 municípios no grupo com $IDHM_{Longevidade}$ Médio, em 94% deles o PT teve mais votos, enquanto o PSL foi preferido em apenas 6%. Em 3.137 municípios no grupo $IDHM_{Longevidade}$ Alto, os valores se invertem para 75% e 25%.

Tabela 6.13 - Resultado do agrupamento induzido, contendo três grupos, caracterizados como Baixo, Médio e Alto, em função dos valores do único atributo $IDHM_{Longevidade}$ que descreve as instâncias, de dados de entrada para o algoritmo. As colunas PSL e PT informam as colunas PSL e PT informam o número de municípios em que o PSL e o PT foram vencedores, respectivamente, em cada um dos três grupos induzidos.

Grupos do agrupamento induzido por $IDHM_{Longevidade}$	PSL	% PSL	PT	% PT	centroides	PSL+PT	Índice Silhouette	Índice Rand
Baixo	287	54	247	46	74,01	534	0,7959	0,6700
Médio	111	6	1.777	94	753,26	1.888		
Alto	2.356	75	781	25	830,23	3.137		

Por fim, a Tabela 6.14 mostra o resultado do agrupamento utilizando o atributo IDHM. Do total de 5.559 municípios brasileiros, 10% deles estão no grupo com IDHM Baixo, outros 41% estão no grupo com IDHM Médio, e 49% estão no grupo com IDHM Alto. Semelhante ao $IDHM_{Renda}$ o resultado do IDHM também mostra uma pequena

diferença entre o número de municípios com IDHM Médio e IDHM Alto. Do total de 2.246 municípios no grupo com IDHM Médio, o PT foi o preferido obtendo 90% dos votos, enquanto o PSL obteve apenas 10%. Em 2.746 municípios no grupo IDHM Alto, os resultados mostram novamente uma inversão de valores, relativos aos obtidos para IDHM Médio, onde 82% preferiram o PSL enquanto 18% preferiram o PT.

Tabela 6.14 - Resultado do agrupamento induzido, contendo três grupos, caracterizados como Baixo, Médio e Alto, em função dos valores do único atributo IDHM que descreve as instâncias, de dados de entrada para o algoritmo. As colunas PSL e PT informam as colunas PSL e PT informam o número de municípios em que o PSL e o PT foram vencedores, respectivamente, em cada um dos três grupos induzidos.

Grupos do agrupamento induzido por IDHM	PSL	% PSL	PT	% PT	centroides	PSL+PT	Índice Silhouette	Índice Rand
Baixo	279	49	288	51	59,86	567	0,8125	0,7052
Médio	222	10	2.024	90	591,01	2.246		
Alto	2.253	82	493	18	714,94	2.746		

Para os experimentos realizados utilizando o atributo População, foram considerados subconjuntos de municípios diferentes, como segue: (1) todos os municípios, (2) municípios com população maior ou igual a 1.000.000 de habitantes, (3) municípios com população menor que 1.000.000 de habitantes, (4) municípios com população inferior a 500.000 habitantes, (5) municípios com população inferior a 100.000 habitantes e (6) municípios com população inferior a 50.000 habitantes. O objetivo dos experimentos referentes ao atributo População foi inspecionar os valores de ambos os índices de validação, relativos aos vários volumes de dados considerados.

A Tabela 6.15 apresenta os resultados dos agrupamentos induzidos utilizando o atributo População, contendo três grupos caracterizados como Pequeno, Médio e Grande. Para o agrupamento induzido considerando o conjunto com todos os municípios, 99,6% deles estão no grupo de municípios pequenos, enquanto 0,4% estão divididos entre Médio e Grande. Um percentual semelhante também é observado para os demais subconjuntos de municípios considerados. Note na Tabela 6.15 que os rótulos Pequeno, Médio e Grande caracterizam grupos do agrupamento associados aos rótulos, levando em conta o tamanho da população considerado, ou seja, a semântica das palavras Pequeno, Médio e Grande está contextualizada ao tamanho da população.

Tabela 6.15 – Resultado do agrupamento induzido, contendo três grupos, caracterizados como Pequeno, Médio e Grande, em função dos valores do único atributo População, considerando vários subconjuntos de municípios definidos pelo correspondente população. As colunas PSL e PT informam as colunas PSL e PT informam o número de municípios em que o PSL e o PT foram vencedores, respectivamente, em cada um dos três grupos induzidos.

Grupos do agrupamento induzido por População	PSL	PT	centroides	PSL+PT	Índice Silhouette	Índice Rand
SOMA DA POPULAÇÃO DE TODOS OS MUNICÍPIOS = 5.559						
Pequeno	2.735	2.800	25.806,27	5.535	0,9938	0,50
Médio	17	5	1.366.544,63	22		
Grande	2	0	8.786.974,50	2		
TOTAL	2.754	2.805	–	5.559		
SOMA DA POPULAÇÃO DE MUNÍCIPIOS COM POPULAÇÃO ≥ 1.000.000 = 15						
Pequeno	9	4	1.737.419,76	13	0,7660	0,5022
Médio	1	0	6.320.446	1		
Grande	1	0	11.253.503	1		
TOTAL	11	4	–	15		
SOMA DA POPULAÇÃO DE MUNÍCIPIOS COM POPULAÇÃO < 1.000.000 = 5.544						
Pequeno	2.508	2.740	15.915,34	5.248	0,9453	0,5022
Médio	192	56	164.448,25	248		
Grande	43	5	541.840,97	48		
TOTAL	2.743	2.801	–	5.544		
SOMA DA POPULAÇÃO DE MUNÍCIPIOS COM POPULAÇÃO < 500.000 = 5.521						
Pequeno	2.352	2.604	12.947,09	4.956	0,9092	0,5024
Médio	87	14	87.984,46	101		
Grande	285	179	293.203,60	464		
TOTAL	2.467	2.797	–	5.521		
SOMA DA POPULAÇÃO DE MUNÍCIPIOS COM POPULAÇÃO < 100.000 = 5.277						
Pequeno	1.777	1.922	7.761,39	3.699	0,8034	0,5002
Médio	559	667	27.784,94	1.226		
Grande	193	159	67.228,58	352		
TOTAL	2.529	2.748	–	5.277		
SOMA DA POPULAÇÃO DE MUNÍCIPIOS COM POPULAÇÃO < 50.000 = 4.951						
Pequeno	1.481	1.502	6.042,29	2.983	0,7858	0,5008
Médio	584	817	18.217,41	1.401		
Grande	284	283	35.785,46	567		
TOTAL	2.349	2.602	–	4.951		

De forma geral, é possível observar que nos resultados de todos os experimentos mostrados na Tabela 6.15, uma quantidade maior de municípios pertence ao grupo caracterizado como Pequeno e que, nesses municípios, o PT obteve a preferência de votos na maioria deles enquanto a preferência pelo PSL está concentrada nos grupos Médio e Grande.

Uma variação da preferência de votos é percebida quando os experimentos são realizados com municípios em que a população é inferior a 100.000 habitantes. Nesses casos o PT também apresenta certa preferência entre os municípios caracterizados como Médio e Grande. Porém, quando os experimentos são realizados apenas com municípios em que a população é superior a 1.000.000 de habitantes, o PSL apresenta preferência majoritária, tanto para os municípios caracterizados como Pequeno quanto para Médio e Grande.

6.5.1 Resultados Obtidos com o Uso dos Índices de Validação

Considerando os resultados dos experimentos mostrados nas tabelas anteriores, as Tabela 6.16 e Tabela 6.17 apresentam um resumo das avaliações dos agrupamentos feitas com o uso dos índices Silhouette e Rand.

Tabela 6.16 Resultado da avaliação dos agrupamentos induzidos para cada atributo: IDHM, $IDHM_{Renda}$, $IDHM_{Longevidade}$, $IDHM_{Educação}$ e População utilizando o índice de validação interna Silhouette.

Atributo	Índice Silhouette
$IDHM_{Renda}$	0,8259
IDHM	0,8125
$IDHM_{Longevidade}$	0,7959
$IDHM_{Educação}$	0,7659
Atributo População	
todos os municípios	0,9938
< 1.000.000	0,9453
< 500.000	0,9092
< 100.000	0,8034
< 50.000	0,7858
≥ 1.000.000	0,7660

Considerando os valores do índice Silhouette mostrados na Tabela 6.16, o agrupamento induzido para o atributo População, considerando todos os municípios, obteve a melhor avaliação entre os experimentos. Entre os índices de IDHM, o atributo $IDHM_{Renda}$ é o que obteve o melhor resultado, seguido por IDHM, $IDHM_{Longevidade}$ e $IDHM_{Educação}$.

Tabela 6.17 Resultado da avaliação dos agrupamentos induzidos para cada atributo de IDHM, $IDHM_{Renda}$, $IDHM_{Longevidade}$, $IDHM_{Educação}$ e População utilizando o Índice Rand.

Atributo	Índice Rand
$IDHM_{Renda}$	0,7300
IDHM	0,7052
$IDHM_{Longevidade}$	0,6700
$IDHM_{Educação}$	0,6463
Atributo População	
< 500.000	0,5024
< 1.000.000	0,5022
≥ 1.000.000	0,5022
< 50.000	0,5008
< 100.000	0,5002
todos os municípios	0,5000

Os resultados da Tabela 6.17 mostram que o agrupamento induzido com os dados descritos pelo atributo $IDHM_{Renda}$, obteve o valor mais alto de índice Rand, o que significa que o agrupamento induzido apresenta uma maior similaridade com relação ao agrupamento de referência (isto é, ao agrupamento induzido usando o mesmo conjunto de municípios, cada um descrito pelo atributo cujo valor representa o partido vencedor das eleições no referido município). Esse fato caracteriza o atributo $IDHM_{Renda}$ como o atributo que particiona o conjunto de municípios de uma maneira que a partição obtida é a que mais se assemelha à partição induzida pelo atributo que caracteriza o partido vencedor. Portanto, os agrupamentos induzidos pelos atributos utilizados, ordenados por valor de índice Rand, do maior valor para o menor valor são: $IDHM_{Renda}$, IDHM, $IDHM_{Longevidade}$, $IDHM_{Educação}$ e População.

6.5.2 Comentários sobre os Resultados dos Experimentos e Considerações Finais

Ao serem comparados aos resultados apresentados nas seções anteriores (Seção 6.2, Seção 6.3 e Seção 6.4), os resultados obtidos com o uso do k-Means nos dois conjuntos de experimentos confirmam a tendência de municípios que apresentam menores índices de IDHM votarem no PT enquanto municípios com índices maiores de IDHM votarem no PSL. Essa tendência também se confirma em relação à População *i.e.*, em municípios com maior número de habitantes a preferência de voto é predominante do PSL e em municípios com menor número de habitantes a preferência de voto é para o PT. A Tabela 6.18 apresenta uma comparação entre os resultados das análises estatísticas anteriores (Seção 6.2) e os resultados obtidos com o uso do k-Means para o atributo População (Seção 6.5). É possível perceber com base nos dados apresentados na tabela, que as duas metodologias utilizadas nas análises apresentam resultados similares.

Tabela 6.18 – Comparação entre os resultados estatísticos anteriores (Seção 6.2) e os resultados obtido com o uso do algoritmo de agrupamento k-Means (Seção 6.5), considerando o atributo População.

Atributo População	Análises estatísticas anteriores		Análises com resultados do k-Means	
Grupo	%PSL	%PT	%PSL	%PT
Pequeno	48	52	49,4	50,6
Médio	80	20	77	23
Grande	73	27	100	0

Como trabalhos futuros pretende-se expandir os experimentos descritos anteriormente por meio de novos experimentos envolvendo descrições de instâncias com vários atributos e, também, da normalização de valores de atributos. Também pretende-se fazer uso de outros algoritmos de agrupamento, tal como o DBScan [Ester et al 1996], para análises comparativas com os resultados obtidos com o k-Means.

Com relação aos dados pretende-se, por meio da experiência adquirida neste trabalho, focalizar o panorama eleitoral por regiões do país, com o objetivo de identificar tendências eleitorais, por região.

6.6 Conclusões

A literatura científica na área desse trabalho disponibiliza um grande volume de estudos e experimentações que foram conduzidos a respeito da previsão de resultados eleitorais. Embora existam alguns trabalhos que usam características difíceis de serem aferidas como traços não-verbais (como a "confiabilidade inerente" de um candidato) como preditores de resultados eleitorais, a maior parte dos trabalhos se concentrou no uso de *feeds* do *Twitter* como a principal fonte de dados para os modelos estatísticos aplicados. Uma infinidade de abordagens diferentes foi experimentada, desde simplesmente usar contagens de *tweet* que mencionam um candidato ou partido político, até abordagens mais sofisticadas usando análise de sentimento e outras métricas mais complexas.

Abordagens baseadas no *Twitter* tiveram resultados promissores coexistindo com outros decepcionantes. Podemos afirmar que parte da questão subjacente às abordagens baseadas no *Twitter* é o viés inerente de usar a população do *Twitter* como representante de todos os eleitores. O uso de robôs do *Twitter* para postar automaticamente milhões de mensagens em nome de um determinado candidato ou partido só piora as coisas. Além disso, a análise de sentimento, que faz parte da maioria das soluções apresentadas, é um desafio para ser ajustada usando *tweets* de 140 caracteres como entrada.

Este trabalho de pesquisa investigou empiricamente o impacto do IDHM nos resultados do segundo turno das eleições presidenciais brasileiras de 2018. Conforme comentado nos capítulos anteriores o IDHM é um índice que visa condensar o nível de desenvolvimento de um município em um único número e é composto por três subíndices: $IDHM_{Educação}$, $IDHM_{Renda}$ e $IDHM_{Longevidade}$, que avaliam de forma independente a situação do município em cada uma das perspectivas representadas por esses três subíndices. É importante lembrar que o IDHM e todos os seus subíndices foram calculados com base nos dados do Censo brasileiro, que estão publicamente disponíveis.

É possível afirmar que, ao usar os dados do Censo como a principal fonte para as variáveis preditoras, o impacto do viés nos preditores utilizados fica limitado, uma vez que os dados do Censo coletam as informações mais precisas disponíveis de cada eleitor potencial no país. Para atingir um dos objetivos desse trabalho de pesquisa, foram

empregados seis algoritmos supervisionados de aprendizado de máquina diferentes para prever o vencedor da eleição em cada município e um algoritmo de aprendizado não supervisionado. Como potenciais preditores, foram utilizados o índice geral de IDHM, cada um de seus três subíndices Educação, Renda e Longevidade e o tamanho da população do município. A análise estatística dos dados revelou que os subíndices, $IDHM_{Educação}$ e $IDHM_{Renda}$, nesta ordem, apresentaram a correlação com a variável alvo vencedor da eleição no município.

Dos algoritmos de classificação utilizados, o SVC e o XGBoost apresentaram os melhores resultados ao preverem corretamente o vencedor da eleição em 87% dos municípios ao utilizarem as instâncias de dados descritas apenas com os atributos de $IDHM_{Educação}$ e de $IDHM_{Renda.}$ Os experimentos evidenciaram que a extensão da descrição das instâncias de dados, com a adição de mais atributos não colaboraram para a indução de um modelo com maior precisão de classificação.

A interpretação dos resultados obtidos é bastante clara: é possível desenvolver um modelo de aprendizado de máquina para prever o candidato presidencial preferido em cada município do Brasil, a partir de um conjunto de dados descritos usando a Escolaridade e a Renda como atributos que descrevem as instâncias de dados. Essa previsão pode ser feita com 87% de acerto e expõe o fato de eleitores em municípios com maiores valores para os atributos Educação e Renda preferirem o candidato do PSL, enquanto aqueles em municípios com valores mais baixos associados à Educação e Renda, tenderam para o PT. Esses resultados foram reforçados com a utilização do algoritmo de aprendizado não supervisionado, onde os agrupamentos evidenciam que em municípios com valores menores para os atributos relacionados ao IDHM e seus subindices ou com populações menores preferiram o PT, enquanto os municípios com valores maiores para os atributos relacionados a IDHM e seus subíndices ou com populações maiores tenderam para o PSL.

Embora as razões para esse comportamento sejam certamente complexas e fora do escopo deste trabalho, os resultados parecem confirmar o senso comum no Brasil, no que diz respeito ao perfil dos eleitores do PT e do PSL.

Caso alguém tenha interesse nos dados utilizados na pesquisa, entrar em contato pelo email ncesar_and@hotmail.com.

Referências

[Ahmed *et al*. 2016] Ahmed, S.; Jaidka, K.; Cho, J. (2016) The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties, Telematics and Informatics, v. 33, no. 4, pp. 1071–1087.

[Alpaydin 2010] Alpaydin, E. (2010) Introduction to machine learning. 2. Ed. Cambridge: MIT Press, 537pgs.

[An *et al.* 2012] An, Y.; Khare, R.; Song, I.-Y.; Hu, X. (2012) Data exploration and knowledge discovery in a patient wellness tracking (PWT) system at a nurse-managed health services center, In: Proc. of The ACM International Health Informatics Symposium (IHI 2012), USA, pp. 661−665.

[Antonakis & Dalgas 2009] Antonakis, J.; Dalgas, O. (2009) Predicting elections: Child's play! Science, 323(5918):1183–1183

[Arlot, & Celisse, 2010] Arlot, S. & Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys Online Journal*, no. 4, pp. 40–79.

[AtlasBrasil 2020] Atlas Brasil, http://atlasbrasil.org.br, acessado em: 01/01/2021.

[Bansala & Srivastava 2018] Bansala, B.; Srivastava, S. (2018) On predicting elections with hybrid topic based sentiment analysis of tweets, *Procedia Computer Science.* v. 135, pp. 346–353.

[Barbera & Rivero 2015] Barbera P.; Rivero, G. (2015) Understanding the political representativeness of Twitter users, *Social Science Computer Review*, v. 33, no. 6, pp. 712–729.

[Barnett & Lewis 1994] Barnett, V.; Lewis, T. (1994) Outliers in statistical data. *John Wiley and Sons*.

[Berkhin 2006] Berkhin P. (2006) A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg.

[Bermingham & Smeaton 2011] Bermingham, A.; Smeaton, A. (2011) On using twitter to monitor political sentiment and predict election results. In Proceedings of the

Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pages 2–10.

[Bessi & Ferrara 2016] Bessi, A.; Ferrara, E. (2016) Social bots distort the 2016 us presidential election online discussion. First Monday, 21(11-7).

[Blum & Mitchell 1998] Blum, A.; Mitchell, T. M. (1998) Combining labeled and unlabed data with co-training, In: Proc. of the *Annual Conference on Learning Theory* (COLT 1998), pp. 92-100.

[Boutet *et al*. 2012] Boutet, A., Kim, H., & Yoneki, E. (2012) What's in your tweets? i know who you supported in the uk 2010 general election. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[Brender 2003] Brender, A. (2003) The effect of fiscal performance on local government election results in israel: 1989–1998. *Journal of Public Economics*, 87(9-10):2187–2205

[Cantu & Saiegh 2011] Cantu, F.; Saiegh, S. M. (2011) Fraudulent democracy? an analysis of argentina's infamous decade using supervised machine learning. Political Analysis, 19(4):409–433.

[Castro & Ferrari 2016] Castro, L.N., Ferrari D.G. (2016) Introdução a Mineração de Dados, Saraiva, 351pgs.

[Coletto *et al*. 2015] Coletto, M., Lucchese, C., Orlando, S., & Perego R. (2015) Electoral predictions with twitter: a machine-learning approach.

[Davis *et al*. 2016] Davis, C.A., Varol, O., Ferrara, E., Flammini A., & Menczer F. (2016) "BotOrNot: A system to evaluate social bots," Developers Day Workshop at World Wide Web Conference (Montreal); version at https://arxiv.org/abs/1602.00975, acessado em: 15/10/2020.

[Dicionário Financeiro] Dicionário Financeiro, https://www.dicionariofinanceiro.com/paridade-poder-compra/, acessado em: 15/12/2020

[Diffen 2020] Diffen (2020) Data vs Information, acessado em: 20/06/2020, https://www.diffen.com/difference/Data_vs_Information.

[Dua & Graff 2019] Dua, D.; Graff, C. (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[Duda *et. al* 2000] Duda, R. O., Hart, P. E., & Stork, D. G. (2000) Pattern Classification, Wiley Publ. 2000.

[Ester *et al* 1996] Ester, M., Kriegel, H. P., Sander, J. & Xu, X. (1996) Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-96 Proceedings. 226-231.

[Everitt et al. 2011] Everitt, B. S.; Landau, S.; Leese, M.; Stahl, D. (2011) Cluster Analysis, 5th Edition, Wiley Series in Probability and Statistics.

[Exploratory 2020] Exploratory (2020), Exploratory, Inc. https://exploratory.io, acessado em: 05/07/2020.

[Facelli *et al*. 2011] Facelli, K., Lorena, A.C., Gama, J., Carvalho, A.C.P.L.F (2011) Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina, LTC, 396pgs.

[Fayyad *et al*. 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, v. 39, pp. 27-34.

[FJP 2020] Fundação João Pinheiro. http://novosite.fjp.mg.gov.br/. acessado em: 10/07/2020.

[Freund & Schapire 1996] Freund, Y.; Schapire, R. E. (1996) Experiments with a New Boosting Algorithm. In International Conference on Machine Learning, pages 148–156.

[Gayo-Avello *et al*. 2011] Gayo-Avello, D.; Metaxas, P. T.; Mustafaraj, E. (2011) Limits of electoral predictions using Twitter, In: Proceedings of the *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 490–493.

[Ginsberg 1993] Ginsberg, M. L. (1993) Essential of artificial intelligence. Morgan Kaufman, San Mateo, Calif, 430pgs.

[Goldstein & Rainey 2010] Goldstein, P.; Rainey, J. (2010) The 2010 elections: Twitter isn’t a very reliable prediction tool. http://lat.ms/fSXqZW.

[Gowda & Diday 1992] Gowda, K. C.; Diday, E. (1992) Symbolic clustering using a new similarity measure, IEEE Transactions on Systems, Man, and Cybernetics, V. 22, no. 2, pp. 368−378.

[Han *et. al*. 2000] Han, J.; Kamber, M.; Pei J. (2000) Data Mining: Concepts and Techniques.

[Halkidi *et al.* 2001] Halkid, M, Batistakis, Y., Vazirgiannis, M. (2001) On Clustering Validation Techniques. Journal of Intelligent Information Systems. p. 107-145.

[Hoffmann 1998] Hoffmann, A. G. (1998) Paradigms of Artificial Intelligence – A Methodological & Computational Analysis, Springer-Verlag Singapore Pte. Ltd.

[IBGE 2020] Instituto Brasileiro de Geografia e Estatística. https://www.ibge.gov.br/. acessado em: 06/11/2020.

[Imperva 2020] https://www.imperva.com/blog/bot-traffic-report-2016/. acessado em: 12/10/2020.

[IPEA 2020] Instituto de Pesquisas Econômicas Aplicadas. https://www.ipea.gov.br/portal/, acessado em: 15/09/2020.

[Jain *et al.* 1999] Jain, A. K.; Murty, M. N.; Flynn, P. J. (1999) Data clustering: a review, *ACM Computing Surveys*, v. 31, no. 3, pp. 264-323.

[Jain 2010] Jain, A. K. (2010) Data clustering: 50 years beyond K-means. Pattern Recognition Letters, v. 31, pp. 651–666.

[Jain & Dubes 1988] Jain, A.K. & Dubes, R.C. (1988) Algorithms for Clustering Data. Prentice Hall.

[Jain & Katkar 2015] Jain, A. P.; Katkar, V. D. (2015) Sentiments analysis of Twitter data using data mining, *2015 Int. Conf. Inf. Process*., pp. 807–810.

[Kanungo *et al*. 2002] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. (2002) An Efficient k-Means Clustering Algorithm : Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 24, pp. 881–892.

[Kaufman & Rousseeuw 2005] Kaufman, L. and Rousseeuw, P.J. (2005) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, Hoboken.

[Kejriwal & Szekely 2007] Kejriwal, M.; Szekely, P. (2007) Data mining in unusual domains with information-rich knowledge graph construction, inference and search, Talk presented at The KDD 2007.

[Kim *et al.* 2007] Kim, H., Adolphs, R., O'Doherty, J.P., Shimojo, S. (2007). Temporal isolation of neural processes underlying face preference decisions. Proc. Natl. Acad. Sci. 104 (46), 18253–18258.

[Korakakis *et al*. 2017] Korakakis, M.; Spyrou, E.; P. Mylonas, P. (2006) A survey on political event analysis in Twitter, In: Proc. of The *12th Int. Work. Semant. Soc. Media Adapt*. Pers. SMAP 2017, pp. 14–19.

[Kovács *et al.* 2001] Kovács, F., Legány, C. and Babos, A. (2001) Cluster Validity Measurement Techniques. Technical report, Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary, pp-107-145.

[Lei do Acesso 2020] Lei do Acesso à Informação - http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, acessado em: 20/07/2020.

[McCarthy 2007] McCarthy, J. (2007) What is Artificial Intelligence? Stanford University. http://www-formal.stanford.edu/jmc/whatisai/whatisai.html, acessado em: 30/07/2020.

[MacQueen 1967] MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, v. 1, no. 14, pp. 281−297.

[Mangin 2007] Mangin, T. (2007) ngram: Textcat implementation in python. http://thomas.mangin.me.uk/. acessado em: 20/08/2020.

[Matte 2020] Matte, M. K. (2020) Impacto do Uso da Desigualdade Triangular para Acelerar o Algoritmo k-Means (Dissertação de Mestrado, Centro Universitário Campo Limpo Paulista).

[Mattes & Milazzo 2014] Mattes, K.; Milazzo, C. (2014) Pretty faces, marginal races: Predicting election outcomes using trait assessments of british parliamentary candidates. *Electoral Studies*, 34:177 – 189.

[Mellon & Prosser 2016] Mellon, J., Prosser, C. (2016) Twitter and Facebook are not representative of the general population: Political attitudes and demographics of social media users.

[Miranda *et al.* 2015] Miranda F. R.; Almeida, J. M. & Pappa, G. L. (2015) Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM) (pp. 1254-1261). IEEE.

[Mitchell 1997] Mitchell, T. M. (1997) Machine Learning. McGraw-Hill Education.

[Monard & Baranauskas 2003] Monard, M. C.; Baranauskas, J. A. (2003) Conceitos de aprendizado de máquina. In. Rezende (2003), p. 89-114.

[Newell & Simon 1976]. Newell, A.; Simon, H. A. (1976) Computer science as empirical inquiry: symbols and search, Communications of the ACM, v. 19, no. 3, pp. 113-126.

[Nguyen *et al.* 2014] Nguyen, D.; Trieschnigg, D.; Meder, T. (2014) Tweetgenie: development, evaluation, and lessons learned. In: Proceedings of the *25th International Conference on Computational Linguistics*, COLING 2014, pp. 62–66. http://doc.utwente.nl/94056/

[Nguyen *et al.* 2013] Nguyen, D.P.; Gravel, R.; Trieschnigg, R.; Meder, T. (2013) “how old do you think i am?" a study of language and age in Twitter, In ICWSM. in Proceedings of the *Seventh International Conference on Weblogs and Social Media*.

[Nicoletti 1994] Nicoletti, M. D. C. (1994) Ampliando os limites do aprendizado indutivo de máquina através das abordagens construtiva e relacional. Tese de Doutorado, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos.

[Nicoletti 2018] Nicoletti, M. C. (2018) Tópicos de Aprendizado de Máquina e Técnicas Subjacentes, CRV Editora, 330 pgs.

[Nuts & Rousseeuw 1996] Nuts, R., Rousseeuw, P. (1996) Computing depth countours of bivariate points clouds. *Journal of Computational Statistics and Data Analysis*, 23:153-168.

[Ortiz-Ángeles *et al.* 2017] Ortiz-Ángeles, S.; Villuendas-Rey, Y.; López-Yáñez, I.; Camacho-Nieto, O. and Yáñez-Márquez, C. (2017) Electoral Preferences Prediction of the YouGov Social Network Users Based on Computational Intelligence Algorithms. *Journal of Universal Computer Science*, vol. 23, no. 3, 304-326.

[Pennebaker *et al.* 2007] Pennebaker, J.; Chung, C.; and Ireland, M. (2007) The development and psychometric properties of LIWC2007. Austin, TX.

[PNUD 2020] Programa das Nações Unidas para o Desenvolvimento (PNUD). https://www.br.undp.org/, acessado em: 14/07/2020.

[PNUD 2020b] Programa das Nações Unidas para o Desenvolvimento (PNUD) – Relatório do Desenvolvimento Humano (2005) http://hdr.undp.org/sites/default/files/hdr2005-portuguese.pdf, acessado em: 14/07/2020.

[Portal da Transparência 2020] Portal da Transparência, http://www.portaltransparencia.gov.br/sobre/legislacao, acessado em: 18/07/2020.

[Poteras *et al*. 2014] Poteras, C. M., Mihaescu, M. C. and Mihai, M. (2014) An optimized version of the Kmeans clustering algorithm. In: Proc. of the 2014 Federated Conference on Computer Science and Information Systems, pp. 695–699.

[Prabhu *et al*. 2019] Prabhu, B. A., Ashwini, B., Khan T. A., & Das, A. (2019) Predicting election result with sentimental analysis using twitter data for candidate selection. In Innovations in Computer Science and Engineering, pages 49–55. Springer.

[Rand 1971] Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, American Statistical Association, v. 66, no. 336, pp. 846-850.

[Raghavan *et al*. 2007] Raghavan, U. N.; Albert, R.; and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. Physical Review E.

[Rosenberg *et al.* 2005] Rosenberg, C.; Hebert, M.; Schneiderman, H. (2005) Semi-supervised self-trainig of object detection models, Inc: Proc, of the *IEEE Workshop on Application of Computer Vision*, pp. 26-36.

[Rousseeuw 1987] Rousseeuw, P. (1987) Silhouetters: a graphical aid to the interpretation and validation of cluster analysis, J. Computational Applied Mathematics, v. 20, no. 1, pp. 53-65.

[R Project 2020] R Project (2020), R Project, Inc. https://www.r-project.org/, acessado em: 12/10/2020.

[Russel & Norvig 2009] Russel, S., Norvig, P. (2009) Artificial Intelligence: a modern approach, 3rd edition. Prentice Hall, 2009.

[Sanders *et al.* 2016] Sanders, E.; Gier, M.; Bosch, A. v. d. (2016) Using demographics in predicting election results with Twitter, In: Proc. of the *International Conference on Social Informatics*, USA.

[Sanders & Bosh 2013] Sanders, E.; Van den Bosch, A. (2013) Relating political party mentions on Twitter with polls and election results. In: *Proceedings of DIR-2013*, pp. 68–71. http://ceur-ws.org/Vol-986/paper_9.pdf.

[Sang & Bos 2012] Sang, E.T.K.; Bos, J. (2012) Predicting the 2011 dutch senate election results with Twitter, In: Proceedings of the *Workshop on Semantic Analysis in Social Media*. pp.53–60.

[Santos 2005], Santos, C. N. (2005) Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro. Rio de Janeiro, 2005. Disponível em: http://www.comp.ime.eb.br/old-pos/?l=0&p=29&q=2005_5, acessado em: 15/07/2020.

[Spezio *et al.* 2012] Spezio, M., Loesch, L., Gosselin, F., Mattes, K., Alvarez, R.M., (2012). Thin Slice decisions do not need faces to be predictive of election outcomes. Polit. Psychol 33 (3), 331–341

[Spyder 2020] Spyder (2020), Spyder, Inc. https://www.spyder-ide.org/, acessado em: 07/07/2020.

[Tan *et al.* 2006] Tan, P.-N.; Steinbach, M.; Kumar, V. (2006) Introduction to Data Mining, Pearson Education, Inc.

[Theodoridis & Koutroumbas 2009] Theodoridis, S.; Koutroumbas, K. (2009). Pattern Recognition, 4th ed., USA: Elsevier.

[TSE 2020] Tribunal Superior Eleitoral. http://www.tse.jus.br/. acessado em: 20/07/2020.

[TSE 2018] Tribunal Superior Eleitoral. *Divulgação do Resultado das Eleições*. http://divulga.tse.jus.br/oficial/index.html, acessado em: 20/07/2020.

[TSE-Estatísticas 2020] Tribunal Superior Eleitoral. http://www.tse.jus.br/eleicoes/estatisticas/estatisticas-eleitorais, acessado em: 20/07/2020.

[TSE-Dados 2020] Tribunal Superior Eleitoral. http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais, acessado em: 20/07/2020.

[Tumasjan *et al.* 2010] Tumasjan, A.; Sprenger, T.; Sandner, P. G.; Welpe, I. M. (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. In Proc. of *4th ICWSM*, pp. 178–185.

[Tuomi 1999] Tuomi, I. (1999) Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory, Journal of Management Information Systems, v. 16, no. 3, pp. 103−117.

[Wang *et al.* 2015] Wang, W.; Rothschild, D.; Goel, S.; Gelman, A. (2015) Forecasting elections with nonrepresentative polls, *International Journal of Forecasting*, v.31, no. 3, pp. 980–991.

[Wesley *et al*. 2003] Wesley et al. (2003) Monitoring sedation status over time in ICU patients, Journal of the American Medical Association (JAMA), v. 289, no. 22, pp. 2983−2991.

[Wikipedia 2020] Artificial intelligence. https://en.wikipedia.org/wiki/Artificial_intelligence. acessado em: 01/07/2020.

[Wilson *et al.* 2005] Wilson T., Wiebe J., Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 347–354.

[Witten *et al.* 2011] Witten, I. H., Frank E., Hall M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[Xu & Wunsch 2005] Xu, R., Wunsch, D.C. (2005) Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, v.16, n.3, 2005, p. 645-678.

[Yasseri & Bright 2016] Yasseri, T.; Bright, J. (2016) Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Science*, 5(1):1.

[Zins 2007] Zins, C. (2007) Conceptual approaches for defining data, information, and knowledge, Journal of the American Society for Information Science and Technology, v. 58, no. 4, pp. 479−493.